CAPSTONE - MSc. In Business Analytics

A Sub-national Prediction Model for Estimating Fatalities in Africa: UN-Engaged Conflict Zones

Date of Submission: July 7, 2023

Authors: Adrià Termes, Foram Pakir, Marten Dybus, Leander Gruss & Sasha Sittek

Project No.: P132 – Group G10

University: ESADE Business School

Recipient Institution: United Nations – Jonathan Stewart

07/07/2023

Table of Contents

Table of Figures

Overview of Tables

Project Introduction

1.0.1. Global conflicts

Conflict and violence have plagued societies throughout history, resulting in immense human suffering and loss. Over the years, the nature and dynamics of conflicts have evolved, leading to changes in their patterns and impacts. This paper aims to develop a predictive model for estimating conflict fatalities in African countries, with a particular focus on nations where the United Nations (UN) has deployed peacekeeping missions. By understanding the factors that contribute to conflict-related deaths, we can improve our ability to mitigate the devastating consequences of such conflicts.

Globally, the distribution of conflict types has undergone significant transformations. Non-state-based conflicts, characterized by multiple armed groups operating within a territory, have experienced a marked increase in prevalence. In 1989, non-state conflicts represented 22% of all conflicts, but by 2016, this figure had risen to 45%. In contrast, one-sided events, where a dominant actor employs violence against weaker opponents, decreased from 31.4% to 18% during the same period (see figure below). It is important to note that conflicts included in this analysis are those in which at least 25 military and civilian deaths occurred as a result of the fighting.

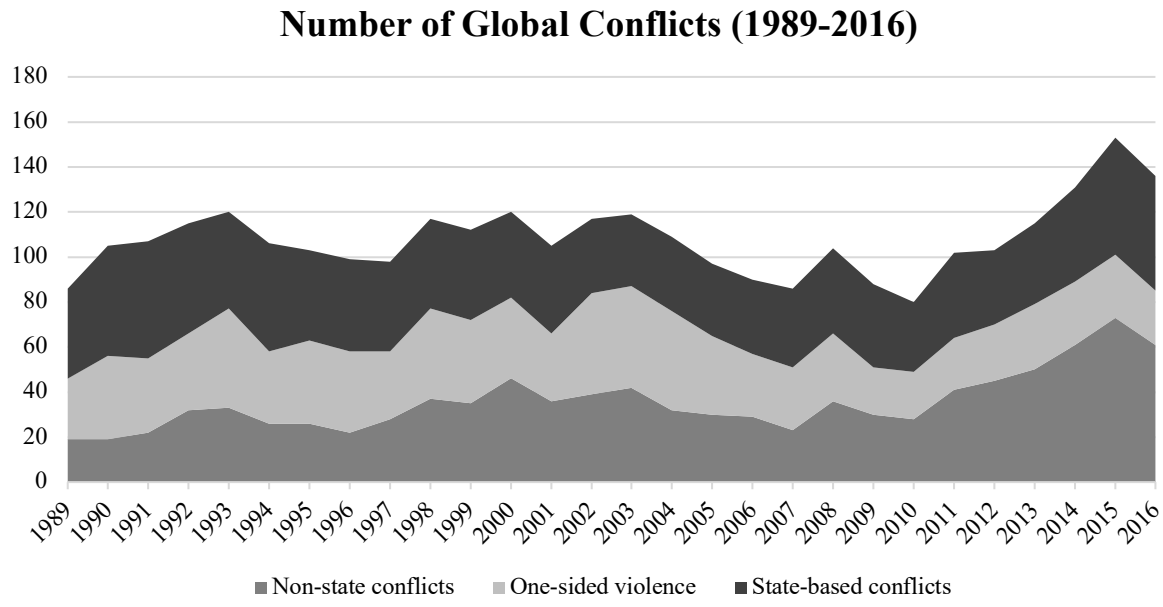

Figure 1 - Global Conflict Development. Note: an incident is considered a conflict when the total fatalities surpasses 25. Source: UCDP/PRIO

1.0.2. Global deaths

When considering the number of fatalities resulting from conflicts, state-based violence has traditionally accounted for 60-70% of total deaths globally. However, a significant shift has been observed in recent years, particularly in non-state and one-sided violence. Initially, non-state conflicts accounted for only 5-15% of conflict-related deaths, but this proportion has now risen to 20-30%. Conversely, one-sided events, which once accounted for 20-30% of total deaths, now represent a smaller percentage of 5-10% on a global scale.

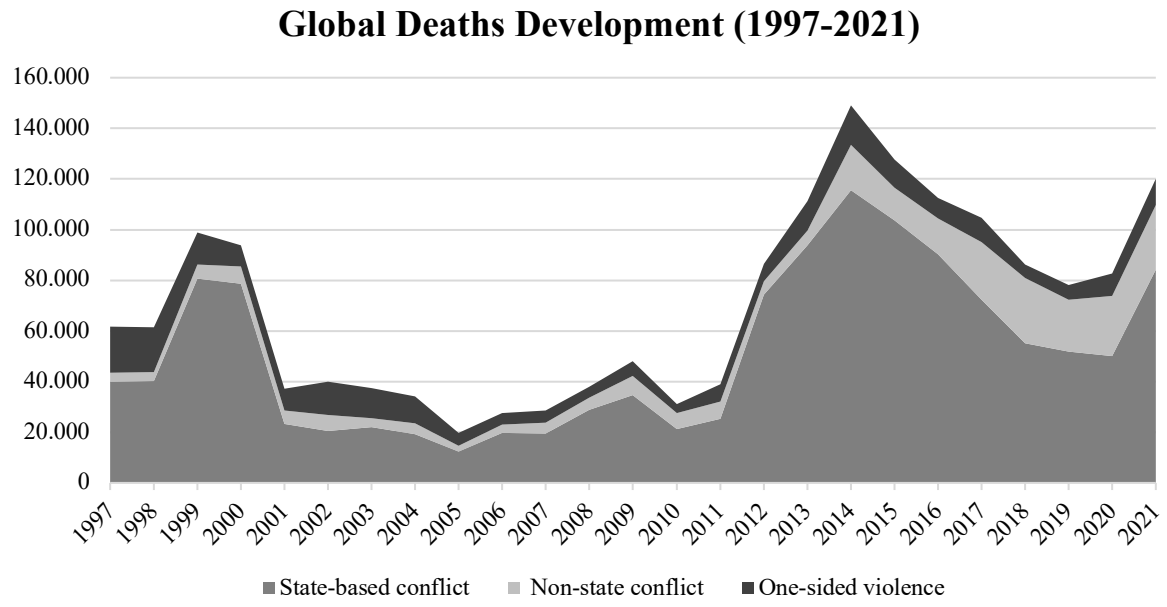

Figure 2 - Global Deaths Development. Source: PRIO/UCDP

1.0.3. African Conflict

Narrowing our focus to the African continent, we find that foreign states have become increasingly involved in state-based conflicts since 2007. This heightened activity by external actors has had far-reaching consequences, influencing the dynamics and severity of these conflicts. Understanding the reasons behind this trend is crucial for developing effective strategies to prevent and mitigate the impacts of state-based violence in Africa.

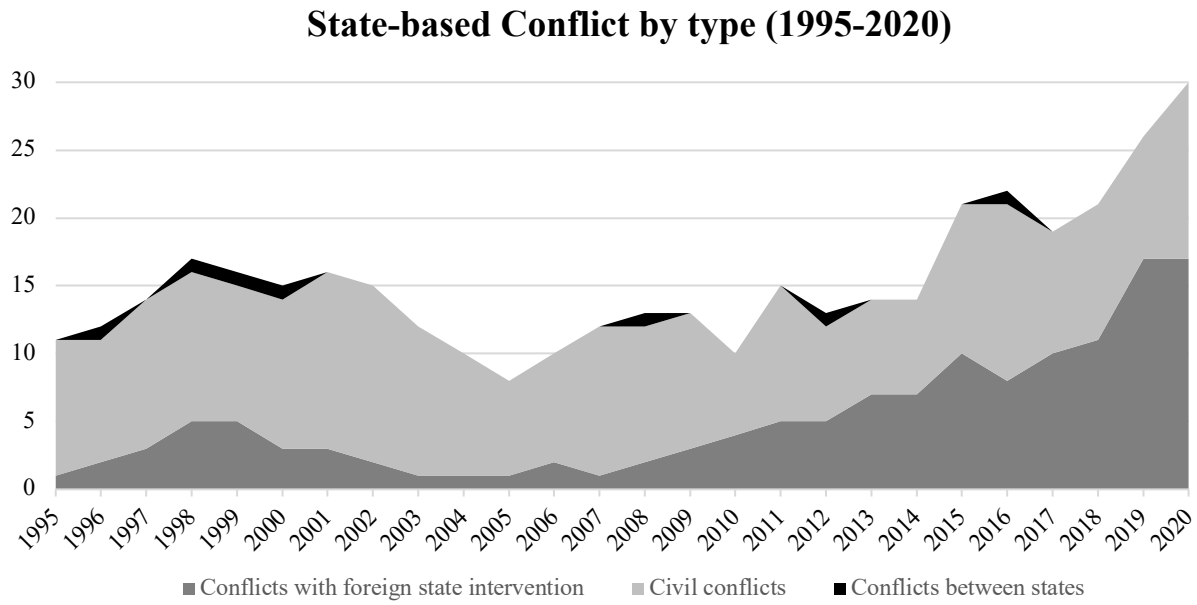

Figure 3 - State-based Conflict Development. Source: UCDP/PRIO

The high number of deaths resulting from state-based conflicts in Africa during the period of 1999-2001 can be attributed to several specific events and factors. One notable example is the Second Congo War (1998-2003), also known as the Great War of Africa, which involved multiple African countries and various armed groups. The conflict stemmed from the aftermath of the Rwandan genocide and the power struggle for control over the rich natural resources in the Democratic Republic of Congo (DRC). The involvement of numerous state and non-state actors, along with inter-ethnic tensions and competition over resources, resulted in a significant escalation of violence and a large number of fatalities. Another event that contributed to the high death toll in state-based conflicts during this period was the civil war in Sierra Leone (1991-2002). The conflict, characterized by rebel groups such as the Revolutionary United Front (RUF), was fueled by the illicit trade of diamonds, political instability, and grievances related to corruption and marginalization. The RUF's brutal tactics, including widespread atrocities and the use of child soldiers, led to a substantial loss of life and displacement of civilians.

The peak in one-sided conflict-related deaths in Africa during 2014-2015 can be attributed, in part, to the activities of extremist groups such as Boko Haram in Nigeria. Boko Haram's violent insurgency, marked by suicide bombings, massacres, and kidnappings, resulted in a significant number of civilian casualties. The group's aim to establish an Islamic state and its targeting of schools, villages, and public spaces led to a sharp increase in fatalities during this period.

The concerning increase in state-based conflict fatalities since 2019 can be linked to various ongoing conflicts and crises in different parts of Africa. For example, the ongoing civil war in South Sudan, which started in 2013, has resulted in a significant loss of life and displacement of populations. Similarly, the conflict in the Tigray region of Ethiopia, which escalated in 2020, has led to widespread violence and a humanitarian crisis, resulting in numerous casualties.

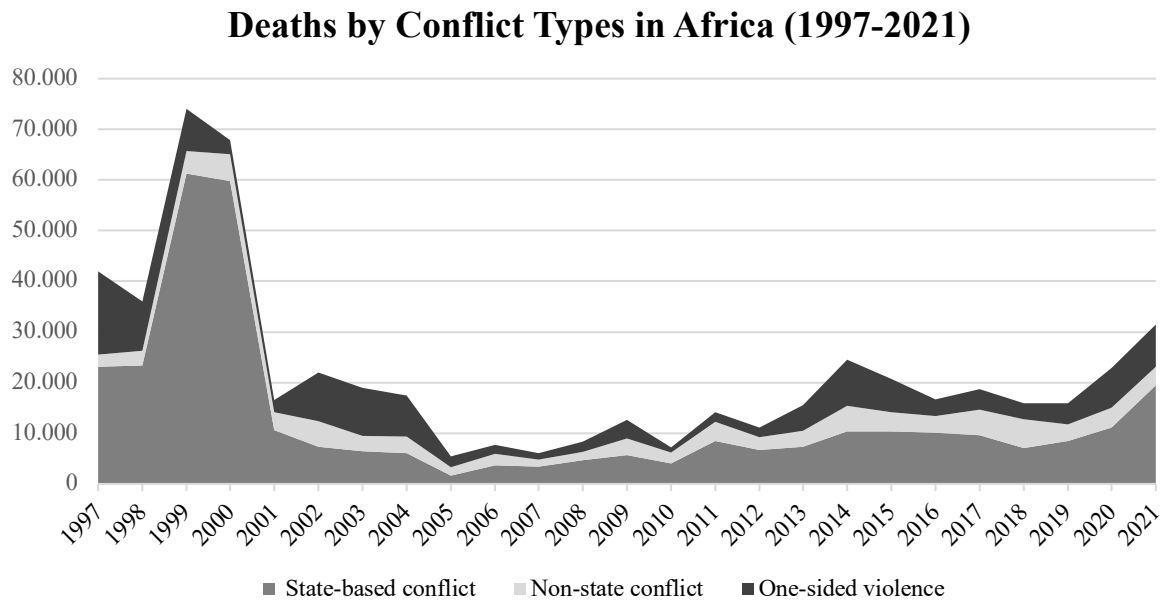

Figure 4 - Deaths by Conflict Types in Africa. Source: UCDP/PRIO

1.0.4. Diverging Fatality Rates

The United Nations (UN) has undertaken peacekeeping missions in certain countries in Africa with the goal of reducing conflict, promoting stability, and addressing underlying issues that contribute to violence. These missions are designed to facilitate peace negotiations, protect civilians, disarm combatants, and support the establishment of sustainable governance structures. However, the fatality

rates resulting from conflicts in these countries have shown variations over time, highlighting the diverse dynamics and underlying causes that exist within each nation. This contrast becomes evident when examining the continental scale, where spikes in fatality rates have been apparent.

Each country where the UN has a peacekeeping mission faces unique challenges and circumstances that shape the nature and intensity of the conflicts experienced. Factors such as historical legacies, political rivalries, ethnic tensions, economic disparities, and resource competition contribute to the dynamics of violence and the fatality rates observed. As a result, the efforts of UN peacekeeping missions must be tailored to the specific contexts and root causes of conflicts in each country.

The differing fatality rates over time within these countries further highlight the evolving nature of conflicts and the changing factors that contribute to violence. These variations can be observed in the temporal patterns of conflict-related deaths, where spikes in fatalities indicate shifts in the intensity or dynamics of the conflicts. Understanding these fluctuations is crucial for identifying the underlying causes and triggers of violence within each country.

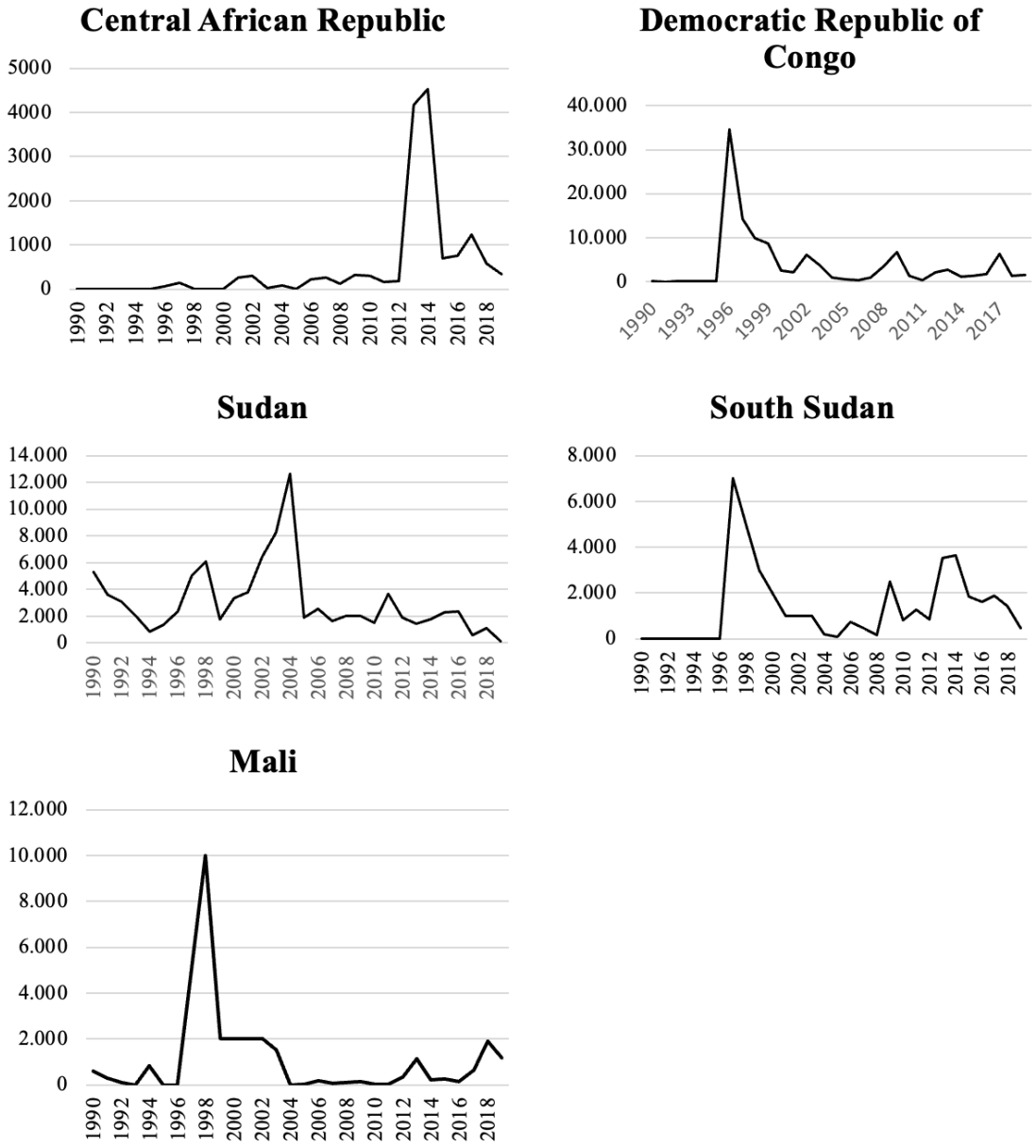

Figure 5 - Fatality Rates for UN-based countries in Africa. Source: UCDP/PRIO

By delving deeper into the factors that contribute to conflict fatalities in African nations with UN peacekeeping missions, this study aims to develop a predictive model that can help identify and prioritize interventions to mitigate the devastating effects of these conflicts. Understanding the patterns and dynamics of conflict fatality can inform policy decisions, enhance peacekeeping efforts, and contribute to the advancement of conflict resolution strategies on the African continent.

