Daniel Racek

Modern Data Science in Conflict Research

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 26.05.2025

Daniel Racek

Modern Data Science in Conflict Research

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 26.05.2025

Erster Berichterstatter: Prof. Dr. Göran Kauermann (LMU München)
Zweiter Berichterstatter: Prof. Dr. Nils B. Weidmann (Universität Konstanz)
Dritter Berichterstatter: Prof. Dr. Xiaoxiang Zhu (TU München)

Tag der Disputation: 30.07.2025

Acknowledgments

I am extremely grateful to my main supervisor Göran Kauermann. You always supported me and guided me through this journey of my PhD. Whenever any issues arose, whether related to research or interpersonal matters, I could count on you. So, I want to say: thank you.

I also want to thank my second supervisor, Paul W. Thurner, who provided indispensable advice from a political science perspective, and always kept me in the loop on new publications and conferences relevant to our work. Furthermore, I would like to thank my third supervisor Xiao Xiang Zhu for her guidance, as well as her team at the Chair of Data Science in Earth Observation, for their valuable advice and support in working with and processing satellite imagery. I also want to thank Nils Weidmann for agreeing to serve as an external reviewer on the examination committee, Brittany Davidson for her advice from a social and general science perspective, and Brigitte Maxa for her help in handling all organizational matters.

I am also truly grateful to all the kind people at the institute, especially at our chair. Even though I wasn't always in the office, you still made me feel appreciated and welcome whenever I came in. A special mention goes to Cornelia Gruber, my office colleague, with whom I loved sharing life and work updates, and to Martje Rave, one of the kindest people I have ever met, for the frequent coffee chats.

Last but not least, I want to thank all my family and friends, both here in Munich as well as in Vienna, for their support and for always believing in me. In particular, I am grateful to Sabrina Pfeffer and Christian Scheucher for taking time out of their day to proofread my work. Finally, a special thanks goes to Thomas Salzer, for putting up with my constant messaging about my paper progress, his advice on visualizations, and reminding me to take time off work and spend it with my friends.

Summary

Driven by digital communication, social media, satellite technology, and the widespread digitization of information, the past few decades have seen a dramatic increase in the volume of data that is being produced and collected every day. As a result, data science, in the form of statistical modelling, machine learning and artificial intelligence, is playing an increasingly important role across both industry and academic research. In recent years, these developments have also begun to impact and transform the field of conflict research. This thesis contributes to this transformation by utilizing modern computational methods and novel data sources to improve the analysis, forecasting and understanding of armed conflict.

Part I of this thesis introduces conflict research and provides the broader context for the contributing articles. It outlines the field's main objectives and challenges, and the potential of data science in addressing them. Following an overview of well-established conflict event databases and best practices for working with them, the first part turns to novel data sources for the field. It first introduces satellite imagery and remote sensing variables, which are derived from the former, and explores their applications in conflict research. It then discusses social media as a data source, highlighting its opportunities and limitations. Next, it provides an introduction into statistical modelling, with a particular focus on generalized additive models (GAMs), which play an important role across all contributions. This is followed by an overview on predictive modelling in the context of conflict forecasting, covering the most widely used machine learning approaches in the field. Part I closes with a summary of the contributing articles and an outlook on the future direction of the field.

Part II of the thesis demonstrates how these novel data sources can be incorporated into both statistical and machine learning models for conflict. The first contribution shows how remote sensing datasets, such as landcover classifications and nighttime lights, can improve the forecasting performance of predictive models in conflict-ridden countries with limited official data sources. The second contribution employs freely available synthetic aperture radar (SAR) satellite images from the European Space Agency (ESA) to detect the destruction of buildings during war. Specifically, a technique called Interferometric SAR (InSAR) is used and combined with a non-parametric median regression and a robust statistical assessment to identify destruction and its timing at the building level. In the third contribution, the language used in tweets from Ukraine is analysed before and during the Russian invasion. Using generalized additive mixed models, the study disentangles sample effects, arising from the in- and outflux of users, from behavioural effects. It identifies a clear shift in language from Russian to Ukrainian with the outbreak of the war, primarily driven by behavioural changes of the users.

In the final part of the thesis, Part III, a statistical model is proposed to capture the diffusion effects of armed conflict across space and time. Specifically, the fourth contribution develops a generalized additive model with a flexible smoothing basis over past conflict, constructed from a set of exponential decay functions with varying decay rates. The model is able to capture the long-lasting and far-reaching spatio-temporal dependencies exhibited by conflict. Further analysis shows that conflict typically breaks out in densely populated areas and from there subsequently diffuses into less populated regions.

Zusammenfassung

Getrieben durch digitale Kommunikation, soziale Medien, Satellitentechnologie und die umfassende Digitalisierung von Informationen hat die Datenmenge, die täglich erzeugt und gesammelt wird, in den vergangenen Jahrzehnten dramatisch zugenommen. Infolgedessen spielt Data Science, in Form von statistischer Modellierung, maschinellem Lernen und künstlicher Intelligenz, eine zunehmend wichtige Rolle sowohl in der Industrie als auch in der akademischen Forschung. In den letzten Jahren haben diese Entwicklungen auch begonnen, die Konfliktforschung zu beeinflussen und zu verändern. Diese Dissertation trägt zu diesem Wandel bei, indem moderne statistische und datenwissenschaftliche Methoden sowie neuartige Datenquellen genutzt werden, um die Analyse, Vorhersage und das Verständnis zu bewaffneten Konflikten zu verbessern.

Teil I dieser Dissertation führt in die Konfliktforschung ein und liefert den übergeordneten Kontext für die Forschungsbeiträge. Es werden die zentralen Ziele und Herausforderungen des Forschungsfeldes skizziert sowie das Potenzial von Data Science zur Bewältigung dieser Herausforderungen aufgezeigt. Nach einem Überblick über etablierte Datenbanken für Konfliktereignisse und bewährte Praktiken im Umgang mit diesen, widmet sich der erste Teil neuartigen Datenquellen für die Konfliktforschung. Zunächst werden Satellitenbilder und daraus abgeleitete Fernerkundungsvariablen vorgestellt und ihre Anwendungsmöglichkeiten in der Konfliktforschung erläutert. Anschließend werden soziale Medien als Datenquelle thematisiert, wobei deren Potenziale und Grenzen diskutiert werden. Es folgt eine Einführung in die statistische Modellierung mit besonderem Fokus auf generalisierte additive Modelle (GAMs), die in allen Forschungsbeiträgen dieser Dissertation eine wichtige Rolle spielen. Im nachfolgenden Abschnitt wird ein Überblick über prädiktive Modellierung im Kontext der Konfliktvorhersage gegeben, einschließlich der am häufigsten eingesetzten maschinellen Lernverfahren in diesem Bereich. Teil I schließt mit einer Zusammenfassung der Forschungsbeiträge und einem Ausblick auf zukünftige Entwicklungen im Forschungsfeld ab.