1.1 Project Context & Goal Definition

This capstone project was conducted by a group of five ESADE students who are pursuing an M.Sc. in Business Analytics. The project was carried out in collaboration with the United Nations (UN) and supervised by their experts. The main objective of the project is to explore the application of machine learning techniques within the specific work domain defined by the UN.

To achieve this objective, the project follows a structured approach known as the Machine Learning Lifecycle. This approach encompasses a series of well-defined steps that ensure a comprehensive and systematic process. By adhering to this lifecycle, the project team can effectively develop a robust model for identifying conflicts, which is of paramount importance for the UN. The Machine Learning Lifecycle approach is crucial for several reasons. Firstly, it provides a framework for organizing and managing the entire machine learning process, from data collection and preprocessing to model development and deployment. This systematic approach helps ensure that no crucial steps are missed and that the project progresses in a well-organized manner.

Secondly, the lifecycle approach emphasizes the importance of continuous evaluation and iteration. This means that the model will be constantly refined and improved based on feedback and new data. By iterating on the model, the project team can enhance its accuracy and effectiveness in identifying conflicts, which is crucial for the UN's mission.

Furthermore, the lifecycle approach promotes collaboration and transparency throughout the project. It enables clear communication and alignment between the project team and the UN, ensuring that the model's development aligns with the specific needs and requirements of the organization. An overview is depicted below.

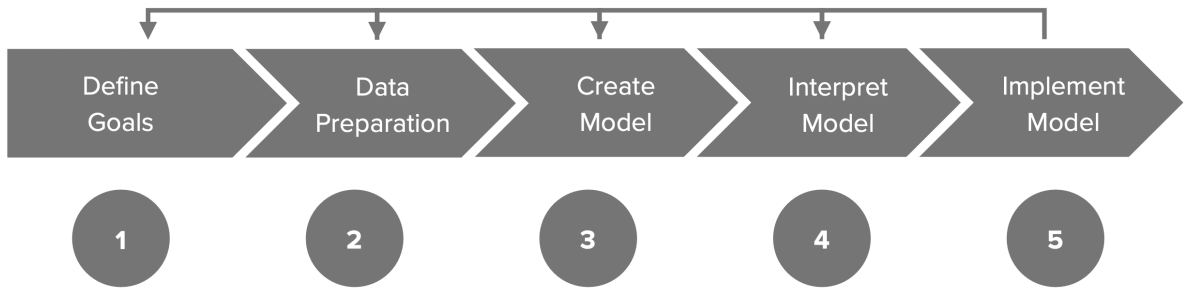

Figure 6 - Machine Learning Lifecycle

In this first chapter, we provide an overview of the project, outline the scope of work, and define the project.

1.2 The United Nations - Structure and Goals

1.2.1 Company Structure

The United Nations (UN) is an intergovernmental organization focused on international peace, security, and cooperation among nations. Within the UN, the Department of Operational Support (DOS) is dedicated to deploying responsible and efficient missions that minimize risks to people, societies, and ecosystems. The DOS prioritizes sustainable practices, considers environmental and social implications, and integrates them into mission planning and operations. By doing so, it aims to have a positive impact on natural resources, society, and ecosystems. The DOS collaborates with analytical subdivisions like the Service for Geospatial, Information, and Telecommunications Technologies (SGITT) to leverage data-driven insights for effective decision-making.

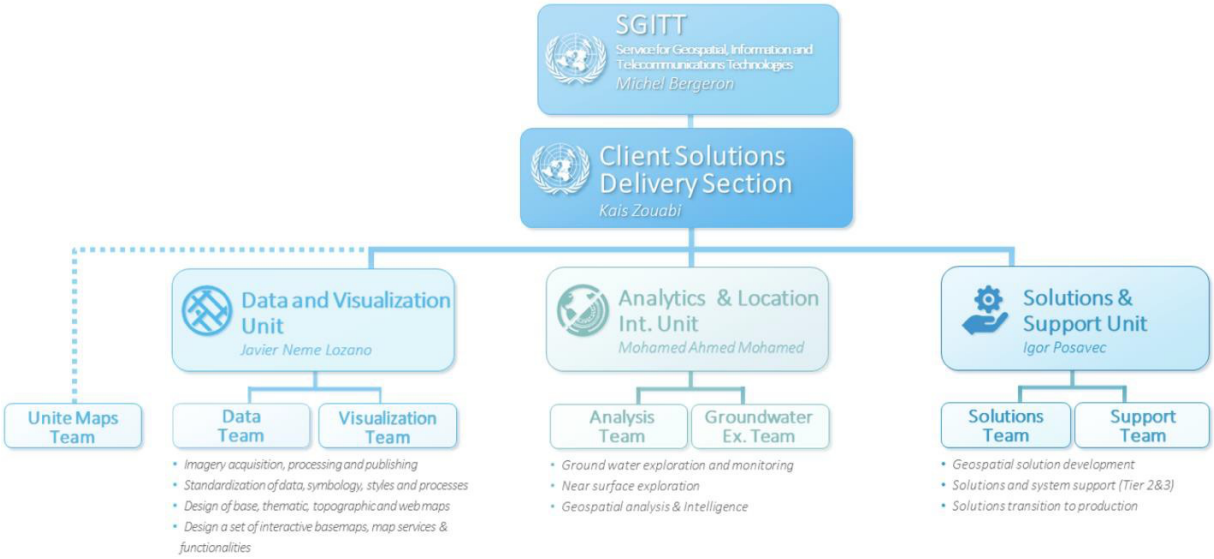

Figure 7 - Organizational Structure of the UN's SGITT Unit (UN, n.d.)

1.2.2. Unit Structure

The Support and Geospatial Information Technology Division (SGITT) serves as the operational hub for delivering ICT services and solutions to UN field operations. It operates from two locations: the United Nations Global Service Centre (UNSGC) in Brindisi, Italy, and the United Nations ICT Facility in Valencia, Spain. SGITT supports Peacekeeping operations, Special Political Missions, and UN Agencies, Funds, and Programs with round-the-clock, resilient ICT and geospatial information services.

Within SGITT, the Client Solutions Delivery Section provides comprehensive corporate field technology solutions and support. They offer a range of services, including data capture, management, visualization, analytics, business intelligence, reporting, solution implementation, support, and environmental services. With two data centers and Big Data Infrastructure tools, they effectively handle large data volumes. The Client Solutions Delivery Section has partnered with ESADE to develop a predictive analytics tool for assessing the rising risk of conflict in countries where the UN is active. This tool aims to proactively identify conflict risks, enabling timely actions to mitigate them. By leveraging ESADE's expertise, the division seeks to create a robust predictive analytics solution that enhances conflict risk assessment and supports informed decision-making. The goal is to enable the UN to

anticipate and respond effectively to emerging conflict situations, contributing to peacebuilding efforts and the organization's overarching mission.

1.3 Objectives & Scope

At the project's inception, defining the scope and objectives was necessary. Aligned with the United Nations guidelines and drawing from previous years' project work, the overarching objective can be succinctly summarized as follows:

"Create a predictive model to estimate the number of fatalities resulting from various conflict types at a sub-national level, specifically in countries where the UN is engaged in missions, with a prediction timeframe ranging from 1 to 12 months in advance."

In order to tackle this objective the project can be further broken down into sub-goals and corresponding process steps.

1) *Define geographic scope* - At the outset of the project, one of the initial steps involved defining the geographic scope. Given the United Nations' presence in various countries through its missions, it was crucial to assess and determine the specific areas that would be addressed. This assessment helped identify the regions where the project's focus and efforts could be directed, ensuring a targeted and meaningful approach in addressing conflicts and related issues.

2) *Collect data sources and pre-process data* - To augment the existing data sets and enhance the depth of information, it is imperative to source additional data that can provide valuable insights into predictors and drivers of conflicts. Subsequently, thorough data cleansing, merging, and establishing a pre-processing pipeline become essential steps to ensure the quality and integration of the data sets.

3) *Create a Conflict Prediction Tool* - The development and definition of indicators is required to evaluate the susceptibility of violent occurrences on a sub-national scale. The goal is to develop a tool for conflict forecasting that makes use of predictive analysis techniques. Exploring and assessing multiple machine-learning models is required for this. The product will be a collection of Jupyter notebook files that will serve as a thorough framework for completing the study and producing useful insights.

4) *Visualize the Predictions* - It is beneficial to properly visualize our findings in order to increase their clarity and dynamism. We can study variances across several dimensions, such as geography, time, and conflict type, using visualization to convey data in a clear and compelling way. We can develop deeper ideas, spot patterns, and convey complex information by using visual representations.

5) *Providing detailed documentation for future use* - The project is carried out in cooperation with the supervisory UN team, however, it's vital to stress that ESADE students are handling the outside execution of it. Therefore, thorough documentation is necessary to record the project's specifics. To ensure the project's repeatability after its completion and to enable others to comprehend and efficiently copy the technique, key steps, and processes will be thoroughly documented.

1.4 Key differentiation & improvement to previous work

Building upon the previous team's work, we identified additional factors, refined methodologies, and incorporated new data sources to expand the scope and depth of our analysis. This allowed us to provide a more comprehensive and robust solution, addressing the challenges and questions at hand with a fresh perspective. By combining the foundation laid by the previous team with our own contributions, we were able to deliver a solution that goes beyond the existing work and offers added value to the field of research. We were not only able to utilize their findings, but also differentiate, extend, and enhance the solution based on our own research and insights.

Current Models	**Value Add**
Models mainly focused on country-regionwide conflicts	Segment the subnational scale more granularly by deploying a grid approach → *Increase in accuracy of conflict predictions*
Current models focus on national conflict indicators primarily	Discover drivers on local scale and potentially enrich the data with national-wide variables → *Find relevant drivers for intranational conflicts*
Previous models focused on two specific countries (Sudan & Mali)	Extend the other geographic regions, while enabling the integration of new countries Subnational Approach Grid Approach (significantly more granular predictions) → *Ensure scalability of working models*
Models focused on specific variables from ACLED, HDI & WDI data sets	Significantly larger dataset in terms of Variables and Geographies and Grid Approach with improved performance and feature relationship conclusions
Results delivered as a notebook, no visualization	Visualization of results in Geospatial software and Increased implantability & explainability

Table 1 - Differentiation to previous Work

1.5 Our Team

Leander Gruss - With a bachelor's degree in business administration from the Rotterdam School of Management and ongoing master's studies in Business Analytics at ESADE, he combines academic excellence with practical experience. His two-year tenure in Berlin's venture capital sector, specializing in AI-based software companies, provided valuable insights into investments and commercial viability. Proficient in assessing project feasibility and profitability, he is committed to leveraging data-driven decision-making for successful project outcomes.

Marten Dybus - Graduated from the Technical University of Munich with a degree in Management & Having graduated from the Technical University of Munich with a degree in Management & Technology, he has a solid academic background. With professional experience in consulting and the German industry, he has developed a strong understanding of business operations. His expertise lies in effectively integrating technical solutions in computer science and AI with strategic business implementations.

Sasha Sittek - With a degree in Business Administration and a master's in finance from Erasmus University Rotterdam, he brings a strong foundation in financial analysis and quantitative methodologies. Having worked in consulting, he has honed his skills in conducting thorough quantitative analyses and finding effective solutions for complex problems.

Adrià Termes Berruga - Having successfully completed a bachelor's degree in business administration at ESADE, he has acquired a solid foundation in the field. Throughout his career, he has held various positions in the dynamic consumer goods market, gaining expertise in marketing, operations, and sales. His comprehensive understanding of these key areas enables him to contribute effectively to analyzing complex business challenges and delivering practical and innovative solutions.

Foram Nikhil - With a degree in Computer Engineering from Pandit Deendayal Energy University, she possesses a strong academic background in the field. Her expertise lies in coding and programming, showcasing excellent coding skills and abilities. Moreover, she has gained practical experience through various positions, successfully handling both front-end and back-end development tasks.

2 Methodology

This chapter on methodology aims to provide a comprehensive overview of the approach employed in analyzing conflict dynamics. It will address key aspects such as conflict variables, geographical scope, limitations of computing resources, frameworks for subnational analysis, and an overview of conflict drivers. Furthermore, it will explore visualization tools utilized for presenting predictive outputs.

2.1 Variable of study: Conflict

2.1.1 Conflict

In the realm of conflict research, the evolution of the definition of conflict utilized in empirical investigations has undergone significant transformation over time. Historically, researchers such as Buhaug & Rod (2007) and Raleigh & Urdal (2007) leveraged the use of thresholds, such as a minimum of 25 deaths annually, to designate a situation as an armed conflict. This allowed them to estimate the number of conflicts rather than the cumulative fatalities resulting from these conflicts, with the focus being placed on state-based conflict (Buhaug & Rod, 2007; Raleigh & Urdal, 2007).

The recognition of the varying types of conflict and their distinct underlying drivers, however, needed a shift toward a more nuanced approach. Gleditsch et al. (2002) notably pioneered this transition by introducing the UCDP/PRIO Armed Conflict dataset, which categorized conflicts as interstate, intrastate, and extra systemic. Further refining this approach, Sundberg & Melander (2013) introduced the UCDP Georeferenced Event Dataset, adding state-based, non-state, and one-sided violence into the analytical lens. This detailed categorization highlighted the heterogeneity of conflict drivers as argued by Hegre et al. (2019), emphasizing the significance of distinguishing conflict types when studying conflict patterns and dynamics (Gleditsch et al., 2002; Sundberg & Melander, 2013; Hegre et al., 2019).

2.1.2 Target Variable construction

Contemporary conflict studies are now transitioning towards capturing the number of conflict-induced fatalities, without imposing a threshold. This approach, as used by Schon & Kon (2022) in their AfroGrid study, allows for a more comprehensive understanding of conflict severity and its human cost. Counting each death, as opposed to setting a minimum threshold, ensures that every instance of lethal violence is included in the analysis. This approach, therefore, accommodates the diverse dynamics of conflicts and the multitude of event types, thereby reflecting a more accurate and inclusive picture of the true human cost of conflicts (Schon & Kon, 2022). For this reason, in this study, we will utilize this approach the UCDP categorization for conflict. In terms of the target variable, the focus will be on the accumulated conflict value across these categories. However, in the provided dataset the three types of conflict will be provided so the dataset can be used and adapted for further research.

2.2 Geographical Focus

2.2.1 Computing Resource Limitation

Our dataset provided includes extensive data on fatality, political, environmental, socio-economic and composite factors on every African country from 1989-2020 for 50x50km grids on a monthly basis. Still, this exceeds our computational resources resulting in that the respective continent-wide models cannot be run.

2.2.2 Adjusting Geographical Scope – UN presence

Consequently, in terms of geographical scope, we adjusted the focus to the appropriate context, namely where The United Nations currently has peacekeeping missions: Central African Republic, Democratic Republic of the Congo, Mali, Western Sahara, South Sudan, and Sudan. Although we initially chose a continental approach, reducing the geographical scope led to stronger consideration of differentiation of the main causes of conflict in the respective countries.

For example, in the Central African Republic (CAR), conflict drivers primarily stem from ethnic and religious tensions, along with competition for political power and control over natural resources. The historical animosities between different ethnic and religious groups, such as the majority Christian and minority Muslim communities, have fueled violence and societal divisions. Additionally, competition for control over valuable resources, including diamonds, gold, and timber, has contributed to the persistence of armed groups and instability in the country. The Democratic Republic of the Congo (DRC) faces a complex set of conflict drivers. The competition over mineral resources, particularly in the eastern provinces, has fueled armed conflict involving both local and foreign actors. Political instability, regional power struggles, and weak governance structures have also played significant roles. Moreover, historical legacies of colonization, exploitation, and resource plundering have left lasting scars on the country, contributing to its protracted conflicts. In Mali, conflict drivers encompass a combination of political, socioeconomic, and ethnic factors. The country has experienced political instability, including military coups, which have created a volatile environment. The rise of militant Islamist groups, such as Al-Qaeda in the Islamic Maghreb (AQIM) and other affiliated factions, has further exacerbated the conflict. Ethnic tensions between different groups, such as the Tuaregs and the central government, have also contributed to ongoing violence and challenges to stability. Western Sahara's conflict revolves around the dispute over self-determination between the Polisario Front and Morocco. The drivers of conflict in this region include historical and political factors, including the colonial legacy of Spanish colonization and the subsequent territorial claims made by both Morocco and the Polisario Front. The competing interests of the various stakeholders, as well as the impact of resource exploitation, further complicate efforts to reach a resolution. In South Sudan, conflict drivers include political power struggles, ethnic tensions, competition for resources, and weak governance. The struggle for power and control over state resources has led to violent clashes between rival factions. Ethnic divisions, particularly between the Dinka and Nuer communities, have played a significant role in the conflicts, as political disputes have often taken on an ethnic dimension. In Sudan, conflicts have emerged from various drivers, including the long-standing Darfur crisis, regional disparities, and competition for resources. In Darfur, the conflict has been fueled by a combination of factors such as marginalization, land disputes, and tensions between nomadic and sedentary communities. Regionally, the disparities between the center and the periphery have contributed to conflicts, including those in South Kordofan and Blue Nile states. Moreover, competition for resources, such as oil and arable land, has been a driving force behind conflicts in different parts of the country (i.e. CIA, 2021; US Department of State, 2021; UN Council, 2021).

By considering these specific differences in conflict drivers, it becomes evident that a tailored approach utilizing country-specific ML models is crucial for accurate analysis, because of diverging feature importance for predicting fatality from violent conflicts. Such an approach enables a deeper understanding of the unique challenges and opportunities present in each country, facilitating the development of precise recommendations for decision-making, policies, and strategies that effectively support peacekeeping missions.