Teil II der Arbeit zeigt, wie diese neuartigen Datenquellen in statistische und maschinelle Lernmodelle für Konflikte integriert werden können. Der erste Forschungsbeitrag veranschaulicht, wie Fernerkundungsdaten, bspw. Landbedeckungsklassen und nächtliche Lichtemissionen, die Prognosegüte von Vorhersagemodellen in konfliktreichen Ländern mit begrenzten amtlichen Daten verbessern können. Der zweite Beitrag verwendet frei verfügbare Synthetic Aperture Radar (SAR) Satellitenbilder der Europäischen Weltraumorganisation (ESA), um die Zerstörung von Gebäuden während Kriegen zu erkennen. Hierzu wird eine Technik namens interferometrisches SAR (InSAR) eingesetzt und mit einer nichtparametrischen Medianregression sowie einer robusten statistischen Evaluierung kombiniert, um Zerstörung und dessen Zeitpunkt auf Gebäudeebene zu identifizieren. Im dritten Forschungsbeitrag wird die Sprachnutzung in Tweets aus der Ukraine vor und während der russischen Invasion analysiert. Unter Verwendung generalisierter additiver Mischmodelle werden Stichprobeneffekte, verursacht durch das Zu- und Abwandern von Social-Media-Nutzer:innen, von Effekten durch Verhaltungsänderungen getrennt. Die Analyse zeigt einen klaren Wechsel von Russisch zu Ukrainisch mit Ausbruch des Krieges, der hauptsächlich auf Verhaltensänderungen zurückzuführen ist.

Im abschließenden Teil der Arbeit, Teil III, wird ein statistisches Modell vorgeschlagen, um die Diffusionseffekte von bewaffneten Konflikten über Raum und Zeit zu erfassen. Konkret entwickelt der vierte Beitrag ein generalisiertes additives Modell mit einer flexiblen Glättungsbasis über vergangenene Konflikte, die aus einer Vielzahl exponentieller Zerfallfunktionen mit unterschiedlichen Ab-

nahmefaktoren besteht. Das Modell kann die langfristigen und weitreichenden räumlich-zeitlichen Abhängigkeiten, die Konflikte aufweisen, adäquat erfassen und abbilden. Weitere Analysen zeigen, dass bewaffnete Konflikte typischerweise in dicht besiedelten Gebieten ausbrechen und sich von dort aus in weniger besiedelte Regionen ausbreiten.

Contents

Part I.

Introduction and background

1. Introduction

Armed political conflict is responsible for thousands of fatalities worldwide each month (Davies et al., 2024). Such conflicts can occur within countries, as in civil wars, or between nations, stemming from a range of political, economic, or ethnic tensions. In both cases, the consequences are severe and far-reaching (Gates et al., 2012). Armed conflict often forces people to migrate within and across borders, triggering refugee crises that can spill over into neighbouring and distant countries, thus destabilizing entire regions. It also disrupts local economies, undermines both domestic and international trade, and weakens governments' capacity to maintain order and provide essential services. As a result, countries caught in cycles of conflict often struggle to escape poverty (Collier et al., 2003).

Research on armed conflict generally pursues three main objectives. First, it seeks to quantify and understand the consequences of conflict, including its economic, political, and humanitarian impacts. Second, it aims to identify the key determinants and mechanisms that drive political violence. Third, a central goal is to develop early warning systems that help predict where and when violence is likely to break out or escalate. All of these can help policymakers and humanitarian organizations to prepare for conflict and design targeted interventions that may prevent or mitigate future conflicts (Hegre et al., 2019; Rohner, 2024).

Historically, conflict research has relied on manually compiled historical records and databases created by individual scholars. However, with the increasing globalization and digitization of information, conflict reporting has improved both in speed and level of detail. News organizations, social media platforms, and NGOs now provide information on conflicts in near real-time. This shift has led to dedicated research teams systematically collecting and curating conflict event data into globally spanning databases (Raleigh et al., 2010; Sundberg and Melander, 2013). Over the past decade, these developments have allowed the field to move away from country-level designs towards more fine-grained subnational analyses that were previously impossible. In addition, the regular and timely updates of these databases have facilitated the development of more sophisticated and accurate early warning systems (Rød et al., 2024).

As conflict research is increasingly moving to subnational levels, scholars are turning to novel data sources such as satellite imagery, remote sensing data, and social media as alternatives to official statistics, which are often unavailable at these levels in conflict-affected areas. An introduction into these data sources, alongside an overview on conflict event databases, will be provided in Chapter 2. The shift towards subnational analyses also requires more advanced modelling techniques. To obtain a better understanding of the determinants and mechanisms of conflict, statistical modelling approaches are essential. Hence, Chapter 3 introduces the key concepts of statistical modelling, with a specific focus on generalized additive models (GAMs), which play a key role in this dissertation. In contrast, forecasting models and early warning systems prioritize predictive performance, thus often relying on black-box machine learning models. The core concepts and most widely used approaches will be discussed in Chapter 4. Concluding remarks and an outlook on future research are provided in Chapter 5.

The remainder of this thesis consists of four contributing articles. Part II demonstrates how these new data sources can be incorporated into both statistical and machine learning models for the study of armed conflict, with the corresponding contributions in Chapter 6, 7 and 8. Part III, with the contributing article in Chapter 9, develops a novel statistical model specifically designed to capture and investigate the diffusion of armed conflict across space and time.

2. Data Sources for Conflict Research

"Without data you're just another person with an opinion."
— William E. Deming
(* 1939, † 1998)

Data is at the heart of any empirical study, not only in conflict research. It provides the empirical grounding on which analyses are built, and enables researchers to draw real-world conclusions substantiated by evidence found in the data. In the context of conflict research, having reliable and detailed information is essential for understanding where, when and how armed conflicts unfold, as well as their wider impacts on societies. Over the past decade, both the volume and variety of available data sources have increased substantially, allowing for more detailed and timely insights into conflict and its dynamics than ever before. This increase in data availability not only advances academic research, but may also help to inform policy decisions, humanitarian efforts, and peacekeeping missions.

This chapter is dedicated to introducing the primary data sources used throughout this thesis. It covers conflict event databases, which have been one of the key advancements for conducting subnational conflict studies. For a long time, researchers had only access to country-level data and events, which severely limited empirical analyses and research (Eck, 2012). This drastically changed roughly a decade ago with the introduction of conflict event databases, which systematically report conflict events across the world on a fine-grained subnational level (Raleigh et al., 2010). With their introduction and the accompanied shift of research to subnational levels, scholars have recently started to explore alternative data sources that can supplement or replace official government statistics, which are often only available at the national level and unreliable in conflict-affected countries. In this context, this chapter introduces both satellite and social media data as "novel" data sources, that have emerged over the past decade, and are beginning to find a foothold in the field. Satellite images provide snapshots of on-the-ground conditions and are available globally, independent of whether a country is experiencing conflict or not. Meanwhile, social media platforms capture public sentiment and discourse in real time, providing a valuable source and continuous stream of information on unfolding events, without requiring traditional surveys or interviews, which are often difficult or impossible to conduct.