2.3 Framework for Subnational Analyses

2.3.1 Introduction

Civil conflicts, like any significant political events, are typically examined and comprehended at the national level. However, it is important to acknowledge that the intensity of various factors associated with civil wars can vary geographically within states. Consequently, employing country-level approximations of local phenomena in statistical studies of civil wars may introduce potential flaws. Hence, Buhaug and Rod (2006) proposed to investigate the extent to which certain spatially varying features, including terrain, population, resources, and identity, contribute to the explanation of sub-national civil war outbreaks.

To achieve this level of granularity, it becomes necessary to employ a unit of observation that goes beyond the country scale. One possible approach could involve utilizing the first-order administrative entity as the unit of observation (e.g. Murshed & Gates, 2005; Østby, Nordås & Rød, 2009). However, unlike international boundaries, which generally remain unchanged over time once established and agreed upon, administrative regions frequently undergo modifications in terms of their shape and composition due to mergers and divisions. Furthermore, the function and size of regions can vary significantly from one country to another, making it impractical to uniformly divide all countries into smaller units.

2.3.2 Grid Squares Methodology

100x100km Grids

As an alternative method, grid cells have been adopted as the units of observation in literature, as initially introduced by Buhaug and Rod (2006) in their estimation of conflicts in the context of African civil wars between 1970 and 2001. In contrast to sub-national political regions, the grid system, once established, maintains a consistent size, shape, and number over time. Additionally, a substantial portion of the relevant geo-referenced data is presented as points, lines, or polygon features (such as diamond sites, main roads, and language), or as raster data (such as population density), which can be readily converted to the predefined grid cell structure. This methodology was further employed and expanded upon by Raleigh and Urdal (2007) in their investigation of the impact of the environment and climate on conflict, utilizing the 100x100km grid developed by Buhaug and Rod (2007) for conflicts occurring between 1990 and 2004. Moreover, this research approach contributes to the existing body of knowledge by analyzing the impact of the environment on internal armed conflicts using georeferenced (GIS) data and small geographical units of analysis, rather than relying solely on political boundaries.

50x50km Grids

The first standardized structure in the field of spatial data storage, manipulation, and analysis was presented by Toleffsen et al. (2012) with the introduction of the PRIO-GRID, a standardized structure for handling high-resolution spatial data. The PRIO-GRID framework encompasses a two-dimensional vector grid network with a resolution of 0.5 x 0.5 decimal degrees or a 50x50km grid, providing comprehensive coverage of all terrestrial areas across the globe. This innovative approach not only facilitates the organization and processing of spatial data but also incorporates a wide array of political, economic, demographic, environmental, and conflict variables for the entire time span between 1946 and 2008. This utilization of a grid-based approach has proven to be highly effective in the literature for studying conflicts on a subnational scale.

In terms of application, Hegre et al. (2019) introduced ViEWS, a political violence early-warning system for Africa that offers monthly forecasts for a duration of 36 months into the future that relies on the PRIO-GRID framework developed by Toleffsen et al. (2012). These forecasts are generated at both the country and subnational levels, encompassing the three types of organized violence identified

by the Uppsala Conflict Data Program (UCDP): state-based conflict, non-state conflict, and one-sided violence. The ViEWS forecasts are constructed based on a variety of constituent models that incorporate insights derived from extensive quantitative research on peace and conflict spanning several decades. These constituent models can be categorized into two types. Firstly, there are thematic models that focus on specific topics such as conflict history, the economy, political institutions, and geography. These models delve into the intricacies of these particular areas to generate predictions and insights regarding future political violence. Secondly, there are more generalized models that combine multiple themes or utilize information from both the country and sub-national levels to generate forecasts. Subsequently, the forecasts generated by these individual models are combined to create ensembles. By aggregating the forecasts, ViEWS aims to provide a more robust and comprehensive understanding of potential future political violence patterns and their spatial distribution.

Despite its reliance on publicly available data, it is important to note that the constituent models integrated into ViEWS may have non-matching time windows for their input data. This temporal mismatch between the various models may result in the potential failure to accurately capture the underlying interaction dynamics. It highlights the need for careful consideration and interpretation of the ViEWS forecasts.

2.3.3 Selected approach for subnational analysis

A recent contribution to the PRIO-GRID framework has been introduced by Schon and Koren (2022). In their study, they address the challenge of time-mismatching and propose a solution that enables the capture of interaction effects. Additionally, they address the limitations of previous studies, which often relied on outdated information and exhibited inconsistencies in terms of temporal granularity (monthly vs. yearly). To overcome these issues, Schon and Koren standardized multiple new data sources related to conflict, environmental stress, and socioeconomic factors. Hereto, they introduced Afro-Grid, an integrated and disaggregated dataset at a resolution of 0.5 degrees and a monthly temporal scale. This dataset encompasses information on conflict, environmental stress, and socioeconomic features in Africa, spanning the period from 1989 to 2020. The standardized approach allows for the continuous processing of new data and model training.

In line hereto in this study the PRIO-GRID framework along the approach of Afro-Grid. By employing grid cells as units of observation, we are then able to overcome limitations associated with country-level approximations and capture the geographical variations in factors contributing to civil wars.

Especially, if along the implementation of PRIO-GRID and its high 50x50km granularity, the complexity of conflict dynamics and patterns increases significantly as found by Hegre et al. (2019). As a result, an extensive selection of drivers is needed to provide coverage and improve the predictive capabilities of the models. Here, we expand on previous work conducted by providing: higher granularity in terms of geography; extended range of dependent variables and; improved model performances.

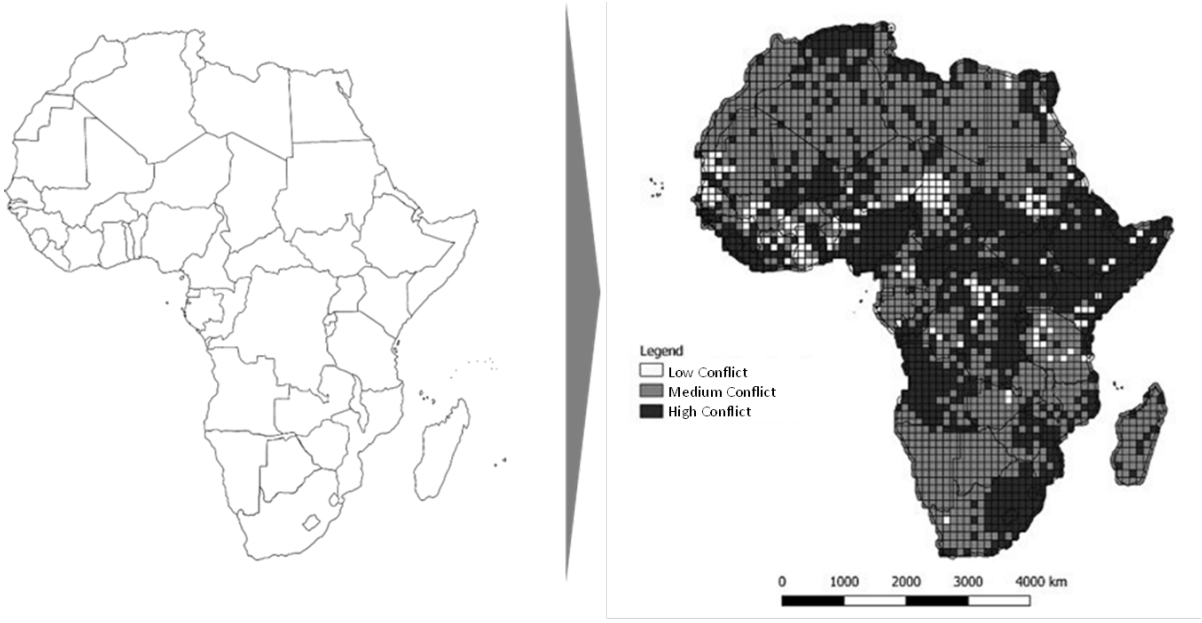

Figure 8 - Standardized approach to subnational analysis – PRIO-GRID
(Note: values are indicative and act as placeholders)

2.4 Drivers of Conflict & Databases

2.4.1 Introduction

The inclusion of drivers to violent conflict prediction models has been extensively informed by conflict research literature. In literature, variables, such as a history of past conflicts, population size, and socio-economic development levels, have shown consistent correlations with civil conflict occurrences (Blattman & Miguel, 2010; Hegre & Sambanis, 2006). Economic conditions significantly impact conflict risk, influencing societal group dependencies, public grievances, rebel recruitment costs, and state counterinsurgency capabilities (Collier & Hoeffler, 2004; Fearon & Laitin, 2003).

2.4.2 Drivers of conflict

Conflict History (Conflict Trap)

Research indicates that conflict history play a crucial role in predicting new armed conflict, as they underscore the "conflict trap" phenomenon, which suggests that historical conflict increases the likelihood of its continuation, recurrence, escalation, and diffusion. Analyses, such as those by Hegre et al. in 2017 and 2019, show the relationship between past and future violence, demonstrating how a large low-income country can reduce its conflict years by successfully containing violence and how regions with a dense history of conflict, such as northern Nigeria, the Kivu provinces in DRC, Somalia, and Darfur, often predict future outbreaks due to their conflict history. This self-perpetuating effect of conflict extends the risk of conflict for a country more than 20 years into the future, even for previously peaceful low-income countries. Furthermore, studies like that of Schon & Koren in 2022 show the significance of including lagged number of fatalities as an environmental variable in conflict prediction models. These findings solidify the importance of including lagged conflict variables in prediction models to anticipate and potentially mitigate future conflicts. Hereto, we utilize a *1, 2, and 3-month lag* for the accumulative value for the three main types of conflict available in the GED/ACLED database.

Political Stability

The integration of social conflict variables into the violent conflict prediction model is underscored by compelling empirical evidence, suggesting their potential to escalate into violent events (Salehyan, 2012; Hegre et al., 2019). Instances like the Arab Spring protests of 2011-2012 in North Africa and the Middle East, and the ethnic riots following Kenya's 2007 elections, highlight the risks embedded in these conflicts (Salehyan, 2012). Such *demonstrations, riots, strikes, and pro/anti-government violence* often intensify into sustained armed conflict, as observed in Syria (2011) and Burundi (2015) when demands are not met, especially if confronted by violent governmental responses (Tilly, 1978; Hegre et al., 2019). These insights, substantiated by data from databases like PITF/ACLED and SCAD, reveal a high probability of conflict throughout Africa linked to protest themes, underscoring the importance of incorporating these variables into the conflict prediction model (Hegre et al., 2019).

Social-economic Development

In order to measure local development Koren & Sarbahi (2018) utilized the nighttime light emissions. The development of *harmonized nighttime light (NL)* indicators as used by Schon & Kon (2022) in their Afro-Grid study, based on Li et al.'s (2020) method, has facilitated the generation of annual estimates for each 0.08-degree pixel, allowing for more accurate socio-economic assessments. The broadened temporal coverage and sensitivity to medium-to-low emissions makes this indicator beneficial for analyzing development and economic impacts at a local level. *Population density* also been identified as a significant driver of conflict. Sustained population growth combined with poor economic performance, especially in developing countries, may increase future conflict risks (Hegre, 2016). To operationalize this, data from the World Population Counts (WorldPop) dataset, which provides annual population estimates between 2000 and 2020 using census data, are utilized.

Environment

Environmental variables hold significant potential as predictors of conflict based on a body of recent scientific research. According to Hsiang, Burke, & Miguel's (2013) meta-analysis, climate change, especially changes towards warmer temperatures or more extreme rainfall, could potentially increase the frequency of both interpersonal violence and intergroup conflict. This study suggests that each one standard deviation change in climate can escalate the frequency of interpersonal violence by 4% and intergroup conflict by 14%. Further reinforcing this point, studies by O'Loughlin et al. (2012) demonstrate that deviations from typical precipitation and temperature patterns can substantially elevate the risk of conflict. They also provide a comprehensive view of how climate change effects can trigger violence at both local and national levels, using a rich database of geolocated violent events for East Africa from 1990 to 2009. Their findings show that warmer-than-average temperatures can exacerbate conflict, while positive rainfall deviations can potentially alleviate conflict in pastoralist sectors.

However, it is essential to note that the relationship between climate and conflict is complex and has yielded mixed and inconclusive results, as highlighted by Buhaug et al. (2014). This complexity is further compounded by seasonal environmental changes in rainfall and temperature, which could have a more significant influence on conflict trends than annual averages. Therefore, it is crucial to include environmental variables in a conflict prediction model to account for these nuances and to provide a more accurate and comprehensive picture of conflict dynamics.

Schon & Koren (2022) integrated the *Normalized Difference Vegetation Index (NDVI; Vegetation density)* into their AfroGrid study, thus presenting a comprehensive set of measures related to environmental fluctuations and stressors. The NDVI is a remote sensing indicator of greenness, reflecting aspects such as grassland areas and agricultural productivity. This information was derived from NASA's MODIS Terra satellite system, which records NDVI data at a 0.08-degree (1 km) pixel level worldwide, on a monthly basis from 1992 to 2018. The NDVI values for each pixel span from 0

(no vegetation) to 1 (full vegetation), although water-covered cells can exhibit negative values (from -1 to 0). The researchers processed the NDVI data to a 0.5-degree cell-month format, generating three distinct NDVI indicators. The first, NDVI (mean), was created by calculating the average NDVI values across all 0.08-degree pixels within a specific grid cell for each month. The other two indicators, NDVI (max) and NDVI (min), were derived by extracting the highest and lowest monthly values from the pixels within a cell, respectively. These time-variant indicators offer a significant enhancement over static measures like cropland coverage, which were used in previous research, as they reflect the dynamic nature of environmental conditions.

In their 2022 study, Schon & Koren further expanded the environmental variables in Afro-Grid by introducing indicators of temperature, precipitation, and drought. These indicators were derived from the CRU TS high-resolution multivariate climate dataset, which provides monthly data. The authors calculated these indicators by averaging the monthly information for each measure across all the pixels within a specific grid cell. In addition to these primary indicators, they also created two anomaly indicators which measure monthly deviations from the norm in *temperature* and *precipitation (mean & anomaly)*. These anomaly measures were generated by calculating rolling 30-year Z-scores, enabling users to ascertain whether the climate in a particular month of a given year deviates from what is typically expected for that time of the year. This anomaly analysis was performed on the entire dataset spanning from 1901 to 2020 prior to its integration into AfroGrid. This approach allowed the authors to provide anomaly measurements – the Temperature (anomaly) and Precipitation (anomaly) indicators – that cover the full temporal range and geographical extent of their data.

The *Standardized Precipitation Evapotranspiration Index (SPEI)* is a drought indicator that quantifies the disparity between precipitation and potential evapotranspiration (PET) over a specified temporal scale. This index is derived by utilizing temperature and precipitation data sourced from the CRU TS monthly gridded dataset. Positive values of the SPEI signify an excess of water, while negative values indicate an insufficiency in water supply. In our study, we employ the indicators calculated by Schon and Koren (2022), who have aggregated the monthly information from all pixels within a grid cell. Alternatively, the SPEI values can be obtained from the SPEIbase, a global database developed by Begueria, Serrano, and Martinez in 2010. The SPEIbase offers comprehensive and reliable information regarding drought conditions, spanning a spatial resolution of 0.5 degrees and encompassing the time period from January 1901 to December 2020.

Natural Resource Availability

The growing importance of environmental variables in conflict prediction models is predicated upon their demonstrated impact on the intensity and spatial distribution of conflicts, and especially geographically specific to Africa. This postulation is substantiated by empirical studies such as Balestri (2012), which identified a distinct spatial propensity of civil conflict, greatly influenced by the proximity and abundance of natural resources like diamonds and gold. The exploitability of these resources, consequential to their 'lootability' and wartime accessibility, exacerbates conflict intensity, thus asserting a significant role in the dynamics of conflict. Complementary research, including that of Ross (2004) and Lujala, Gleditsch, and Gilmore (2005), shed light on the economic aspect of lootable resources, particularly secondary resources like diamonds. These studies underscored the role these resources play in fomenting civil wars, especially in ethnically fragmented nations. Additionally, Lujala (2007) indicated that oil-producing countries frequently exhibit increased susceptibility to conflict, owing to the lucrative potential of wealth accumulation via hydrocarbon exploitation. The compelling evidence from these studies accentuates the imperative of including environmental variables in conflict prediction models, given their profound and direct influence on conflict trajectories and outcomes. Hereto we consider the presence of *petroleum, diamonds* and *gold.* The data for this study will be derived from the PRIO-GRID (Hegre et al., 2019), where information is consolidated at a granularity of 50x50km grids.

This data amalgamation leverages underlying datasets including the Petroleum Dataset from Lujala et. Al (2007), Diamond Resources by Gilmore et al. (2005), and GOLDATA from Balestri, Sara (2015).

Geographical Features

According to Hegre et al. (2019), conflicts are more likely to occur in border regions rather than in close proximity to countries' capitals. The adjacency of border regions creates opportunities for interaction and potentially increases the willingness to engage in conflict (Brochmann, Rod & Gleditsch, 2011). Therefore , we the PRIO-GRID data on the distance to the nearest border for each grid in our analysis. Additionally, the presence of rugged terrain appears to heighten the risk of conflicts, potentially due to disagreements during the demarcation of borders, as suggested by Gibler (2007). This implies that the ruggedness of borders plays a more significant role in determining conflict risk than the length of the borders themselves. Furthermore, states that share rivers experience a higher incidence of conflict. Scholars argue that this may be attributed to an increased willingness to fight over a valuable resource, such as water, or the inherent ambiguity and variability of river borders (Furlong, Gleditsch, and Hegre, 2006; Gleditsch et al., 2006; Toset, Gleditsch, and Hegre, 2000). For this we also integrated data on the *proportion of area covered by mountains*, originating from United Nations Environment Programme (UNEP) 'mountain watch' (available in PRIO-GRID).

National Indicators

Kim and Conceicão (2010) demonstrated a clear association between low *Human Development Index (HDI)* scores below 0.5 and political instability, as measured by indicators like Political Stability and Absence of Violence (Kaufman et al., 2006). This indicator assesses the 'likelihood of government destabilization or overthrow through unconstitutional or violent means, including politically motivated violence and terrorism'. Consequently, the presence of low HDI scores increases the risk of conflict outbreaks and recurrence, which is strongly linked to low human development. In result, self-reinforcing violence cycle may appear, similar to the previously discussed concept of "conflict traps". Hereto we utilized the HDI as set by the UN. The HDI is a composite measure three crucial dimensions: health, education, and standard of living, which are combined to generate a score ranging from 0 to 100%. The health dimension reflects factors such as life expectancy and access to healthcare, offering insights into the overall well-being of a population. Education is measured by indicators such as expected years of schooling, providing an understanding of educational opportunities and their potential impact on development. The standard of living dimension is captured through the indicator of Gross National Income (GNI) per capita. Areas with significant income disparities and widespread poverty are more susceptible to conflict, due to higher economic instability.

Category	Variables	Variable notation	Database
No. Of Fatalities	Fatality estimates for **state conflict**	*ged_nonstate_fatal_best*	GED
	Fatality estimates for **non-state conflict**	*ged_nonstate*	
	Fatality estimates for **one-sided violence incidents**	*ged_viol_fatal_tot_best*	
	Fatality estimates for **all incidents (Aggregated value)**	*target*	Own
	Lagged fatality estimates for **all incidents (T - 1 month)**	*lag1*	Calculation
	Lagged fatality estimates for **all incidents (T - 2 month)**	*lag2*	(GED)
	Lagged fatality estimates for **all incidents (T - 3 month)**	*lag3*	
Composite Index	**Human Development Index**	*hdi_index*	UNDP
Political Stability	Organized/Spontaneous **demonstrations**	*scad_(org/spont)_demo*	SCAD
	Organized/Spontaneous **violent riot**	*scad_(viol/spont)_riot*	
	General/Limited **strike**	*scad_(gen/lim)_strike*	
	Pro/Anti-**government violence**	*scad_(pro/anti)_gov*	
	Extra/Intra-**government violence**	*scad_(nsa/int)_gov*	
	Incidents defined as **battles** by actor	*acled_battle_(state/rebel…)*	ACLED
	Incidents defined as **remote violence** by actor	*acled_remote_(state/rebel..)*	
	Incidents defined as **violence against civilians** by actor	*acled_viol_(state/rebel..)*	
	Incidents defined as **riots**	*acled_riots*	
	Incidents defined as **protests**	*acled_protests*	
	Distance to own **border**	*own_borders_dist*	PRIO-GRID
Environment	**Vegetation** density indicator (NDVI)	*NDVI_mean*	AFROGRID
	Average **tempereature**	*t.avg*	
	Mean **deviation** (Z-score)of 30-year rolling **temperature** average	*t.nom*	
	Average **precipitation**	*p.avg*	
	Mean **deviation** (Z-score)of 30-year rolling **precipation** average	*p.anom*	
	Drought indicator (SPEI)	*droughtstart_speibase*	
	Area covered with **mountains**	*mountains_mean*	PRIO-GRID
	Gold placer/surface	*gold(placer/surface)_s*	PRIO-GRID
	Diamonds secondary	*diamsec_s*	
	Petroleum	*petroleum_s*	
Social-economic	Total nighttime light emission	*NL_sum*	AFROGRID
	Population density	*population*	
ID Indicators	**Month**	*month*	AFROGRID
	Year	*year*	
	Year & Month	*ym*	
	Latitude of grid centroid	*latitude*	
	Longitude of grid centroid	*longitude*	
	Grid ID (PRIO-GRID)	*gid*	
	Correlates of War ID	*COWCODE*	
	Name of **country**	*country.name*	

Table 2 - Feature Overview

2.5 Visualization for Sub-National Conflict Prediction

2.5.1 Prediction Output

In the field of conflict prediction, accurate and timely information is crucial for effective decision-making and resource allocation. Hereto, PRIO-GRID provides a highly granular and consistent geographical reference system for analyzing conflicts at a subnational level. In terms of output, our proposed prediction model estimates fatality counts in each grid cell on a monthly basis for the suggested 12-month period.

Despite the availability of extensive data points, grids with high conflict fatalities can still be challenging to identify accurately. Moreover, assessing time trends and identifying hotspots for conflict escalation can be complex without proper visualization tools. Due to the large number of grids and the challenge to identify where they are located subnational conflict prediction requires comprehensive data analysis and visualization techniques to uncover patterns and trends that may go unnoticed at the national level. Visualizing subnational conflict data allows analysts and decision-makers to identify patterns, trends, and correlations that may not be apparent through raw data alone. As a result, localized variations in conflict dynamics and areas with higher levels of violence can be identified.

2.5.2 GIS Software for Visualization

In the context of visualizing prediction output for all grid cells within the PRIO-GRID framework, GIS software proves to be an ideal solution. GIS (Geographic Information System) software offers a powerful platform for analyzing and displaying geospatial data, making it well-suited for visualizing the prediction output across the multitude of grid cells. GIS software provides several advantages for visualizing prediction output in this context:

- **Geospatial Integration:** GIS software enables the integration of diverse geospatial data layers, including the PRIO-GRID framework, prediction outputs, and additional relevant data such as demographic information or socio-economic factors. This integration allows for a comprehensive view of the prediction output in relation to the underlying geographic context.
- **Spatial Analysis:** GIS software provides a range of spatial analysis tools that can be employed to examine the prediction output across the grid cells. These tools facilitate spatial queries, overlay analysis, proximity analysis, and other operations that help identify spatial patterns, hotspots, and correlations within the prediction output data.
- **Customized Visualization:** GIS software offers a wide array of visualization techniques and customization options. Users can employ various symbology options, such as color ramps, graduated symbols, and thematic maps, to effectively represent the prediction output values for each grid cell. This enables the visualization of patterns, trends, and variations in conflict prediction across the subnational regions of interest.
- **Temporal Visualization:** GIS software platforms support time-enabled visualization. This functionality allows for the creation of animated or interactive maps that illustrate the prediction output over time. By animating the temporal aspect, decision-makers can easily comprehend how conflict predictions evolve and identify significant changes or patterns across different time periods. This precisely solves the matter of detecting conflict hotspots, allowing the UN to timely allocate resources in order to potentially mitigate resulting fatalities.
- **Spatial Interactivity:** GIS software provides interactive capabilities that enable users to explore the prediction output and associated data at different scales. This interactivity allows decision-makers to zoom in on specific areas of interest, click on individual grid cells to access detailed information, and perform on-the-fly spatial analysis to gain deeper insights into the prediction output.

2.5.3 GIS Software for Visualization – QGIS

We have selected QGIS as the primary tool for visualizing the prediction output within the PRIO-GRID framework for this project. QGIS, being a free open-source GIS software, offers a comprehensive set of functionalities that perfectly align with our project objectives. It provides a robust platform for integrating and analyzing diverse datasets, including the PRIO-GRID framework, prediction output data, and additional relevant layers.

In QGIS, we can effortlessly combine and overlay these datasets to develop a holistic understanding of the spatial patterns and trends in conflict prediction across the subnational regions of interest. The software's advanced visualization capabilities enable us to represent the prediction output with customized symbology, thematic mapping techniques, and dynamic time-enabled visualizations.

A significant advantage of creating visualizations in QGIS is the ease of utilizing the output files in other GIS software. QGIS supports a wide range of file formats, such as shapefile, GeoJSON, KML, and GeoTIFF, among others, for exporting visualizations. This compatibility ensures seamless sharing and utilization of the visualizations in various GIS software platforms, and is thus suitable for a variety of stakeholders that might use our methodology and output.

2.5.4 Setting up QGIS

To effectively utilize the output in QGIS, a series of steps were conducted to visualize and analyze the monthly prediction data for 1-12 months into the future. Initially, a PRIO-GRID global shapefile was integrated to visualize country borders. Subsequently, the shapefile was filtered to display only the continent of Africa as the base layer to address the region of focus.

Historical layers were incorporated into the visualization to provide context and demonstrate the transition into the prediction period. The inclusion of a subjective number of historical months allowed for a temporal progression leading up to the prediction phase. EPSG:4326 (WGS 84) is chosen as the coordinate reference system (CRS) for this project. It accurately represents the Earth's shape and ensures precise spatial representation of grid cells and associated data. WGS 84 is an adopted global standard for mapping and geospatial reference and supported by GIS software and promotes interoperability, allowing seamless integration with other geospatial data. By utilizing the EPSG:4326 CRS, designed for latitude and longitude coordinates, we ensure precise representation of the grid cells and associated data. This is particularly important when working with global datasets like PRIO-GRID, as it guarantees accurate positioning of the grid cells on the Earth's surface. In result, it allows consistency, enabling accurate spatial analysis and visualization of the prediction output.

For each prediction month, a separate layer was added on top of the shapefile, the global layer which is filtered for the African continent. The layers were sourced from CSV files containing prediction values and characteristics for each grid cell of each month. This approach facilitated the visualization of prediction values along with associated characteristics such as country, longitude, latitude, and distance to borders, among other features of choice. The incorporation of these layers allowed for a comprehensive representation of the prediction output within the spatial context. Utilizing the longitude and latitude data of each grid cell, the centroids were visualized to represent the spatial distribution of the predicted fatalities. This approach facilitated the assessment of monthly patterns and the identification of areas with higher fatality rates.

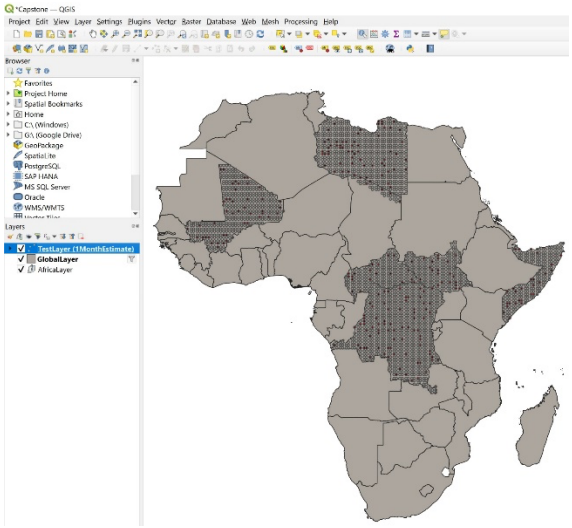

Figure 9 - QGIS – Geospatial visualization

Initially a global shapefile is added to QGIS in order to visualize the borders of the respective countries, and filter for the area of interest: Africa. By adding the respective historical/predicted fatality values, along the respective longitude and latitude the grids can be visualized. (Note: values are indicative and act as placeholders)

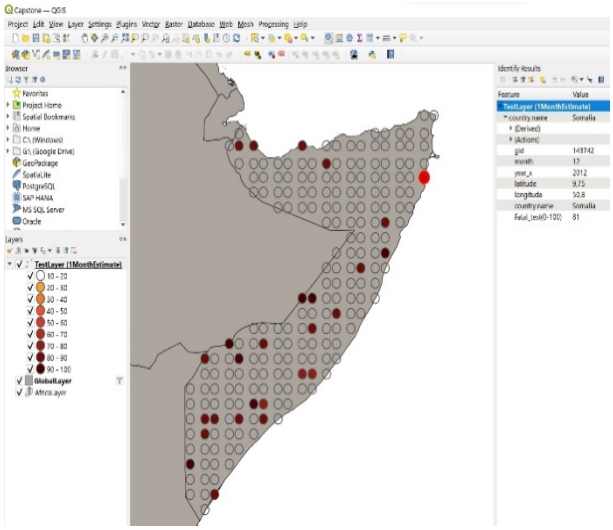

Figure 10 - QGIS – Grid Features

By clicking on the respective grid, all grid information can be found: country, longitude & latitude, historical/predicted fatality count etc. This can be extended by integrating more features in the respective CSV file that acts as a layer. Along the gradient scale is become immediately apparent which locations will see strong or little conflict (Note: values are indicative and act as placeholders).

To assess change over time, a sum vector was introduced. This vector was derived by summing the fatality counts from the past three months, providing a visual representation of the progression from the first month to the accumulated values of subsequent quarters. This visualization allowed for the observation of evolving dynamics and the identification of emerging hotspots. Utilizing the *"Time Manager"* function in QGIS, the changes over time were effectively animated and visualized. This feature enabled seamless transitions between monthly, quarterly, and yearly views, facilitating the identification of patterns, trends, and potential resource requirements or emerging hotspots at different temporal scales.

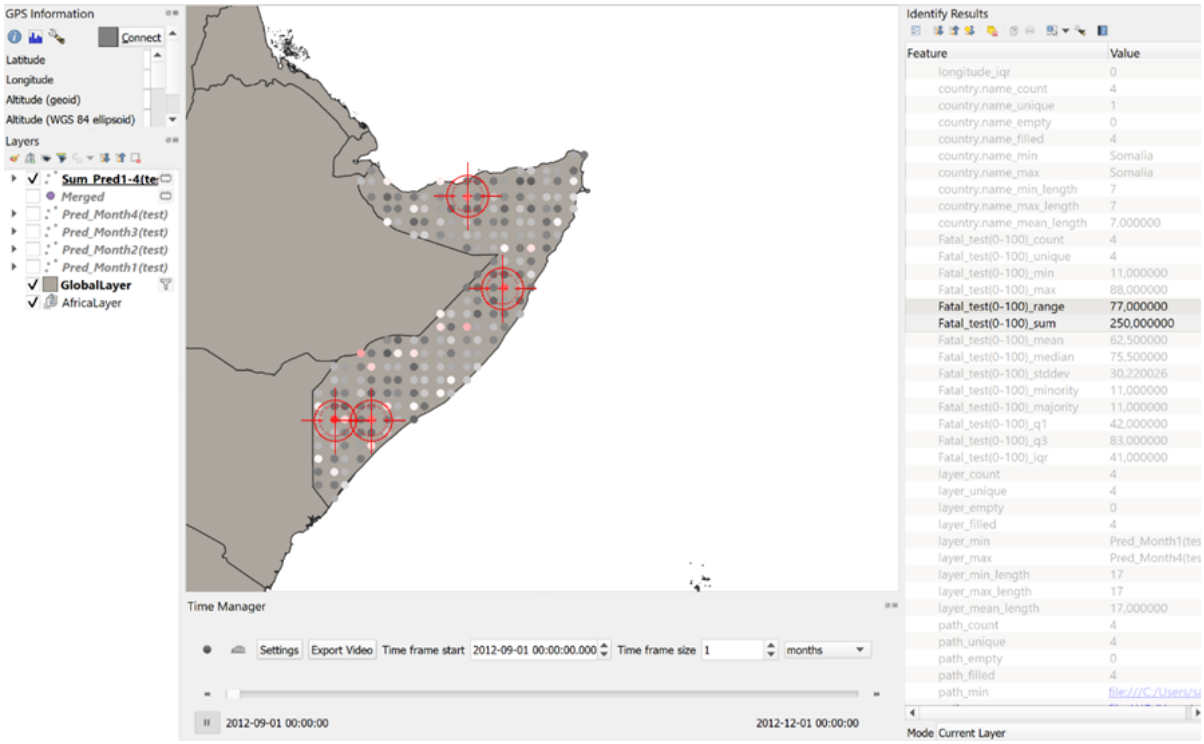

Figure 11 - QGIS – Hotspot Detection
By aggregating several prediction layers, it can be assessed and easily visualized which grids will see strong increases in fatalities arising from violent conflicts. Using the Time-Manager this can even be visualized over time. (Note: values are indicative and act as placeholders)

Overall, through the utilization of QGIS and the aforementioned steps, a comprehensive and dynamic visualization of the prediction output was achieved. The integration of CSV files containing prediction values and characteristics, along with the layers depicting centroids and sum vectors, provided a robust framework and standardized process for analyzing and interpreting the spatial and temporal dynamics of conflict fatalities. This visualization approach enhanced the ability to identify critical areas requiring urgent resources and anticipate emerging hotspots, thereby facilitating more efficient and effective decision-making in conflict prediction and resource allocation

3 Data Preparation

In the following section, we will outline the steps taken for data cleaning, preparation, and merging. This process ensures the quality and coherence of the data before conducting further analysis and model exploration. We handled missing values, outliers, and inconsistencies, transformed variables, scaled features and merged multiple data sources. These steps lay the foundation for deriving meaningful insights from the data.

3.1 Merging of Data Sources

In order to conduct a comprehensive model creation it is important to merge the respective data sets to find relationships between variables. To do so we merged the following data sets in that order:

1) **Afro-Grid**: Multi-faceted and holistic dataset covering multiple African countries aspects (violence variables, environmental, political...) from 1989 to 2020 at different levels of granularity. This dataset is the bedrock for all future mergings and analyses done in this project.

	gid	year	month	ym	ged_state	ged_nonstate	ged_viol_tot	ged_viol_state	ged_viol_nonstate	ged_state_fatal_best	...	scad_anti_gov	scad_n
0	62356	1989	1	1989-01-01	0	0	0	0	0	0	...	NaN	
1	62356	1989	2	1989-02-01	0	0	0	0	0	0	...	NaN	
2	62356	1989	3	1989-03-01	0	0	0	0	0	0	...	NaN	
3	62356	1989	4	1989-04-01	0	0	0	0	0	0	...	NaN	
4	62356	1989	5	1989-05-01	0	0	0	0	0	0	...	NaN	

Figure 12 - Afrogrid Data Set

2) **HDI**: HDI stands for Human Development Index, which is a statistical measure used to assess and compare the overall development levels of countries. The HDI takes into account indicators such as life expectancy, education, and income to provide a broader picture of human well-being and progress.