The remainder of this chapter is organized as follows. Section 2.1 introduces conflict event databases, which form the backbone of subnational conflict research. Section 2.2 then discusses satellite imagery and remote sensing datasets, along with their applications in conflict research. Finally, Section 2.3 explores social media data and its use cases.

2.1. Conflict Databases

Conflict event databases are essential for studying political violence, as they provide access to structured and systematically collected datasets on events of armed conflict on a subnational level. A conflict event typically refers to an individual occurrence of political violence, such as a battle, an attack on civilians, or a bombing. Each event is assigned an approximate date and location, a classification of the violence type (e.g., "battle" or "violence against civilians"), and an estimated number of fatalities. Depending on the dataset, these core attributes are often supplemented by additional details, such as information on the actors involved, more granular event categories, high and low estimates for the fatalities, time and spatial precision codes, and source references.

Among the many available databases, the Uppsala Conflict Data Program (UCDP) Georeferenced Event Dataset (GED) (Sundberg and Melander, 2013) and the Armed Conflict Location & Event Data Project (ACLED) (Raleigh et al., 2010) are the most widely used datasets in the field due to their global coverage and systematic data collection efforts. To collect up-to-date information on conflict events, both rely on media monitoring, reports from international organizations, and non-governmental organizations (NGOs), many of which operate at the local level. ACLED additionally incorporates information from trusted social media accounts for their coverage. Both UCDP GED and ACLED are publicly available, allowing researchers to freely download their data or access it via their respective APIs. The two datasets and their history are discussed in more detail next.

2.1.1. Uppsala Conflict Data Program Georeferenced Event Dataset

The Uppsala Conflict Data Program (UCDP) Georeferenced Event Dataset (GED) was officially introduced in 2013, initially covering only Africa before expanding to a global coverage. Maintained by the Uppsala Conflict Data Program, the dataset is updated annually and provides event-level data on armed conflict dating back to 1989. UCDP GED systematically records instances of armed force involving organized actors, such as governments and rebel groups, provided the event results in at least one direct (estimated) fatality. To align with UCDP's definition of armed conflict, events in which at least one actor does not surpass a total of 25 fatalities within a calendar year are excluded from the dataset.

In 2020, UCDP introduced a candidate version of their dataset (Hegre et al., 2020), which provides preliminary monthly event data with a delay of roughly 1.5 months and which does not apply the annual 25-fatality threshold. Most of these candidate events are later incorporated into the annual release, which is usually published approximately 18 months after the end of the calendar year. Hence, the candidate dataset offers a much more timely version of the GED for time-sensitive studies and analyses.

2.1.2. Armed Conflict Location & Event Data Project

The Armed Conflict Location & Event Data Project (ACLED) was officially launched in 2010, initially covering 50 "unstable" countries before expanding globally. Unlike UCDP GED, ACLED

does not apply fatality thresholds in its reporting on political violence, and includes events involving unknown or generic actors (e.g., "unidentified armed men") (Eck, 2012), thus also capturing incidents of unorganized or semi-organized activities. Furthermore, ACLED records events of riots and protests, even when they are non-violent, offering an additional source of information for studying political unrest. As of the time of writing, ACLED is updated on a weekly basis, making it the most up-to-date, manually curated and globally-spanning conflict event dataset available.

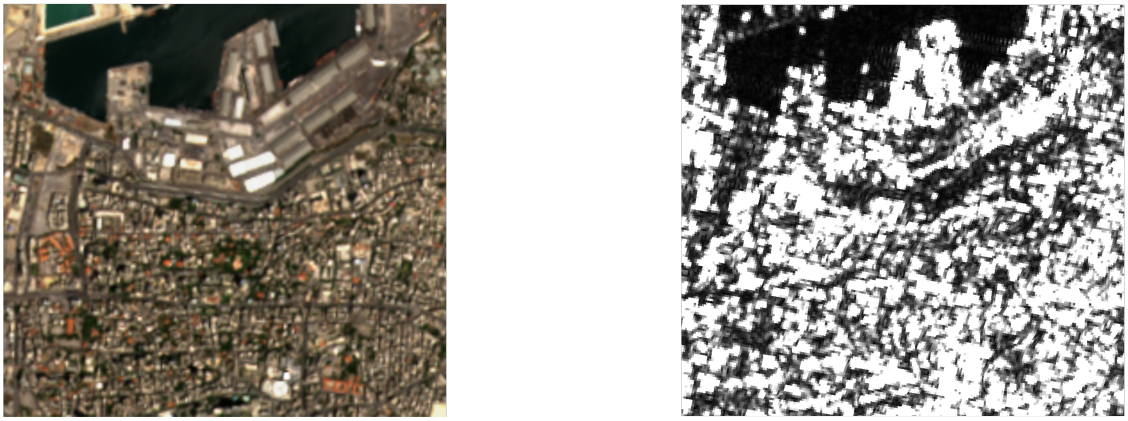
2.1.3. Precision Levels & Units of Analysis

Independent of the dataset, the precision of recorded conflict events varies substantially. Both UCDP GED and ACLED assign spatial and temporal precision codes to indicate the accuracy of event locations and dates. These precision levels depend on the availability and reliability of the original sources used to document each event. On the spatial level, each event is assigned geographic coordinates. These can be highly precise, such as a specific town, but in many cases are not. When precise coordinates are unavailable, events are typically geocoded to the nearest town, the capital of the region, or the closest natural location (e.g., a border area or forest), hence introducing spatial uncertainty. On the temporal level, the exact date of an event is not always known. Some events are reported with a precise day, while others are only known to have occurred within a certain week or month. In UCDP GED, events may even be recorded as having taken place at some point within an entire year. The exact coding strategies vary between the datasets, and thus they need to be carefully accounted for to avoid any misinterpretations. Detailed documentation is provided in each dataset's respective codebook.

Due to these lower levels of reporting precision for many events, in conflict research, events and/or fatalities are generally aggregated into larger spatial and temporal units of analysis. Common approaches are to study and forecast conflict at the level of grid cells or administrative regions, large enough to mitigate biases potentially arising from the spatial imprecision. It is important to note that the varying sizes of administrative zones within and across countries, as well as their irregularity in shape, complicates spatial inference. This issue is further discussed in the contribution in Chapter 6. In a similar fashion, events are also aggregated temporally, typically to monthly or yearly counts. These aggregation strategies inherently involve a trade-off between introducing biases and obtaining more fine-grained insights. Biases can arise from incorrectly assigning events to specific spatio-temporal units of analysis. For example, an event may be assigned to a grid cell covering the capital, even though its spatial precision code suggests it could have also taken place in one of the surrounding cells. This trade-off is further explored and discussed in Cook and Weidmann (2022).