	Country	Continent	ISO_Code	Level	GDLCODE	Region	1990	1991	1992	1993	...	2012	2013	2014	2015	2016	2017	2018	2019
0	Afghanistan	Asia/Pacific	AFG	National	AFGt	Total	0.273	0.279	0.287	0.297	...	0.466	0.474	0.479	0.478	0.481	0.482	0.483	0.488
1	Afghanistan	Asia/Pacific	AFG	Subnat	AFGr101	Central (Kabul Wardak Kapisa Logar Parwan Panj...	0.332	0.339	0.349	0.361	...	0.548	0.552	0.553	0.548	0.551	0.553	0.555	0.561
2	Afghanistan	Asia/Pacific	AFG	Subnat	AFGr102	Central Highlands (Bamyan Daikundi)	0.281	0.288	0.297	0.308	...	0.480	0.483	0.483	0.477	0.479	0.479	0.480	0.484
3	Afghanistan	Asia/Pacific	AFG	Subnat	AFGr103	East (Nangarhar Kunar Laghman Nooristan)	0.287	0.293	0.301	0.311	...	0.468	0.469	0.466	0.459	0.461	0.463	0.464	0.469
4	Afghanistan	Asia/Pacific	AFG	Subnat	AFGr104	North (Samangan Sar-e-Pul Balkh Jawzjan Faryab)	0.259	0.265	0.274	0.284	...	0.466	0.480	0.492	0.497	0.500	0.501	0.502	0.507

Figure 13 - HDI Data Set

3) **PRIO-GRID**: PRIO-GRID refers to the Peace Research Institute Oslo (PRIO), an independent research institution based in Norway that focuses on studying peace and conflict. The PRIO-GRID data set contains information related to armed conflicts, peace processes, and other conflict-related variables.

	gid	row	col	xcoord	ycoord	diamsec_s	diamprim_s	goldplacer_s	goldsurface_s	mountains_mean	petroleum_s	ttime_mean
0	49182	69	222	-69.25	-55.75	NaN	NaN	NaN	NaN	NaN	NaN	1012.0180
1	49183	69	223	-68.75	-55.75	NaN	NaN	NaN	NaN	NaN	NaN	933.2906
2	49184	69	224	-68.25	-55.75	NaN	NaN	NaN	NaN	NaN	NaN	765.0629
3	49185	69	225	-67.75	-55.75	NaN	NaN	NaN	NaN	NaN	NaN	712.4025
4	49186	69	226	-67.25	-55.75	NaN	NaN	NaN	NaN	NaN	NaN	643.4555

Figure 14 - Afrogrid merged with HDI and Prio Data Set

4) **Lagged**: In the context of data analysis, lagged data refers to values from a previous time period used to predict or analyze trends in subsequent time periods. It involves shifting or delaying the data by a certain time interval, which allows for the examination of relationships and dependencies over time. As previously explained, we consider the lagged values for number of fatalities for T-1, T-2 and T-3 months for the reason that past conflict, or a conflict-trap is often one of the main drivers for future conflict.

:l_viol_tot	COWCODE	country.name	latitude	...	population	hdi_index	bdist2	bdist3	nearest_country_dist	own_borders_dist	target	lag1	lag2	lag3
0	482.0	Central African Republic	5.75	...	8793.0	0.337	173.98240	173.98240	173.98240	173.98240	0	NaN	NaN	NaN
0	482.0	Central African Republic	4.75	...	31534.0	0.337	50.25557	50.25557	50.25557	50.25557	0	0.0	NaN	NaN
0	482.0	Central African Republic	4.75	...	31534.0	0.337	50.25557	50.25557	50.25557	50.25557	0	0.0	0.0	NaN
0	482.0	Central African Republic	4.75	...	31534.0	0.337	50.25557	50.25557	50.25557	50.25557	0	0.0	0.0	0.0
0	482.0	Central African Republic	4.75	...	31534.0	0.337	50.25557	50.25557	50.25557	50.25557	0	0.0	0.0	0.0

Figure 15 - Afrogrid merged with HDI, Prio and lagged Variables

The final data set was constructed by integrating the other two data sets in a stepwise manner, using the Afro-Grid data set as the foundational base (with data points from 1989 to 2020).

3.2 Setting Focus on Time Period

Due to the merging of diverse data sets with varying time spans, we encountered a challenge regarding missing values for specific periods. Each data set operates on its own scale in terms of time length, resulting in gaps and inconsistencies within the merged data. Consequently, certain periods lack sufficient data points, leading to missing values in our analysis. This issue necessitates careful consideration and appropriate handling techniques to ensure the accuracy and reliability of our findings. Addressing these missing values becomes crucial in order to maintain the integrity of our analysis and draw meaningful conclusions from the available data.

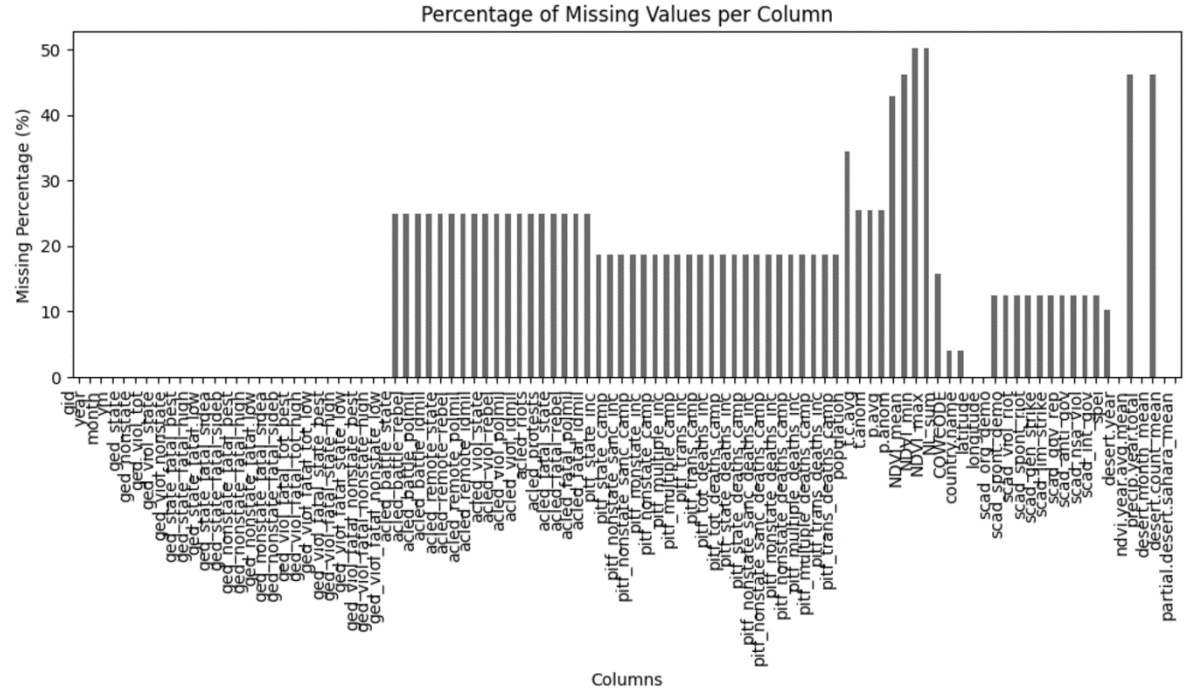

Figure 16 - Missing Values per Variable

The first analysis we ran showed that there are many missing values for certain databases. Therefore, it was decided to set up a first filter after the year 1998 to eliminate a huge proportion of missing values and simplify the data. Diving deeper, a pattern can be identified over time as shown below:

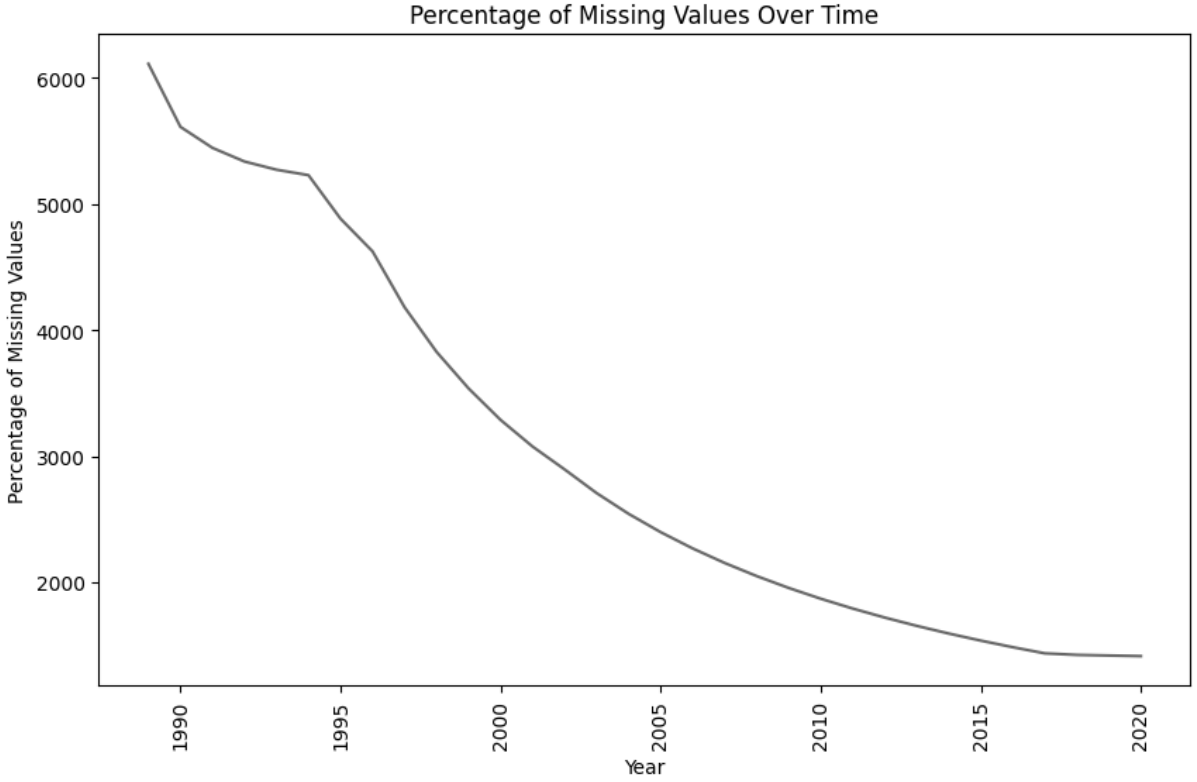

Figure 17 - Missing Values over Time

The number of missing values over time decreased meaning the more recent the case, the more

likely it is to find actual data on it. Therefore, to streamline model creation and improve efficiency, a period from *2003* onwards was chosen, in accordance with the UN. This focused approach allows us to analyze a more recent and relevant time frame, reducing complexity and optimizing computational resources. Therefore, at the end of the merging process and the filtering down to 2003 onward as a time period, a final dataset has been created consisting of 5239608 instances and 113 columns.

3.3 Further Handling of Missing Values

Finally, the data set still contained some missing values which needed to be imputed. The following graph shows the distribution of missing values in the final merged and filtered data set indicating the number of missing values per column per year:

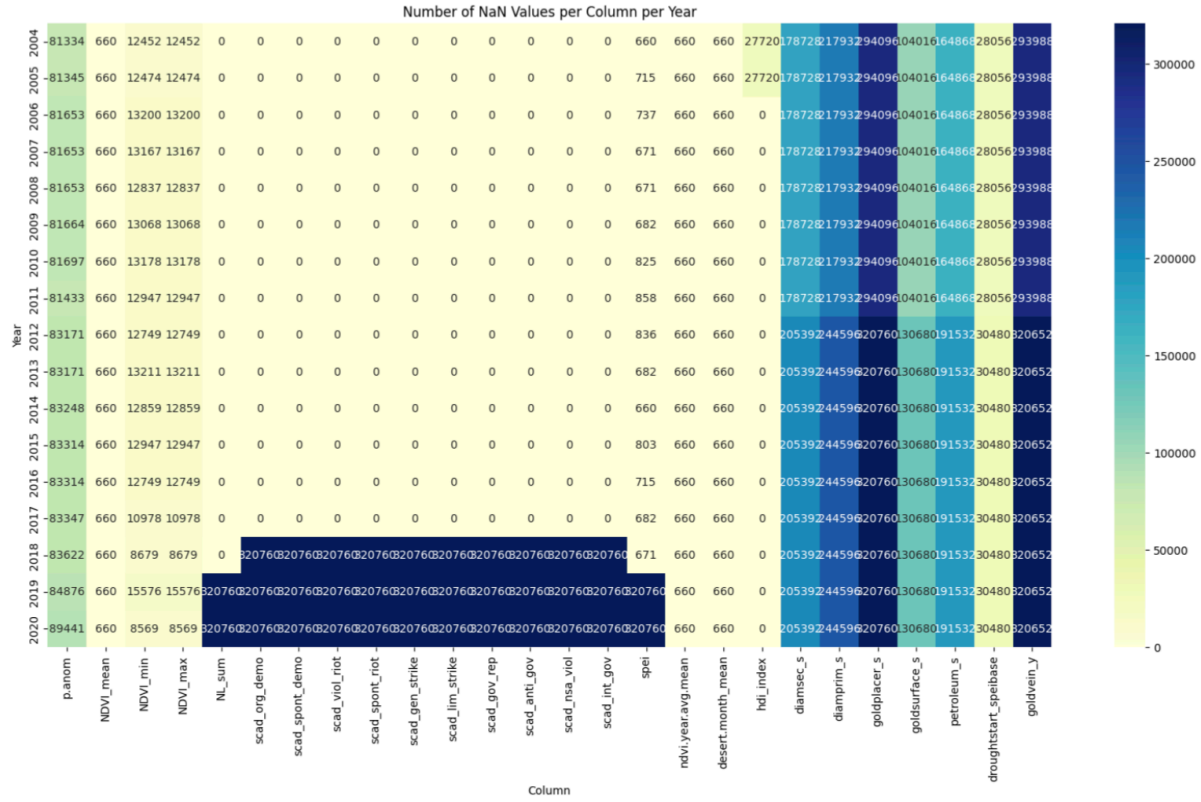

Figure 18 - Missing Values after Final Merging

Breaking down missing values as a percentage of the overall number of data points gives the following result (top 10 variables in terms of percentage of missing values):

	Variable	Missing Percentage
0	Goldplacer_s	100
1	Goldvein_y	99.96
2	Diamprim_s	75.28
3	Diamsec_c	62.57
4	Petroleum_s	58.07
5	Goldsurface_s	38.33
6	p.anom	26.91
7	Scad_int_gov	18.37
8	Scad_org_demo	18.37
9	Scad_spont_demo	18.37
10	Scad_viol_riot	18.37

Table 3 - Percentage of Missing Values after Final Merging

Having tried a couple of different imputation approaches, in order to avoid extreme noise in imputation we cleaned up variables superior to 30% and performed some simple imputing. For numerical values, we substituted for the mean, and for categorical variables, we chose the mode. By doing so, we follow literature review practices and ensure bias reduction and increase in statistical power.

3.4 Merging with PRIO-GRID distances to border and other countries

In the final stage of our data processing, an additional dynamic merging process incorporating the distances to other countries was undertaken. This data, which is complete and void of any missing values, can provide valuable insights and enhance the explainability of our models.

The variables introduced through this merge are:

- **bdist2**: This variable denotes the distance to the second nearest neighboring country or location. It reflects the spatial closeness or remoteness between the geographical area under study and its second nearest neighbor. By assessing these distances, we can gain insights into potential spatial influences or connections that are often crucial in geographical and social science research. The measurement units for these distances can vary, with the most common being kilometers or miles, depending on the original dataset.

- **bdist3**: Similarly, the 'bdist3' variable signifies the distance to the third nearest neighboring country or location. It indicates the spatial proximity or remoteness between the given area or region and its third nearest neighbor. By considering this additional layer of geographical context, we can further enhance our understanding of the spatial dynamics at play.

By incorporating these variables into our dataset, we can create more nuanced models that take into account the geographic and spatial aspects of our data. These new variables can contribute to a more comprehensive understanding of our research question, considering not just immediate geographic neighbors, but extending to the second and third nearest entities as well. Including these variables, we also manage to control for the fact that each grid is attributed to one country (the bigger part of the grid) only even though it might be shared by multiple countries at the same time.

3.5 Multicollinearity & Feature Distribution General Data Set

In our exploratory data analysis, we applied different visual techniques based on the nature of our variables. For discrete variables, count plots were used. This visualization method allows to observe the frequency of each category in the discrete variables, providing a clear understanding of the distribution and potential skewness in our data. For continuous variables, the chosen visualization form is histograms. A histogram allows to see the shape of the data distribution, which can indicate any skewness, kurtosis, or outliers present in our data. Overall, the data set seems balanced. Due to the sheer number of variables, the plots are not visualized here but can be found in the notebooks.

These visualizations guided us in planning the individual analyses for each country. Furthermore, if sufficient computational resources are available, these exploratory steps can help in running a more comprehensive model that incorporates all the countries in our dataset. By examining the data through these two different lenses, we can gain a deep understanding of our data, enabling us to make informed decisions during the subsequent analysis phases.

Next, we conducted a comprehensive analysis to test for multicollinearity among the predictor

variables. The process involved checking the correlation matrix to discern any potential issues of multicollinearity. This analysis was performed to ensure that our predictor variables were not highly correlated with each other, as such a correlation could compromise the integrity of our statistical model. By addressing multicollinearity, we aimed to improve the reliability and interpretability of the model coefficients. The results from the analysis helped us to refine ourthe model, ensuring its robustness and accuracy, and allowing us to derive more meaningful and valid conclusions from ourthe dataset. The following graphs depict the multicollinearity matrices for the general data set and specifically for the Central African Republic dataset (see next sections)..

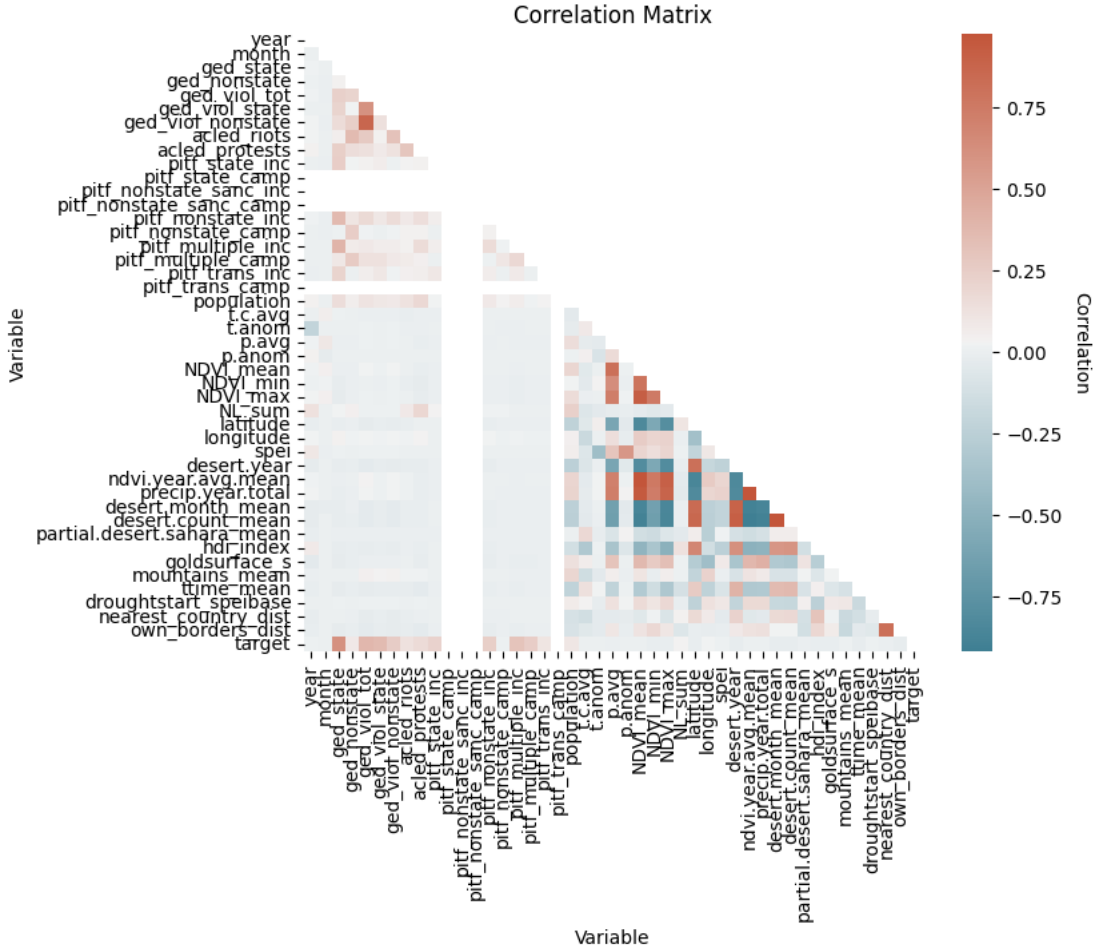

Figure 19 - Correlation Matrix General Data Set

Despite the multitude of variables in the dataset, the analysis indicates a relatively low degree of multicollinearity. In the subsequent analytical stages, the issue of multicollinearity is addressed individually in each notebook dedicated to a specific country. This approach allows for a nuanced examination of the unique correlations and dependencies that might exist among the variables within each country's dataset. By tackling multicollinearity at this granular level, a more accurate and reliable statistical analysis can be ensured, taking into account the specific contexts and characteristics of each country.

In the final phase of the data preparation process, the dataset was segmented into seven distinct subsets. Each subset corresponds to one of the countries where the United Nations currently has an active peacekeeping mission. This step was undertaken to tailor the analysis to the unique circumstances and data patterns of each country, acknowledging that each may have distinct influencers and dynamics. By analyzing these subsets individually, the specificity and relevance of the findings can be enhanced, yielding insights that are both nuanced and contextually grounded in the realities of each peacekeeping

mission. It is important to note that the code will create a folder called “countries” in the current notebook working directory where all seven counties datasets will be stored as csv files.

3.6 Multicollinearity in Central African Republic

As a next step, the individual data sets are checked, in this case using Central African Republic as a showcase. A low fidelity model was constructed as a preliminary step to assess the importance of the various variables. This model will not only provide a preliminary sense of how each predictor might contribute to the outcome but also assist in better identifying and eliminating multicollinear variables. By focusing on feature importance, it becomes easier to discern which variables might be creating unnecessary redundancy due to high collinearity. This approach will streamline the model, preserving only those variables that provide unique and valuable information, thereby enhancing the efficiency and interpretability of the eventual high fidelity model. Running a first linear regression (with quite good performance metrics, 66% R-squared) leads to the following feature importance distribution:

	Feature
1	Ged_viol_tot
2	Ged_viol_nonstate
3	Ged_viol_state
4	Nearest_country_sist
5	Own_borders_dist
	…
41	NDVI_mean
42	t.anom
43	NDVI_mean
44	latitude
45	Partial.desert.sahara_mean

Table 4 - Feature Importance of Initial Low Fidelity Model

Checking for it can be said that many of them are of little importance to the model performance and predictive power.

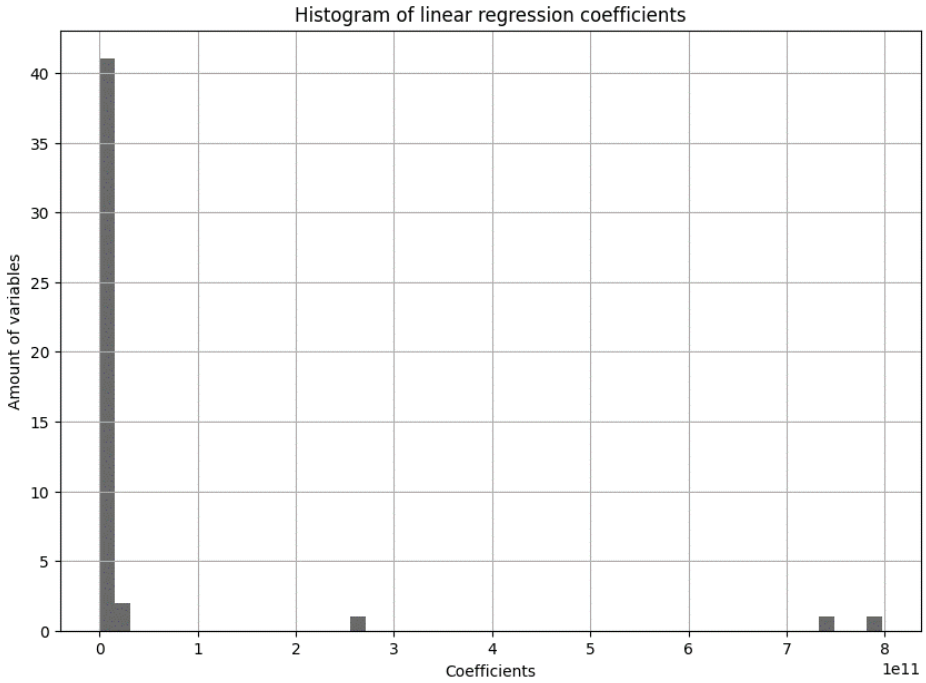

Figure 20 - Distribution of Coefficients in Intitial Low Fidelity Model

This results in 45 features, with 35 coefficients below our importance threshold. The following step is to pair the features to perform an actual correlation analysis leading to the following multicollinearity matrix:

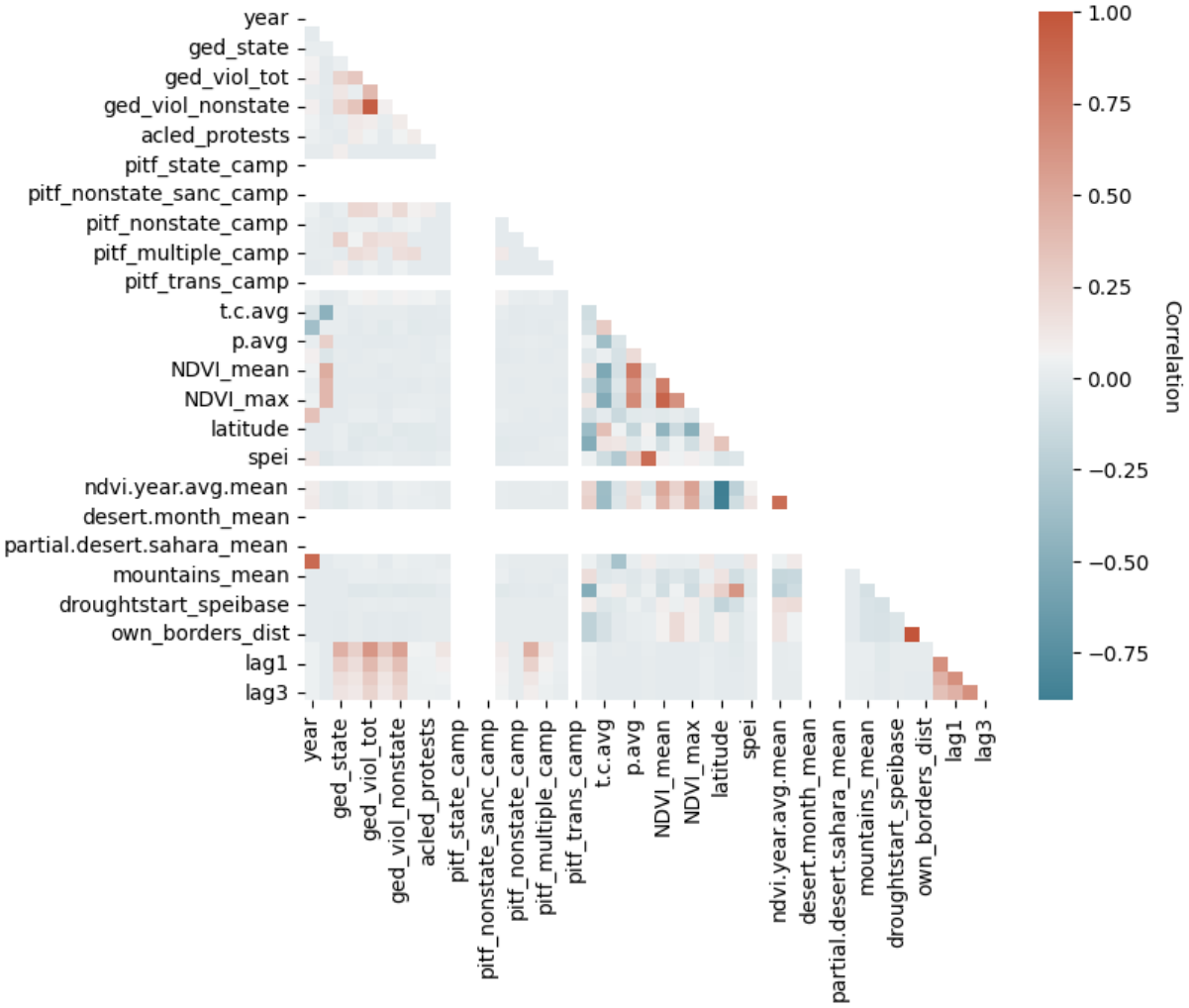

Figure 21 - Correlation Matrix Central African Republic

In a next step, the most associated pairs are listed for further analysis:

Ged_viol_tot	Ged_viol_nonstate	0.95
NVDI_mean	NVDI_max	0.94
latitude	ndvi.year.avg.mean	0.99
year	Hdi_index	0.87
p.anom	Ged_viol_fatal_nonstate_low	0.87
	…	
Precip.year.total	latitude	0.87
spei	p.anom	0.87
Ndvi.year.avg.mean	Precip.year.total	0.85
p.avg	NDVI_mean	0.78
NDVI_min	NDVI_mean	0.76

Table 5 - Associated Variables with Pearson Correlation above 75% for Central African Republic Data Set

The cutoff threshold chosen is a *Pearson* correlation of 75% which includes 13 variables. A s a final step, the two insights are combined and variables >0.75 of correlation and <1 of coefficient ar e dropped, which leaves the data set with 36 variables (11 are dropped).

This cautious approach might seem conservative, but it is an essential step in the data preparation process. It is geared towards preserving as much information as possible and maintaining the predictive power of the model. By removing less important yet highly collinear features, we prevent

these variables from artificially inflating the precision of our estimates or causing issues with the model's interpretability. As such, this approach ensures the robustness and reliability of our subsequent analyses, safeguarding the quality of our predictions and findings.

4 Model Creation and Interpretation

This chapter focuses on the creation of the actual machine learning model, including its implementation, execution, and evaluation metrics. Here, we delve into the practical aspects of running the model and measuring its performance using various metrics. The chapter aims to provide a comprehensive understanding of the model's creation process and its evaluation, allowing for a thorough assessment of its effectiveness in solving the given problem.

4.1 Model Development & Evaluation

In our analysis, we compared three different models: a Random Forest Classifier with grid search, a Gradient Boosted Tree model, and a Keras Sequential Neural Network. These models were evaluated based on their R-squared and Mean Squared Error (MSE) performance metrics. Let's examine the results and discuss the potential technical reasons behind the varying performance using the following visuals for the different performance metrics:

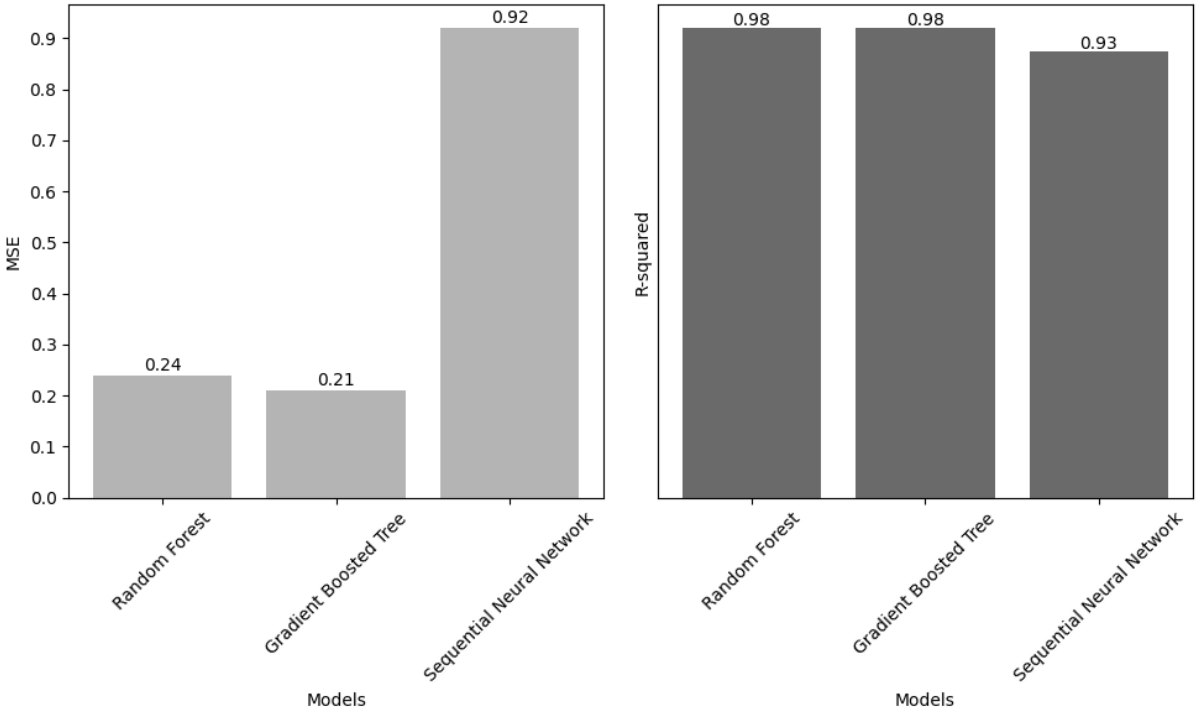

Figure 22 - Performance Metrics of Final Models

The Random Forest Classifier, including grid search, achieved an R-squared of 0.98 and an MSE of 0.24, indicating a relatively high level of explained variance and a moderate average prediction error. The Random Forest algorithm combines multiple decision trees, leveraging the power of ensemble learning and optimizing the predictions based on the collective decisions of the trees. The grid search technique helps fine-tune the model's parameters, ensuring the Random Forest Classifier can capture complex relationships and interactions in the data. The hyperparameters put are little, in order to optimize the already large processing time required. The best performing hyperparameters used for this model included the following, allowing in min_samples_split to better soften overfitting by wanting a minimum set of instances per split, and having a total of 100 decision trees with bagging of data and variables.

{'random_forest__max_depth': None, 'random_forest__max_features': 'sqrt', 'random_forest__min_samples_leaf': 1, 'random_forest__min_samples_split': 5, 'random_forest__n_estimators': 100}

The Gradient Boosted Tree model achieved an R-squared of 0.98 and an MSE of 0.21, outperforming both the Random Forest Classifier and the Neural Network. The Gradient Boosted Tree model sequentially adds decision trees, with each tree focusing on correcting the errors made by the previous trees (or weak learners). This iterative process allows the model to learn from its mistakes and make more accurate predictions. The Gradient Boosted Tree model's ability to capture complex interactions and handle non-linear relationships likely contributes to its superior performance.

In contrast, the Neural Network model achieved an R-squared of 0.93 and an MSE of 0.92, demonstrating lower predictive performance compared to the other models. Neural Networks consist of interconnected layers of artificial neurons, enabling them to learn intricate patterns and relationships in the data. The architecture design of the Neural Network, including the number of layers and neurons, plays a crucial role in its ability to learn and generalize from the data. In addition, Neural Networks typically require large amounts of diverse training data to effectively capture the underlying patterns. Insufficient training data or limited representation of the true data distribution could hinder the model's performance. The best performing hyperparameters used for this model included 3 hidden layers, with 64, 64 and 32 neurons (respectively), a ReLU activation function, and the Adam optimizer.

In summary, the Gradient Boosted Tree model demonstrated the best performance among the models considered, achieving the highest R-squared and the lowest MSE values. The Random Forest Classifier, including grid search, also performed well, showcasing its capability to handle complex data patterns. The Neural Network model showed a very good performance, especially in catching time correlations as it can be observed in multiple time variables being highly valued in feature importance terms.