2.2. Satellite Data

Satellite imagery captures the Earth's surface and is taken by satellites, orbiting the planet, operated by both governments and businesses. Over the past decades, advancements in satellite technology have substantially improved image quality, temporal availability, and affordability, and thus satellite images are an increasingly powerful tool in many fields, including research on armed conflict. Satellite images vary in resolution, frequency and spectral coverage. They can broadly be categorized into two main types: optical imagery and radar-based imagery.



(a) Optical image from Sentinel-2A.

(b) SAR image from Sentinel-1.

Figure 2.1.: Publicly available satellite images from the Sentinel program of the European Space Agency (ESA) of the same location. (a) is an optical image with 10m resolution from Sentinel-2A, using spectral bands 4, 3 and 2. (b) is a multi-looked SAR image with 20m resolution from Sentinel-1. For visualization of the two-dimensional image of complex samples with a real and imaginary part, a γ_0 greyscale visualization of the VV polarization is used.

Optical Satellite Images Optical images are captured through passive sensors that detect reflected sunlight from the Earth in multiple spectral bands, both in the visible and near-visible part of the electromagnetic spectrum. Optical images are widely used for applications such as land cover classification, vegetation and agricultural monitoring, as well as urban mapping. However, since optical sensors rely on sunlight, they can only capture images during the day. Furthermore, cloud cover and shadows can obstruct the view of the ground, thus often requiring multiple satellite passes to obtain a clear image of a given area. An excellent introduction to optical satellite imagery is provided by Enright (2022), survey articles are available from Cheng and Han (2016) for object detection and Gómez et al. (2016) for landcover classification.

Synthetic Aperture Radar (SAR) Satellite Images SAR images are captured by active sensors that emit microwave signals and then record the backscattered signals. Due to the movement of the satellite, signals are sent out and received from different sensor positions, allowing for two- and three-dimensional reconstructions of objects. SAR imagery is particularly valuable for surface mapping, forestry applications (see e.g., Kugler et al., 2015), and monitoring urban infrastructure. Due to microwaves' longer wavelengths, SAR images are not affected by cloud coverage and only minimally by weather conditions. Furthermore, since these active sensors do not rely on sunlight, images can be captured both during the day and at night. A tutorial paper for SAR imagery and its processing techniques is for example provided by Moreira et al. (2013).

Many public satellite programs provide free access to their imagery. For example, the European Space Agency (ESA) offers open access to data from all of its Sentinel satellite missions, including for commercial usage (ESA, 2025). A comparison of an optical and a SAR image from the Sentinel program is shown in Figure 2.1. NASA's Landsat program (Wulder et al., 2022) is also providing free access to their imagery since 2008 (USGS, 2018). Notably, these publicly available images are only of medium resolution, ranging from 10m to 60m per pixel, which can limit the

ability to capture fine-grained details. In contrast, commercial providers such as Maxar and Planet offer high-resolution imagery with spatial resolutions as fine as 20cm, enabling much more detailed analyses. However, these high-resolution images are proprietary and often costly to obtain (Planet Labs, 2025), making widespread usage unfeasible for both researchers and humanitarian organizations.

Generally, working with raw satellite imagery requires considerable technical expertise in both remote sensing and image processing. Consequently, instead, researchers often rely on already processed remote sensing datasets, derived from satellite imagery, discussed next. However, recent advancements in foundational earth observation models, pre-trained in an unsupervised manner on millions of satellite images across various modalities, are reducing this technical entry barrier (Cong et al., 2022; Jakubik et al., 2025). These models can often be used out-of-the-box, enabling researchers to directly extract meaningful insights from satellite imagery.

2.2.1. Remote Sensing Datasets

Remote sensing datasets provide large-scale structured features about the Earth's surface. They are generated by extracting relevant information from satellite imagery using trained models and algorithms, sometimes combined with additional geospatial data sources such as census records. These datasets enable downstream analyses to use already pre-processed features and thus eliminate the need for researchers to work directly with raw satellite images, which can be complex and challenging to process.

Commonly used freely available remote sensing datasets include but are not limited to:

- **Land Cover:** Datasets such as Copernicus' Global Landcover Map (Buchhorn et al., 2020) and NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover dataset (Friedl and Sulla-Menashe, 2022) classify the Earth's surface into categories such as urban areas, forests, cropland, and water bodies.
- **Topography and Elevation:** Detailed information about topography and elevation, including measures of terrain ruggedness and slope, are for example available globally from Amatulli et al. (2018).
- **Population Estimates:** Products like WorldPop (Tatem, 2017) and Meta's High-Resolution Settlement Layer (HRSL) (Tiecke et al., 2017) estimate population densities using satellite imagery combined with other geospatial data sources.
- **Nighttime Lights:** Datasets such as the VIIRS Nighttime Lights (Elvidge et al., 2021) capture nighttime illumination and are often used as a proxy for economic activity and infrastructure development.
- **Vegetation Monitoring:** Indices such as the Vegetation Health Index (VHI) and the Vegetation Condition Index (VCI), for example available from the Food and Agriculture Organization (FAO) of the UN (FAO, 2025), track vegetation health over time. These datasets are often used for assessing drought conditions and monitoring agriculture.
- **Temperature and Precipitation:** Satellite images also allow to estimate both surface temperature and rainfall. Remote sensing datasets on precipitation are for example available from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) (Funk

et al., 2015), whereas both day- and nighttime temperature recordings can be obtained from NASA's MODIS Land Surface Temperature and Emissivity dataset (Wan et al., 2021).

These datasets are usually available in raster format, which is a grid of equally sized cells, all containing values for one or more remote sensing variables of the area they are covering. An overview about the various spatial data types and introduction into spatial data manipulation with *R* is for example freely available (online) from Pebesma and Bivand (2023).

2.2.2. Satellite Data in Conflict Research

Satellite images and remote sensing datasets can be extremely valuable for research on armed conflict. Many conflict-affected regions lack reliable official statistics on, for example, population, urban infrastructure, economic conditions, or agricultural activity, as long-term violence disrupts data collection and often prevents government agencies from conducting censuses and/or maintaining records about their economic activity and population. In these settings, remote sensing data offers a potential alternative. Its global coverage makes it possible to draw on up-to-date information even when no other data sources are available. The contribution in Chapter 6 demonstrates that remote sensing datasets can substantially improve the predictive performance of both machine learning and statistical models forecasting conflict during the Syrian civil war from 2011 to 2020. Syria presents itself as an ideal use case for evaluation, as the only other available data source are demographics from the 2004 census. In particular, the contribution showcases that remote sensing data leads to more accurate predictions of conflict onset.