4.2 Overfitting Check

To ensure that our models were not overfitting the training data, we implemented a train-test split methodology. This involved dividing the available data into two distinct sets: a training set used to train the model and a testing set used to evaluate its performance on unseen data. By comparing the model's performance on these two sets, we could assess its ability to generalize beyond the training data. The following graphs depict the results:

Random Forest

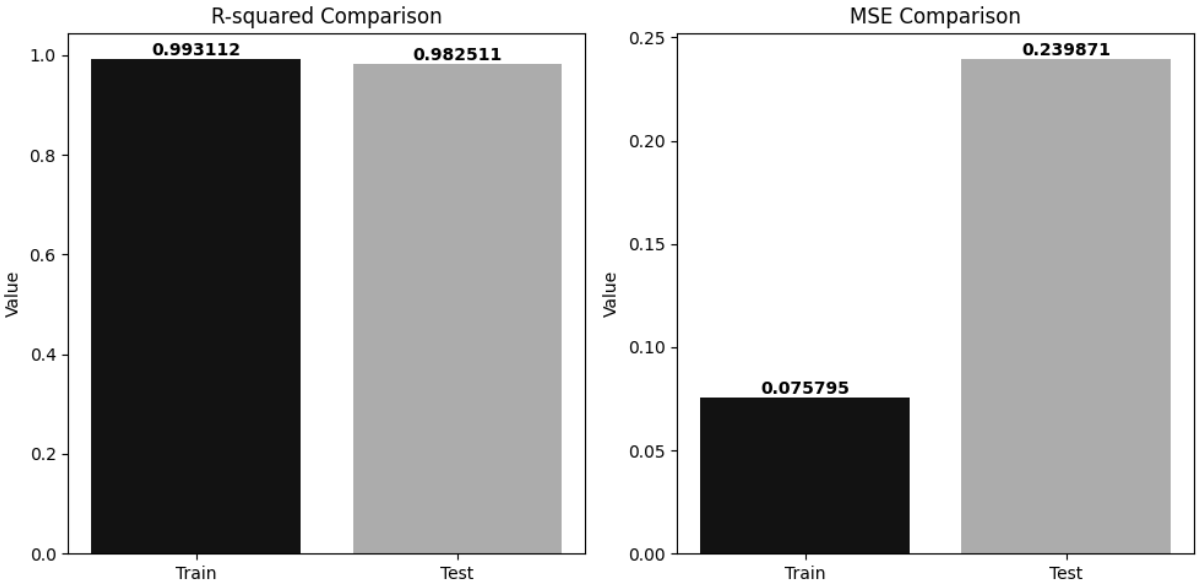

Figure 23 - Random Forest Train vs. Test Metrics

Gradient Boosted Tree

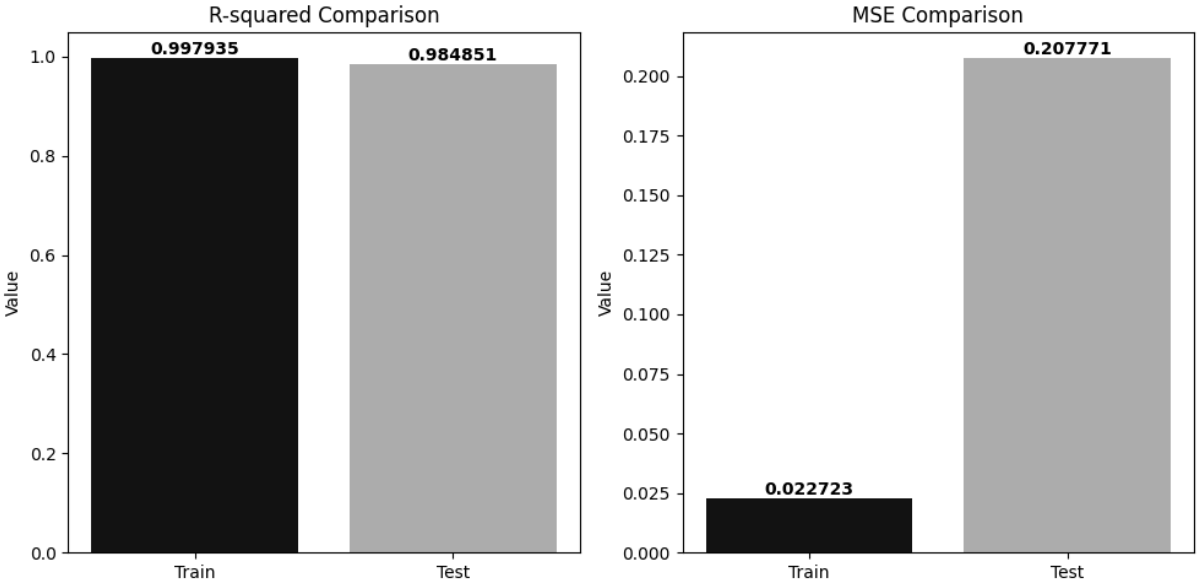

Figure 24 - Gradient Boosted Tree Train vs. Test Metrics

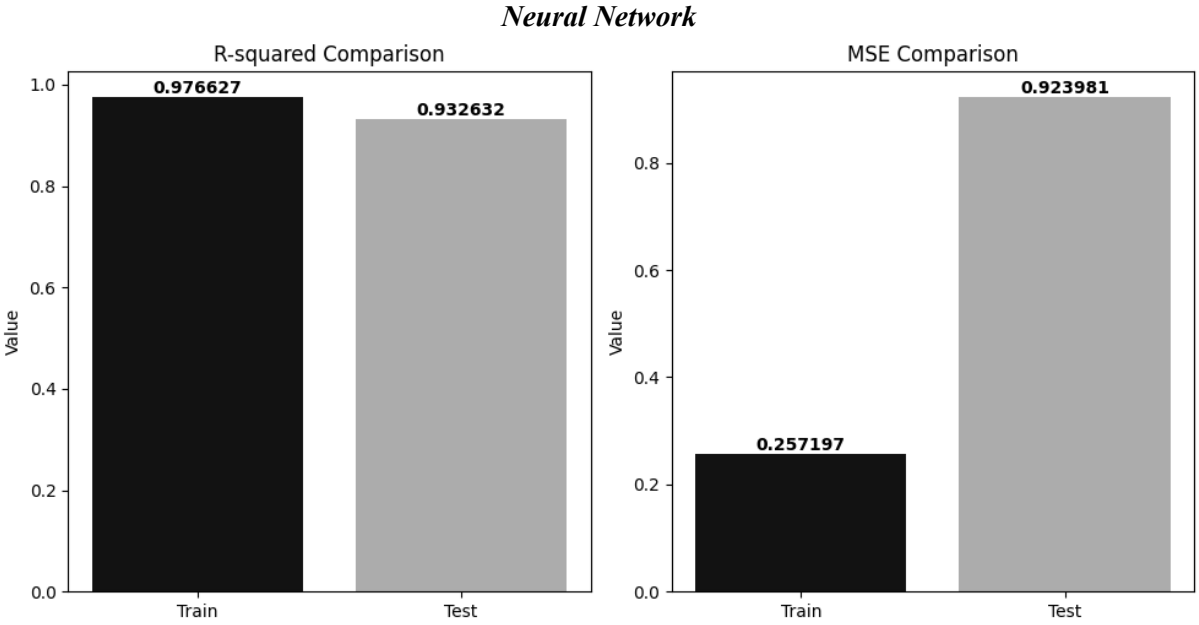

Figure 25 - Neural Network Train vs. Test Metrics

Collectively, all models demonstrate a commendable ability to generalize well, exhibiting solid performance even on the test data. While there is a marginally increased disparity in the Mean Squared Error (MSE) comparison between training and testing datasets, this difference remains within an acceptable range. Consequently, the models uphold their reliability and predictive power, affirming their utility in the applied context.

Focusing on our best-performing model, it can be said that the XGBoost model exhibited a very small difference in performance between the training set and the testing set. The difference in MSE wasis only 0.15, indicating that the model's predictions on the testing set were quite close to the actual conflict levels. This suggests that the model did not overfit the training data excessively, as it demonstrated a similar level of performance on unseen data. Nonetheless, similar things could be said about the other 2 models, with special attention to the Deep Neural Network, which surprisingly arrived to perform even better on unseen data than in the training data! This demonstrates checking for feature importance across all different models may provide a good view of what may be important from different perspectives, all of them reliable in predictability terms.

4.2 Results Interpretation

When performing regression analysis to predict conflict in a region and assist the UN in deploying resources, it is important to assess the accuracy and reliability of the regression model's predictions using key performance metrics. The main metrics commonly used in regression analysis are the Mean Squared Error (MSE) and R-squared (R^2). Which were both evaluated for the best performing model and taken as criteria for model selection.

- **MSE:** measures the average squared difference between the predicted values and the actual values. It provides an overall measure of the model's predictive accuracy, where lower values indicate better performance. In the context of predicting conflict levels, minimizing prediction errors is crucial to ensure accurate resource allocation.
- **R-squared:** represents the proportion of the variance in the dependent variable (i.e., conflict levels) that can be explained by the independent variables included in the regression model. Higher R-squared values indicate a better fit of the model to the data. This metric reflects the model's ability to capture and explain the patterns and trends in conflict levels, contributing to

its reliability in predicting and guiding resource deployment. [OBJ]

Based on the information provided, an XGBoost model demonstrated the best performance among the evaluated models. It achieved an MSE of 0.21, indicating a relatively low average prediction error. This suggests that the model's predictions were consistently close to the actual conflict levels observed in the region, being off by 0.21 deaths on average. Additionally, the XGBoost model obtained an R-squared value of ~ 0.98, indicating a high proportion of the variance in conflict levels can be explained by the independent variables in the model. This high R-squared value implies that the model provided a strong fit to the data, accurately capturing the factors influencing conflicts in the region.

Considering the overall performance of the XGBoost model, it demonstrates both accuracy and precision in predicting conflict levels. The low MSE signifies the model's ability to generate predictions close to the actual values, supporting effective resource allocation. The high R-squared indicates that a significant portion of the variation in conflict levels can be attributed to the predictors used in the model, enhancing its reliability.

While the XGBoost model shows promise, it is important to consider other factors such as the model's robustness, interpretability, and the theoretical relevance of the predictors employed. Continuous validation and monitoring of the model's performance are also necessary to ensure its ongoing effectiveness in predicting conflicts and guiding resource deployment decisions.

4.4 Feature Importance

Central to our entire analytical approach is understanding feature importance and interpretation. In the context of our project, crafting an accurate prediction model for conflicts is indeed crucial. However, the model's true value emerges from uncovering and interpreting the underlying drivers of conflict on a subnational scale.

While we aim for a highly accurate model, we also focus on decoding the multifaceted, localized factors that fuel these conflicts. These drivers, such as socio-economic discrepancies, political unrest, or resource constraints, serve as critical pieces of the larger conflict puzzle. By comprehending their roles, we gain a more nuanced understanding of conflicts.

The three different models run for the country of Central African Republic, naturally return slightly different results in terms of variable importance. However, certain patterns are visible in all three models.

Feature	Random Forest	Gradient Boosted Tree	Neural Network
1	Lag1	Ged_viol_tot	t.anom
2	Ged_viol_tot	Lag1	NL_sum
3	Ged_viol_nonstate	Ged_nonstate	Month
4	Ged_state	Ged_state	Population
5	Lag2	Ged_viol_state	Lag1
6	Lag3	longitude	Ttime.mean
7	Pitf_multiple_inc	t.c.avg	Mountains_mean

Table 6 - Feature Importance of Final Models

The first intriguing insight from our analysis is the striking similarity in feature importance as indicated by both the random forest and gradient-boosted tree models. This semblance may be due to the analogous nature of these two algorithms, as both essentially construct multiple decision trees and aggregate their outputs. On the other hand, deep learning operates on a markedly different paradigm, utilizing artificial neural networks and hence, showing distinct patterns in its treatment of feature importance.

Neural networks, despite their complexity in interpretation, excel in capturing time-sensitive data, particularly related to dates. This is attributed to their inherent nature and ability to recognize patterns in temporal information. The emphasis on data in neural networks aligns with their effectiveness in capturing time dynamics and identifying temporal relationships within the data.

On the other hand, the other two models demonstrate similar importance to the variables "Ged_viol_tot" and the lagged variables. The substantial emphasis placed on these variables indicates a higher likelihood of conflicts, especially if conflicts have occurred in the past. This importance aligns with findings from the conducted literature review, further reinforcing the significance of these variables in understanding and predicting conflict-related events.

Additionally, the variable "Ged_viol_state" clearly highlights a specific aspect of violence. This variable provides a focal point for further investigation and analysis, as it signifies an area of violence that warrants attention and deeper exploration. By directing focus towards this variable, insights into violent occurrences and their implications can be gained.

Understanding conflict dynamics is crucial for comprehending patterns of violence and implementing targeted interventions for peacebuilding and conflict resolution efforts. To further enhance our understanding, future studies should delve deeper into the importance of specific variables in our model, particularly in the context of high fatality predictions. These variables encompass a range of factors, including temperature, population density, past conflict (with a time delay), and even terrain structure. Investigating the role of these variables can provide valuable insights for policymakers, humanitarian organizations, and peacebuilders, enabling them to allocate resources effectively and implement targeted interventions to mitigate violence, protect vulnerable populations, support peace negotiations, and promote conflict resolution efforts. By considering these diverse factors, we can gain a more comprehensive understanding of conflict dynamics and inform evidence-based strategies for sustainable peace.

5 Predictions - Visualization

As shown our model has strong performance. However, as stated in the previous chapter accurate and easily usable information is crucial for effective decision-making and resource allocation. Although PRIO-GRID provides a consistent geographical reference system, its high granularity makes it hard to identify grids with trending or high conflict fatality counts. In this section, the model predictions for the selected geographic scope and prediction time window (<12 months) will be outlined using GIS-software visualization techniques.

5.1. Prediction Output & Layer Integration

In the methodology section, the extensive set-up of the selected QGIS software can be found. In summary, EPSG:4326 (WGS 84) is chosen as the coordinate reference system (CRS) since it accurately represents the Earth's shape and has the precise spatial representation of grids needed. To visualize the prediction values and associated characteristics for each grid cell, separate layers were added to the shapefile. The global layer, filtered for the African continent, was used. Prediction values and cell characteristics were sourced from CSV files. This approach enabled the representation of prediction values along with features like country, longitude, latitude, and distance to borders. By incorporating these layers, the spatial context of the prediction output was effectively captured. The centroids of each grid cell, based on their longitude and latitude data, were visualized to show the spatial distribution of predicted fatalities. This method facilitated the identification of areas with higher fatality rates and the assessment of monthly patterns.

In order to track changes over time, a sum vector was utilized. This vector was created by adding up the fatality counts from the preceding three months, offering a visual depiction of the progression from the initial month to the cumulative values of subsequent quarters.

In line with our adjustment for the geographical scope along the presence of the UN, we further presented our results for only the Central African Republic, although the model training, evaluation and prediction steps could be applied to each other country in the African continent as all data is readily available and the steps taken can be repeated. Also note that our predictions are only for a 12-month period ahead. As stated, previous studies found that predicting further ahead comes with significant increases in uncertainty.

Utilizing a gradient scale with a fixed interval size of 10 in QGIS for visualizing predictions per month offers several benefits. Firstly, it promotes consistency in the representation of data, making it easier to compare and interpret predictions across different time periods. The fixed interval size ensures that the color range remains constant, providing a standardized visual reference. Furthermore, this approach aids in quickly identifying areas of interest, such as conflicts with high fatalities. By assigning distinct colors to different prediction ranges, it becomes easier to pinpoint regions with higher fatality rates. Users can visually distinguish areas that require immediate attention or further analysis.

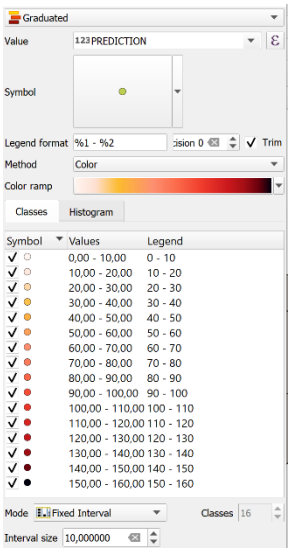

Figure 26 - QGIS – Gradient Scaling

As can be seen in the figure below in Central African Republic we find that in the first month for the data provided the grid will see fatalities of an estimated 151, a significantly high fatality count. Overall,

conflict in CAC are shown to be more in the western region. For the 12 month estimated, we see that although overall conflict fatalities are lower, some border conflict becomes apparent as well in the north-east of CAC.