Satellite images on the other hand can help to facilitate near real-time monitoring of on-the-ground conditions during conflict, enabling the detection of, for example, building destruction (Mueller et al., 2021), population displacement (Rufener et al., 2024), and disruptions to critical infrastructure such as bridges, power plants, or hospitals. The contribution in Chapter 7 showcases how freely available SAR images from the ESA can be used to detect the destruction of individual buildings in conflict zones. Combining interferometric SAR (InSAR) (Bamler and Hartl, 1998) with a robust statistical assessment of each building pixel's stability over time, the approach is able to correctly identify destroyed buildings in near real-time.

2.3. Social Media

In the past decade, social media platforms, such as Twitter (now known as X), Facebook, Reddit and Instagram, have become an important source of real-time information and public discourse. Unlike traditional (or sometimes also "old") media, social media content is primarily user-generated. This can provide additional insights into both past and ongoing events, not (yet) reported by newspapers or other agencies. Social media data offers a constant stream of information that researchers can analyse to detect trending topics, shifts in sentiment, and more generally, behavioural patterns. Depending on the platform, the available data can include text, images, videos, geolocations, timestamps, social interactions (likes, shares, reposts), network relationships, and user metadata. This breadth of data allows researchers to study not only what people say, but also how information spreads, how communities form and evolve, and how individuals and groups engage with specific topics over time. An influential study that showcases how social media data

can be used for research is the work by Vosoughi et al. (2018), who investigated the spread of true and false news online.

Working with social media also presents several challenges. The large volume of user-generated content can be difficult to store and process, and extracting relevant pieces of information can be challenging. Moreover, it is important to keep in mind that social media users are not representative of the general population, leading to potential selection biases in analyses relying on social media posts (see, e.g., Ruths and Pfeffer, 2014). In recent years, bots have played an increasingly prominent role on these platforms (Geissler et al., 2023). They can distort measures of user activity, amplify certain messages or narratives, and complicate analyses of public opinion, and therefore need to be carefully accounted for. Social media platforms have also started to impose strict limits on their APIs, making data collection for large-scale studies substantially more difficult, and in some cases even impossible (Davidson et al., 2023). For example, Twitter (now X), closed its official research API in 2023, which had granted academic researchers free access to millions of tweets per month. Broad access to the API now costs several thousand euros per month (X, 2025). Finally, researchers need to be aware of ethical considerations when working with social media data, including questions of user privacy, informed consent, and the responsible handling of potentially sensitive information.

2.3.1. Social Media in Conflict Research

In the context of conflict research, social media offers opportunities to monitor public discourse and behaviour during war or conflict in near real time, without the need for resource-intensive surveys on the ground, which may be dangerous or impossible to conduct. Natural language processing (NLP) techniques can be used to automatically identify language, topics, and sentiment, sometimes even at a granular geographic level through geotagged content, enabling researchers to efficiently process large volumes of data for further analysis. The contribution in Chapter 8 demonstrates how millions of tweets can be processed and analysed to track the language use of Ukrainian users before and during the Russian invasion of Ukraine. Using a form of generalized additive models (GAMs; introduced in Chapter 3) to separate sample effects, arising from the in- and outflux of users on the platform, from behavioural effects, the analysis identifies a clear shift in language use from Russian to Ukrainian. This change is mainly attributable to behavioural shifts, hence can be interpreted as users' conscious choice towards a more Ukrainian identity as a result of the Russian invasion.

Social media data can also be used to draw on information about conflict events themselves. As discussed earlier, for example, ACLED has started incorporating reports from trusted social media accounts into its conflict event collection and coding process. Notably, more automated solutions are also starting to be developed, to utilize social media at scale. For instance, Scholz et al. (2025), develop a transparent classifier that distinguishes protest images from non-protest images on social media, while Sobolev et al. (2020) use social media posts to estimate the size of protests.

Finally, social media data may also help to improve the forecasting performance of early warning systems. As discussed later in Chapter 4, forecasting conflict is particularly challenging, and while many data sources are available, only a few contain signals that actually improve predictive accuracy. In this aspect, near real-time shifts in social media posts, such as rising hostility, hate speech, or a spike in activity in contested areas, seem promising to help identify regions at risk for

political violence. However, due to the aforementioned challenges with social media data, at the time of writing, research in this area is still limited to exploratory studies (Zeitzoff, 2017; Dowd et al., 2020).

3. Statistical Modelling

> "The statistician cannot evade the responsibility for understanding the process he applies or recommends."
> — Ronald Fisher
> (* 1890, † 1962)

Statistical modelling plays an essential role in identifying patterns in data and making informed inferences about the underlying processes that generate them. At its core, it provides a framework for reasoning under uncertainty, and a way to quantify relationships between variables while accounting for randomness. This stands in contrast to algorithmic approaches, usually in the form of black-box machine learning models, which focus on the final output, i.e., a prediction, and treat the underlying data generating processes as unknown. Their primary goal is optimizing the performance in making this prediction, but they are typically difficult to interpret. This distinction is further discussed in both Breiman (2001b) and Kauermann et al. (2021).

In the context of conflict research, statistical models are used to investigate the drivers and consequences of armed conflict, identify causal relationships, and more broadly improve our understanding of conflict. Well-known examples in the literature include the work by Von Uexkull et al. (2016), who analyse the impact of growing-season droughts on conflict, and Bazzi and Blattman (2014), who examine the relationship between price shocks and violence. In contrast, machine learning models are primarily employed for forecasting conflict and developing early-warning systems (Hegre et al., 2017). The latter is discussed in more detail in Chapter 4.

Following the view outlined in Cox (2006) and Kauermann et al. (2021), a data generating process can be divided into a systematic component and a stochastic component. The former is deterministic, and captures the structured relationships between variables. The stochastic component, the randomness when observing data in the real-world, allows one to separate the two, draw conclusions and quantify the uncertainty. To capture this data generating process more formally, one assumes that the data comes from a statistical model. Such models can be broadly categorized into parametric and non- (or semi-) parametric models. The former assume a fixed functional form and a set of parameters for the data generating process, whereas the latter lift this restriction for parts of, or the whole data generating process, to allow for more flexibility. In the following section, first parametric models are introduced in general, before discussing linear regression and generalized linear models (GLMs) as specific examples. Then, in Section 3.2, non- and semiparametric models are covered, with a specific focus on generalized additive models (GAMs).

3.1. Parametric Statistical Models

Formally, a random variable Y comes from a parametric probability model

$$Y \sim F(y; \boldsymbol{\theta}).$$

The term $F(y; \boldsymbol{\theta})$ denotes the cumulative distribution function of the random variable defined as $F(y; \boldsymbol{\theta}) = \mathbb{P}(Y \leq y)$, which gives the probability that Y takes a value less than or equal to y. This known distribution function is parametrized by the vector $\boldsymbol{\theta}$, and may contain multiple components p, with $\boldsymbol{\theta} = (\theta_1, \theta_2, ...\theta_p) \in \mathbb{R}^p$. The distribution of the random variable Y can be either discrete or continuous, depending on the nature of the values it can take. A discrete random variable takes values from a countable set, such as in the Binomial or Poisson distribution. In contrast, a continuous random variable can take any value over an interval of real numbers. Examples include both the Normal as well as the Exponential distribution. For an introduction into probability theory, including fundamental concepts and commonly used probability distributions, see for example Fahrmeir et al. (2016).