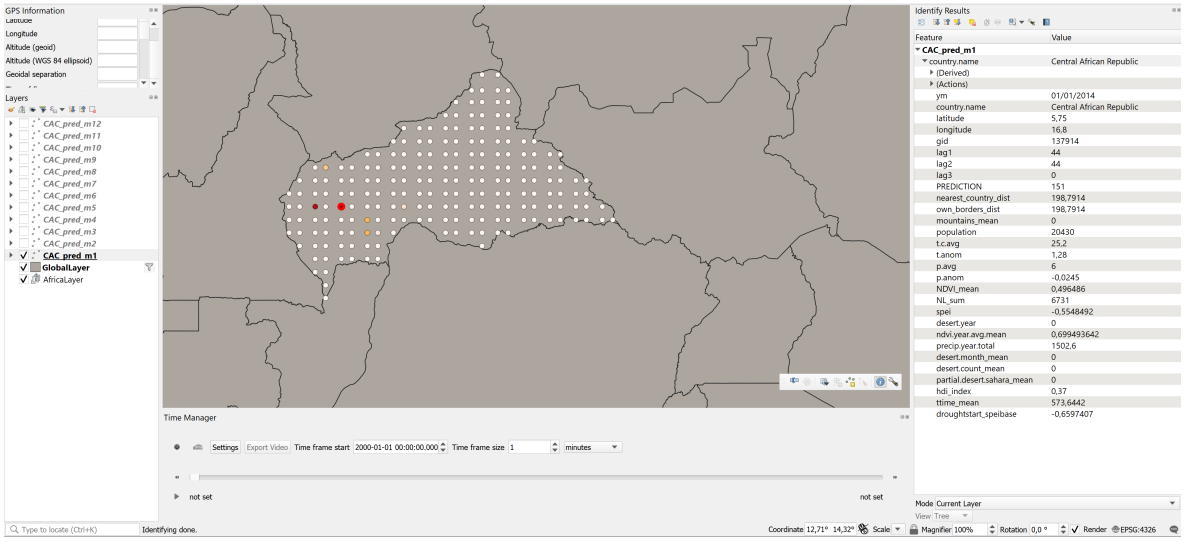

Figure 27 - QGIS – Prediction Output CAC 1

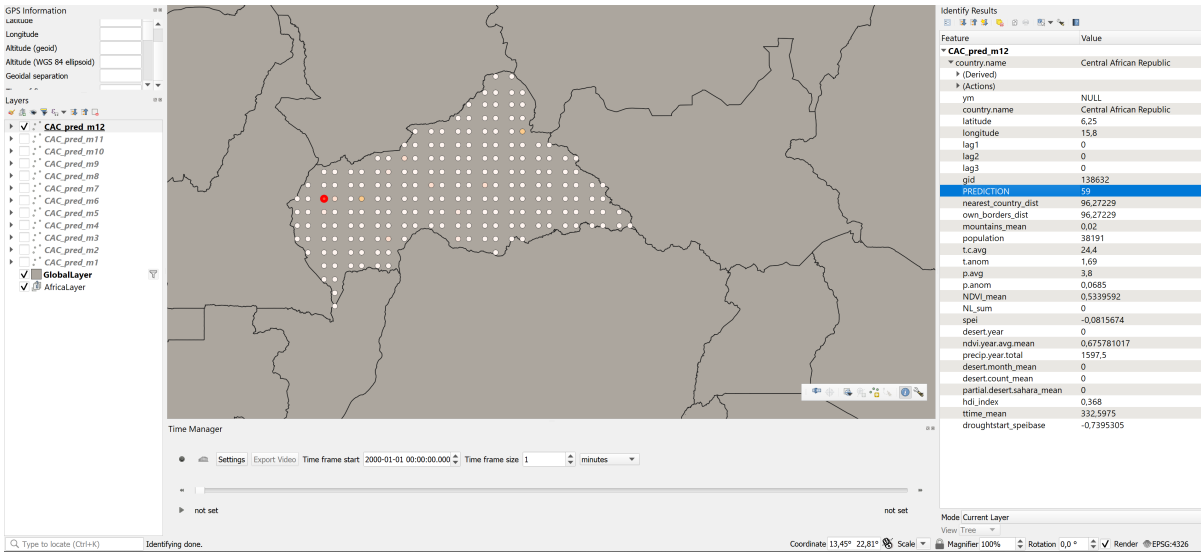

Figure 28 - QGIS – Prediction Output CAC 2

Figure shows prediction values for the next 12 months for the UN-based Central African Republic (CAC)

If we then look at the sum vector that accumulates the number of fatalities quarterly, several grids show strong accumulation. The "Time Manager" function in QGIS can further be utilized to animate and visualize these changes over time. This feature allowed for smooth transitions between monthly, quarterly, and yearly views, making it easier to identify patterns, trends, and potential resource needs or emerging hotspots at different time scales.

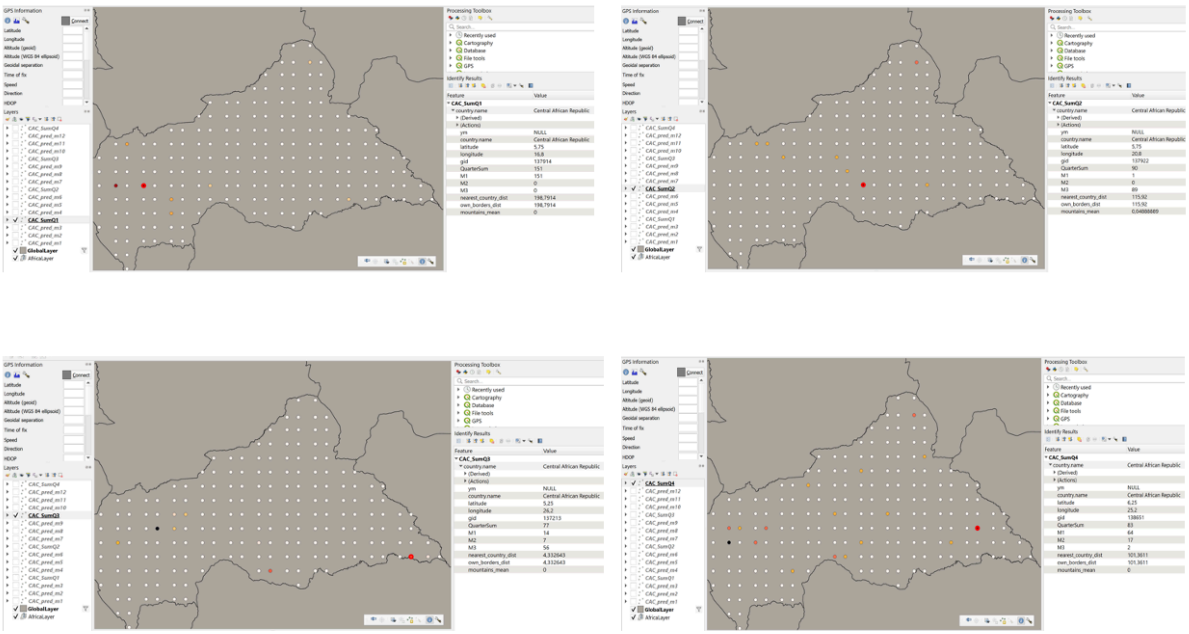

Figure 29 - QGIS – Quarterly Sum Prediction CAC

In our analysis, we observed that for the first quarter, there is a significant expectation of high fatalities only in the first month within a specific grid (#137914). However, when examining another grid in the 3rd and 4th quarters, we noticed a more continuous and prolonged estimation of fatalities. This indicates the need for a more dynamic analysis to identify conflict hotspots.

Conflict hotspots are regions or areas characterized by a concentrated or intense occurrence of conflicts. Our model and visualization tool have proven effective in efficiently identifying regions with high fatality rates and even predicting where hotspots may emerge in the coming months. This capability is valuable for developing early warning systems aimed at detecting and preventing potential escalations of violence in affected areas.

Understanding conflict hotspots is crucial for comprehending patterns of violence, analyzing conflict dynamics, and implementing targeted interventions for peacebuilding and conflict resolution efforts. To further enhance our understanding, future studies should delve deeper into the importance of specific variables in our model, particularly in the context of high fatality predictions. This analysis can provide valuable insights for policymakers, humanitarian organizations, and peacebuilders, enabling them to allocate resources effectively and implement targeted interventions to mitigate violence, protect vulnerable populations, support peace negotiations, and promote conflict resolution efforts.

6 Conclusion and Outlook

This final chapter serves as a summary of the main findings and takeaways of the project. Next to an outlook giving hints to potential further applications of our solution it also shows a personal reflection on the concepts learned and the project in general.

6.1 Project Results

The primary highlight and value addition of this project lies in the utilization of the Afrogrid dataset. Leveraging its granularity has significantly enhanced the prediction accuracy of the model, presenting a more detailed picture of the dynamics at play. This refined approach serves to yield richer, more precise insights. The adaptability of the model across different countries is another key advantage. Although the Central African Republic has been used as a demonstration case, the model can be effectively applied to any other country, particularly within Africa. This flexibility is facilitated by the comprehensive coverage provided by the Afro-Grid dataset.

Additionally, the models have demonstrated excellent generalizability and robustness against overfitting. This ensures their performance remains reliable and accurate across diverse scenarios and data, further affirming their applicability in a range of settings and for various research purposes.

6.2 Future Outlook

To provide continued value to the organization and excel in the capstone project, future students should consider implementing the following actions:

- *Expand to more Geographies*: To enhance the scope and impact of the ML project, students can extend their analysis to include additional countries within the UN's operating focus or even those outside. This expansion would provide a broader understanding of conflict dynamics across various regions, enabling the UN to make informed decisions on resource allocation and intervention strategies.
- *Add Additional Features*: While conducting the initial ML project, certain variables were excluded due to data limitations or other considerations. Students can further enrich the analysis by including these additional variables. This inclusion would provide a more comprehensive understanding of the factors influencing conflict and help uncover new insights and patterns.
- *Add Additional Databases*: To enhance the predictive power of the ML models, students can incorporate supplementary databases that provide relevant information. For instance, integrating stock prices, energy prices, or other public conflict databases can offer valuable contextual data that may contribute to more accurate conflict predictions. By leveraging these additional data sources, the models can capture a wider range of factors that influence conflict dynamics.
- *Integrate Notebooks via API*: To ensure seamless implementation and utilization of the ML project's findings, students can develop APIs to integrate their analysis notebooks with other UN systems or platforms. This integration would enable easy access to the ML models and allow decision-makers within the organization to make real-time use of the predictions and insights generated.
- *Multi-class Classification*: Expanding the ML project beyond predicting the number of conflicts, students can explore the classification of conflict types. By categorizing conflicts into different classes based on their characteristics or drivers, the UN can gain a deeper understanding of the underlying causes and tailor their interventions accordingly. Implementing multi-class classification models would provide a more comprehensive and nuanced view of conflict dynamics.
- *Big Hadoop Ecosystem*: To handle the increased scale of data and ensure efficient processing,

students can leverage the Big Hadoop ecosystem. This ecosystem comprises technologies such as Hadoop, Spark, and related tools that facilitate distributed computing and storage of large datasets. By utilizing these technologies, students can effectively manage the storage, processing, and analysis of extensive datasets, allowing for more robust and scalable ML models.

By considering these actions, future students can build upon the existing project for the UN, providing added value to the organization and delivering more comprehensive and impactful insights. These enhancements, including expanding the geographic scope, incorporating additional features and databases, integrating notebooks via API, exploring multi-class classification, and leveraging the Big Hadoop ecosystem, would contribute to a more holistic and powerful ML solution for conflict analysis and resource allocation within the UN.

6.3 Personal Learnings

Our project on conflict prediction has provided us with invaluable learnings that have far-reaching implications. Through our rigorous research and development process, we have gained a deep understanding of the complex nature of conflicts and the dynamics that drive them. This knowledge has equipped us with the skills to approach intricate problems, fostering innovative thought processes and highlighting the importance of dynamic strategies.

The primary takeaway from this project is that it offered an engaging opportunity to witness a real-world application of a machine learning model. The project not only imparted significant knowledge about the complex nature of conflicts but more importantly, it educated about the strategies to approach such intricate problems, illustrating the dynamism and innovative thought processes required. We have realized the power of data-driven decision-making. By harnessing historical data, socio-economic indicators, and relevant factors, we were able to develop a robust predictive model. This experience has emphasized the significance of leveraging data and analytics to inform strategic decision-making processes across various domains combining them for the greater good. Thereby, we have deepened our understanding of risk assessment and mitigation strategies. Through the analysis of patterns and identification of key risk factors, we gained valuable insights into the early indicators of conflicts. This knowledge enables us to proactively address potential conflicts, allocate resources effectively, and implement preventive measures to mitigate risks.

More importantly, our project has taught us the value of interdisciplinary collaboration and the importance of ethical considerations. Through collaborating with experts from the field, we have cultivated innovative thinking and a holistic approach to addressing complex societal challenges. We have explored the ethical implications of developing and deploying a predictive model in sensitive contexts, emphasizing the need for responsible and fair use of technology. A point of pride in this endeavor is the belief that the developed model holds the potential to contribute towards greater societal good. The hope is that this model, with its predictive capabilities, could be built upon further to potentially improve lives in less fortunate parts of the world. If it could contribute even in a small way towards saving or improving lives, that would signify the true success of this project. It is deeply gratifying to consider the potential impact this model could have on individual lives. The potential to positively affect lives underscores the immense value and profound personal fulfillment derived from this work.

Overall, our practical knowledge and understanding drive positive change through data-driven decision-making, effective risk assessment, and continuous model improvement. Thereby, we contribute to conflict prediction, peace-building, and the well-being of conflict-affected regions.

Bibliography

Amnesty International. (2021). Ethiopia: Nowhere is Safe for Civilians as Violent Tigray Conflict Spreads

Balestri, S. (2015). The Gold Standard: Does Gold Mining Contribute to Conflict in the Democratic Republic of the Congo? World Development, 66, 232-249. (Note: This reference is based on the author and title you provided. However, please ensure that you verify the accuracy and availability of the specific source.)

Beguería, S., Serrano-Notivoli, R., & Saz, M. A. (2010). SPEIbase: A global soil moisture database for drought analysis. Water Resources Research, 46(12)

Blattman, C., & Miguel, E. (2010). Civil war. Journal of Economic Literature, 48(1), 3-57.

Buhaug, H., & Rod, J. K. (2007). Local determinants of African civil wars, 1970-2001. Political Geography, 26(6), 715-735.

Central Intelligence Agency. (2021). The World Factbook: Central African Republic. Retrieved from https://www.cia.gov/the-world-factbook/countries/central-african-republic/

Collier, P., & Hoeffler, A. (2004). Greed and grievance in civil war. Oxford Economic Papers, 56(4), 563-595.

Deaths in conflicts. (2022). Our World in Data. https://ourworldindata.org/grapher/ucdp-deaths-in-conflicts-by-violence-type?facet=none

Fearon, J. D., & Laitin, D. D. (2003). Ethnicity, insurgency, and civil war. American Political Science Review, 97(1), 75-90.

Furlong, K. R., Gleditsch, N. P., & Hegre, H. (2006). Geographic Opportunity and Neomalthusian Willingness: Boundaries, Shared Rivers, and Conflict. International Studies Quarterly, 50(4), 733-756.

Gleditsch, N. P., Salehyan, I., & Schultz, K. A. (2008). Fighting at Home, Fighting Abroad: How Civil Wars Lead to International Disputes. Journal of Conflict Resolution, 52(4), 479-506.

Gleditsch, N. P., Wallensteen, P., Eriksson, M., Sollenberg, M., & Strand, H. (2002). Armed conflict 1946-2001: A new dataset. Journal of Peace Research, 39(5), 615-637.

Hegre, H., & Sambanis, N. (2006). Sensitivity analysis of empirical results on civil war onset. Journal of Conflict Resolution, 50(4), 508-535.

Hegre, H., Ellingsen, T., Gates, S., Gleditsch, N. P., & Tollefsen, A. F. (2013). Toward a democratic civil peace? Democracy, political change, and civil war, 1816-1992. American Political Science Review, 107(3), 387-406.

Hegre, H., Hultman, L., & Nygård, H. M. (2019). Simulating the effect of civil conflict on interstate disputes. Journal of Conflict Resolution, 63(5), 1129-1153.

Hsiang, S. M., Burke, M., & Miguel, E. (2013). Quantifying the influence of climate on human conflict. Science, 341(6151), 1235367.

Kaufman, R. R., Kraay, A., & Mastruzzi, M. (2006). Governance Matters V: Aggregate and Individual Governance Indicators, 1996-2005. World Bank Policy Research Working Paper No. 4012

Kim, J., & Conceicão, P. (2010). Conflict, Development, and Security. World Development, 38(9), 1231-1242.

Li, X., Zhou, Y., & Zhang, Y. (2020). A harmonized nighttime light dataset for detecting and modeling fine-scale variations in human settlements. Earth System Science Data, 12(4), 2455-2466.

Lujala, P. (2007). The Spoils of Nature: Armed Civil Conflict and Rebel Access to Natural Resources. Journal of Peace Research, 44(6), 611-629.

Lujala, P., Gleditsch, N. P., & Gilmore, E. (2005). A Diamond Curse? Civil War and a Lootable Resource. Journal of Conflict Resolution, 49(4), 538-562.

Murshed, S. M., & Gates, S. (2005). Spatial-horizontal inequality and the Maoist insurgency in Nepal. Review of Development Economics, 9(1), 121-134.

O'Loughlin, J., Linke, A. M., Witmer, F. D., Laing, A., Gettelman, A., Dudhia, J., ... & Tiedtke, M. (2012). Climate variability and conflict risk in East Africa, 1990–2009. Proceedings of the National Academy of Sciences, 109(45), 18344-18349.

Østby, G., Nordås, R., & Rød, J. K. (2009). Regional inequalities and civil conflict in sub-Saharan Africa. International Studies Quarterly, 53(2), 301-324.

PRIO-GRID. (n.d.). Retrieved from https://grid.prio.org/#/

Prunier, G. (2009). Africa's World War: Congo, the Rwandan Genocide, and the Making of a Continental Catastrophe. Oxford University Press

Raleigh, C., & Urdal, H. (2007). Climate change, environmental degradation, and armed conflict. Political Geography, 26(6), 674-694.

Richards, P. (1996). Fighting for the Rain Forest: War, Youth and Resources in Sierra Leone. Heinemann.

Ross, M. L. (2004). How Do Natural Resources Influence Civil War? Evidence from Thirteen Cases. International Organization, 58(1), 35-67.

Salehyan, I. (2012). Rebel fragmentation and civil war termination. American Journal of Political Science, 56(2), 526-540.

Schon, J., & Koren, O. (2022). Introducing AfroGrid, a unified framework for environmental conflict research in Africa. Scientific Data, 9(1). https://doi.org/10.1038/s41597-022-01198-5

Sundberg, R., & Melander, E. (2013). Introducing the UCDP Georeferenced Event Dataset. Journal of Peace Research, 50(4), 523-532.

Tollefsen, A. F., Håvard Hegre, H., Buhaug, H., Calvin, K. V., Davis, K., & Team, P. R. I. O. D. A. T. A. (2012). Armed conflict and climate change: Evidence from the PRIO-GRID. Climatic Change, 117(4), 647-663.

Toset, H. P., Gleditsch, N. P., & Hegre, H. (2000). Shared Rivers and Interstate Conflict. Political Geography, 19(8), 971-996.

U.S. Department of State. (2021). Democratic Republic of the Congo. In Country Reports on Human Rights Practices for 2020. Retrieved from https://www.state.gov/reports/2020-country-reports-on-human-rights-practices/democratic-republic-of-the-congo/

UCDP - Uppsala Conflict Data Program. (n.d.). https://ucdp.uu.se/exploratory

United Nations Security Council. (2021). Report of the Secretary-General on the situation in Mali

United-Nations (n.d.). Department of Operational Support.. https://operationalsupport.un.org/en

Zenn, J. (2014). Boko Haram's International Connections. Combating Terrorism Center at West Point