In practice, the true parameter vector $\boldsymbol{\theta}$ of the distribution is unknown and can only be estimated from observed realisations $y_1, ..., y_n$ of the random variable, i.e. the data sample. Naturally, this means the estimate $\hat{\boldsymbol{\theta}}$, typically indicated through the hat, is dependent on the realisations and thus itself a random variable. Several approaches exist for parameter estimation, with two of the most widely used classic statistical approaches being maximum likelihood and Bayes estimation. Alternatively, one could also take a more traditional optimization route and minimize a loss between predicted and observed values, as often done in predictive tasks (see Chapter 4). An overview on different estimation approaches and the decision which one to choose is for example provided in Wasserman (2013).

3.1.1. Linear Regression

In a regression setting, one is interested in modelling a random Variable Y, the so-called target, response or dependent variable, conditional on one or more explanatory variables, also known as covariates or features $\boldsymbol{x}$. The idea is that the covariates influence the distribution of Y through the parameters $\boldsymbol{\theta}$, and one wants to learn about and model this relationship. Formally, this means modelling the conditional distribution of Y given $\boldsymbol{x}$

$$Y|\boldsymbol{x} \sim F(y; \boldsymbol{\theta}(\boldsymbol{x})).$$

The most simple regression is the classic linear regression model. Here, the mean of the variable of interest Y is modelled as

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where β_0 is the intercept and β_1 the slope of the regression line. The latter measures the influence of x on Y. The term ϵ is known as the error term and captures random deviations with a mean of zero from this relationship. This means the expectation of Y given x can be written as

$$E(Y|x) = \beta_0 + \beta_1 x.$$

In the classic linear regression model, the error term is assumed to follow a normal distribution $\epsilon|x \sim N(0, \sigma^2)$ and thus the conditional distribution

$$Y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

also follows a normal distribution. The parameters of the model, β_0 and β_1, which are unknown, can be estimated from the data through, for example, maximum likelihood. This simple linear regression model can easily be extended to k covariates $x_1, x_2, ..., x_k$ through

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon,$$

where $\beta_1, \beta_2, ..., \beta_k$ are the associated coefficients. Each coefficient measures the effect on the target Y, when the corresponding covariate is increased by one unit, while holding all other covariates constant. This so-called multiple linear regression model is sometimes more compactly written in vector notation as

$$Y = \boldsymbol{x}^T \boldsymbol{\beta} + \epsilon,$$

where $\boldsymbol{x} = (1, x_1, x_2, ..., x_k)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_k)^T$. Note that the covariates $\boldsymbol{x}$ can only linearly affect the target variable, hence the name linear regression. However, the included covariates can be transformed in various ways, thus allowing for more complex relationships with the target. For example, one can use a log-transformation, include polynomials, or also account for interaction effects between variables, by including their product as a covariate in the model.

3.1.2. Generalized Linear Regression

Generalized linear regression models (GLMs) extend standard linear regression by relaxing the assumption that the conditional distribution of Y needs to follows a normal distribution. Instead, GLMs allow Y to follow any distribution from the exponential family, which includes most of the commonly used distributions. This allows one to model a wide variety of target variables, such as discrete counts or strictly positive continuous values.

As in standard linear regression, GLMs model the conditional expectation $E(Y|x)$. However, for transformation, a response function $h(.)$ is required, with

$$E(Y|\boldsymbol{x}) = h(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k), \tag{3.1}$$

which links the linear predictor $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ to the conditional expectation $E(Y|\boldsymbol{x})$. This transformation ensures that the linear predictor η is mapped onto the appropriate scale for Y. Hence, the response function $h(.)$ needs to be chosen accordingly. For example, when modelling count data using a Poisson distribution, an exponential function is typically used as a response function to ensure positivity and to additionally simplify estimation (compared to other possible options; see e.g., McCullagh (2019) for details). Generalized linear models retain their linearity in the linear predictor η. Equivalently to standard linear regression models, the included covariates can be transformed to also account for more complex effects. Notably, due to the transformation of the linear predictor with the response function $h(.)$, covariates may have a non-linear effect on the conditional expectation $E(Y|x)$, even though the linear predictor itself remains a linear combination of the covariates. Estimation can similarly be carried out through maximum likelihood. Generalized linear models were first introduced by Nelder and Wedderburn (1972). A comprehensive overview is, for example, available in Fahrmeir et al. (2022).

GLMs are widely used across disciplines such as economics, social science and political science, to analyse the relationships between variables of interest, often with causal interpretations. Naturally, they also play an important role in conflict research to improve our understanding of armed conflict. For instance, the earlier mentioned work by Bazzi and Blattman (2014) employs classic linear regression models. Buhaug (2010) use both linear and logistic regression, the latter assuming a Bernoulli distribution of the target variable, to examine the effect of climate on African civil wars. Similarly, Fjelde and Hultman (2014) investigate the role of ethnic affiliation in violence against civilians in Africa using GLMs, by employing a negative binomial regression with zero-inflation, the latter to account for excess zeros (Dunn et al., 2018).

3.2. Non- & Semiparametric Statistical Models

While parametric models assume a fixed functional form for the distribution of a random variable Y, non- and semiparametric models relax this assumption, to allow for more flexibility in capturing the sometimes complex relationships observed in the data. Thus, they are used when the true form of the data generating process is unknown or too complex to be captured by a parametric model. However, both require larger sample sizes for a reliable estimation. Nonparametric models make minimal assumptions about the data generating process and allow the data to almost entirely determine the structure of the (estimated) distribution and its relationship between the variables of interest. Semiparametric models, on the other hand, have nonparametric elements while also containing a parametric component. A well-known example of a semiparametric regression model is the generalized additive model, discussed next. A general overview on nonparametric methods is provided in the textbook by Hollander et al. (2013).

3.2.1. Generalized Additive Models

Generalized additive models (GAMs), first introduced by Hastie and Tibshirani (1990), extend GLMs by allowing the linear predictor η to include non-parametric smooth functions of the covariates. The model has the form

$$E(Y|\boldsymbol{x},\boldsymbol{z}) = h(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + m_1(z_1) + m_2(z_2) + ... + m_l(z_l)), \tag{3.2}$$

where, as earlier in (3.1), $\boldsymbol{x}$ are the covariates that are linearly included, and $\boldsymbol{z}$ are the newly included covariates with an unknown nonlinear effect on Y, determined by the smooth functions $m_1(.), ..., m_l(.)$. In order to be able to use the same estimation routines as for GLMs, (3.2) needs to be representable as a linear model (Wood, 2017). This is achieved by choosing a so-called basis, made up of known basis functions, that define the space of possible functions of which $m(.)$ is an element. More specifically, for a covariate z, the unknown smooth function $m(z)$ is defined as

$$m(z) = \boldsymbol{b}(z)^T\boldsymbol{\theta} = \sum_{j=1}^{J} b_j(z)\theta_j, \tag{3.3}$$

where $b_j(z)$ is the j^{th} basis function with an associated unknown coefficient θ_j. By substituting (3.3) into (3.2), one obtains a linear predictor.

There are various possibilities on how to define this basis and estimate the smooth function, such as P-splines (Eilers and Marx, 1996) or regression splines. They all involve setting up a

sufficiently large basis and penalizing $\boldsymbol{\theta}$ in some form, in order to both obtain a good fit as well as smooth function. Possibilities for optimization include cross validation, as well as in-sample criteria such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). An introduction and overview on GAMs is provided by Wood (2017).

GAMs are particularly useful when the relationships between the target variable and (some of) the covariates are non-linear, and difficult or cumbersome to capture through a combination of linear transformations, or when they are completely unknown a priori. They can also model nonlinear interaction effects between covariates, making them highly flexible in a variety of settings. For these reasons, they are a powerful tool for conflict research, to better understand the often complex patterns observed in the data. Despite these reasons, and their prominent use in other fields such as ecology (Guisan et al., 2002) or epidemiology (Schneble et al., 2021), GAMs have seen relatively little application in the conflict literature, with only a few notable exceptions (Zhukov, 2012; Fritz et al., 2022).

Generalized additive models play a key role across all contributions in this thesis. In the contribution in Chapter 6, which systematically evaluates the conflict forecasting performance of remote sensing datasets during the Syrian civil war, a GAM is one of the four models used in the performance evaluation. Notably, the difference in performance to black-box machine learning models, which are covered next in Chapter 4, is only minor. In the contribution in Chapter 8, which analyses language use of Ukrainian users before and during the Russian invasion of Ukraine on Twitter, a generalized additive mixed model (GAMM), which adds random effects into a GAM (see e.g., Faraway, 2016), is used to disentangle sample effects, arising from the in- and outflux of users on the platform, from behavioural effects.

In the contribution in Chapter 7, a non-parametric median regression is used to detect the destruction of buildings in conflict zones. This median regression differs from traditional regression models, by modelling the median instead of the mean, making it more robust to outliers (see Koenker (2005) for details). Specifically, median regressions with a flexible non-linear trend (as in GAMs) are fitted to interferometric coherence scores (Bamler and Hartl, 1998) of individual building pixels over time, in order to separate actual destruction from random background noise in SAR satellite images. Finally, the contribution in Chapter 9 designs a novel generalized additive model to capture the spatio-temporal diffusion effects of armed conflict. The model utilizes a basis of exponential decay functions with varying decay rates, which smooth across the spatio-temporal conflict history of each observation.

4. Forecasting Conflict

"Prediction is very difficult, especially if it's about the future."
— Niels Bohr
(* 1885, † 1962)

Anticipating the future is a central task across a wide range of disciplines, from epidemiology and meteorology to economics and businesses making demand forecasts. The basic idea behind forecasting is to use past and present data to predict future outcomes. This typically means identifying patterns or structures in the observed data and inferring what is likely to happen next. The main goal is to generate the best possible predictions, often without explicitly modelling or understanding the underlying data-generating process. For this reason, algorithmic approaches such as black-box machine learning models are commonly employed.

The prediction of armed conflict has been one of the most important tasks in peace and conflict research for decades (Singer, 1973). As early as 1963, the Correlates of War Project (Small and Singer, 1982) started systematically collecting quantitative data on war, adhering to scientific principles. Since then, the development of early-warning systems for conflict has become one of the main goals of the field. For a long time, most forecasting efforts focused on large-scale, country-level events such as civil wars (Harff and Gurr, 1998; King and Zeng, 2001; Goldstone et al., 2010). Only with the introduction of global disaggregated conflict event datasets such as ACLED and UCDP in the 2010s (see Chapter 2), studies increasingly moved to more fine-grained subnational levels (Hegre et al., 2019). More recently, forecasting competitions hosted by the Violence & Impacts Early-Warning System (ViEWS) project (Vesco et al., 2022; Hegre et al., 2024) have highlighted current advances in the field, both in terms of methodology and the use of new data sources.

As these competitions and many other studies show, forecasting conflict remains a notoriously difficult task. Conflict events are rare, especially at the subnational level, which leads to highly unbalanced datasets. Moreover, only a few of the available data sources actually carry signal that helps improve forecasting performance. Indeed, past conflict remains the best predictor of future conflict (Bazzi et al., 2022). But past conflict usually does not help to anticipate new outbreaks, making the forecasting of conflict onset particularly challenging, as thousands of covariates, or features, as denoted in the machine learning community, are usually employed in an attempt to obtain any performance improvement (Hegre et al., 2021b). For these reasons, black-box machine learning models have typically outperformed more interpretable statistical models in conflict forecasting applications. The following section introduces the core concepts of such predictive models, before providing a short summary of the most popular machine learning approaches in Section 4.2, with a focus on models that have been particularly successful in conflict research. The chapter closes by providing a brief introduction into interpretable machine learning methods, which aim to open the black-box and make machine learning models more interpretable.

4.1. Core Concepts

Any predictive model is built on the idea of minimizing the prediction error. This is formalized through a so-called loss function, which, for each observation, quantifies how well a model's prediction matches the observed target variable. The loss function defines the objective the model minimizes when it is trained, i.e., fitted, and thus is crucial in determining the model's parameters.

One of the most commonly used loss functions for continuous target variables is the L_2-loss, also known as squared error loss, defined as

$$L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2,$$

where y_i is the observed outcome, i.e., the prediction target, and $\hat{y}_i$ the predicted value. Summing this across all observations gives the total loss, which the model seeks to minimize during training, typically via optimization routines such as gradient descent. Notably, a key characteristic of the L_2-loss is that it penalizes larger errors more heavily due to the squaring, making it more sensitive to outliers. For this and other reasons, alternative loss functions such as the L_1-loss, defined as

$$L(y_i, \hat{y}_i) = |y_i - \hat{y}_i|,$$

where the $|.|$ operator denotes taking the absolute value, are sometimes preferred. In classification settings, where the target is categorical (e.g., predicting whether conflict will occur in a given region), different loss functions are used. A common example is the log loss, also known as cross-entropy loss, defined as

$$L(y_i, \hat{p}_i) = -\left(y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)\right),$$

where $\hat{p}_i$ is the predicted probability that $y_i = 1$. Naturally, the choice of loss function depends on the nature of the problem and the particular use case at hand. Importantly, there is a direct link between minimizing loss functions and maximizing the log-likelihood in classical statistical modelling. For instance, maximizing the log-likelihood in a linear regression model under the assumption of normally distributed errors corresponds to minimizing the L_2-loss of the linear predictor. More generally, the assumed distributional form determines the loss function that is minimized (Murphy, 2012).

When fitting a machine learning model, the available data is typically split into a training set and test set. The model is trained, meaning its parameters are chosen optimally, on the training set by minimizing the total loss. The test set, which remains unseen during training, is then used to evaluate the model's predictive performance on new, previously unseen data. This is required, as machine learning models, due to their large number of parameters, can easily overfit the training data, thus capture noise instead of the underlying structure, and hence perform poorly on new observations. Notably, in forecasting applications, this split becomes substantially more challenging, as the temporal dimension of the data needs to be taken into account. Specifically, the training set must not contain any information from future time points, i.e., observations from the evaluation period, to avoid data leakage and to ensure a realistic performance assessment. An overview over more elaborate evaluation strategies, such as cross-validation, is provided in Arlot and Celisse (2010). Best practices for forecasting and model evaluation in forecasting settings are, for example, discussed in Petropoulos et al. (2022). Among others, the latter work covers

time-series cross validation, which is used for model evaluation in the contributions in Chapter 6 and Chapter 9.

Loss functions are closely related to, but not equivalent to evaluation metrics, the latter of which assess model performance. For continuous outcomes, a common evaluation metric is the mean squared error

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

which corresponds directly to the average L_2-loss on the test set. Other metrics include the mean absolute error (MAE) for regression, or accuracy, F1-score, and AUC for classification. While evaluation metrics are only chosen based on the specific goals of an application, loss functions also need to have desirable optimization properties that allow for efficient model training. As a result, loss functions may not always align perfectly with the evaluation metric used to assess model performance. Further information and a general introduction into predictive modelling and machine learning is for example provided in James et al. (2013).

4.2. Common Machine Learning Models

4.2.1. LASSO Regression

One of the simplest yet often surprisingly effective prediction models, especially in settings with a large number of features, is Least Absolute Shrinkage and Selection Operator (LASSO) regression, originally proposed by Tibshirani (1996). Like generalized additive models (GAMs), LASSO belongs to the broader class of penalized regression models. It extends the (generalized) linear regression framework by adding an L_1 penalty to the loss function.

Formally, in a linear regression setting, LASSO solves the optimization problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{k} |\beta_j| \right\},$$

where y_i is the prediction target and $\boldsymbol{x}_i$ is the vector of k covariates for observation i in the training set. The first term denotes the standard L_2-loss, which is minimized in linear regression, the second term imposes an L_1 penalty on the coefficients $\boldsymbol{\beta}$. The tuning parameter $\lambda \geq 0$ controls the strength of this penalty. Hence, when $\lambda = 0$, this reduces to a standard linear regression model. As λ increases, coefficients $\boldsymbol{\beta}$ are shrunk towards zero, with some (the least important ones for the prediction task), being set to exactly zero and thus effectively excluding them from the model. This property makes LASSO particularly attractive in sparse high-dimensional settings, where the number of covariates k is large, but only a few have an influence on y, as often the case in conflict forecasting applications. The exact value of λ is typically selected optimally via cross-validation.

In a way, LASSO bridges the gap between classical statistical modelling and more complex machine learning approaches. While the model retains the interpretability of a linear model, the regularization effect of the penalty generally improves predictive performance compared to standard linear regression. Its main limitation is the linearity in effects between covariates and target. As in GLMs, non-linear and interaction effects need to be explicitly included as covariates to be captured. Additionally, one needs to be aware that in case of highly correlated covariates,

LASSO selects one at random, while all others are (nearly) zeroed out. Closely-related methods include Ridge Regression (Hoerl and Kennard, 1970), which replaces the L_1 penalty with an L_2 penalty, and Elastic Net (Zou and Hastie, 2005), which includes both penalties in the optimization problem.

LASSO regression has been successfully employed across a wide variety of disciplines such as ecology (Tredennick et al., 2021) and medicine (Gotlieb et al., 2022), and has been used in several conflict forecasting studies (Mueller and Rauh, 2018; Bazzi et al., 2022). In the contribution in Chapter 6, which systematically evaluates the conflict forecasting performance of remote sensing datasets during the Syrian civil war, LASSO logistic regression is one of the four models used in the performance evaluation.

4.2.2. Decision Trees

Decision trees, popularized by Breiman et al. (1984), recursively partition the feature space into disjoint regions, with the aim of improving predictive accuracy at each split. Each split is defined by a simple decision rule based on the value of one of the features. The resulting model resembles a tree structure, hence the name, where internal nodes define the splitting rules, and terminal nodes, the so-called leaves, assign a predicted value to all observations that fall into the corresponding partition. An example is depicted in Figure 4.1.

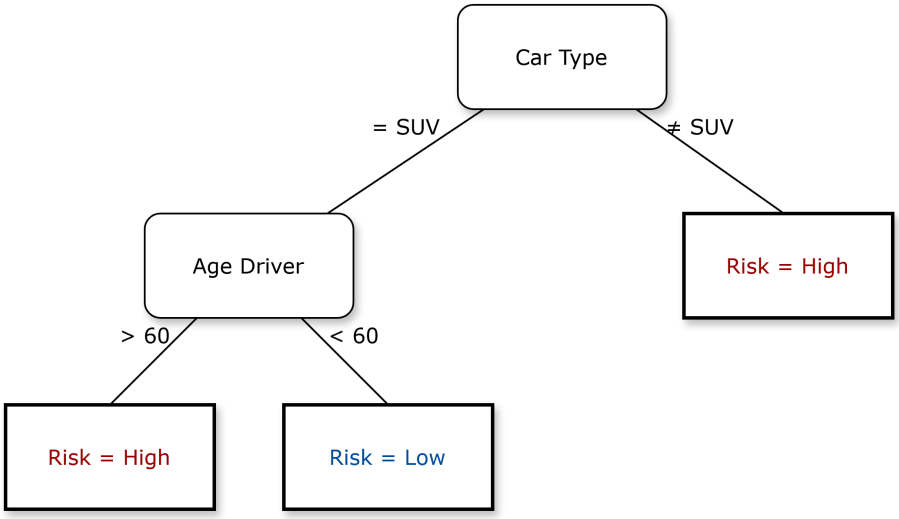


Figure 4.1.: Exemplary depiction of a decision tree. Here, the decision tree classifies driver risk based on car type and age. If the car is not an SUV, the risk is classified as high. If the car is an SUV, the driver's age determines the outcome. Then, over 60 leads to a high risk, while under 60 results in a low risk.

The tree is constructed in a top-down recursive manner. At each step, a feature and corresponding threshold is chosen that best split the data into two homogenous groups based on a chosen impurity criteria or loss function. For regression trees, where the target variable is continuous, the L_2-loss is typically employed. Hence, at each step, the tree seeks to find the split that minimizes

$$\sum_{i:x_i \in P_1} (y_i - \hat{y}_{P_1})^2 + \sum_{i:x_i \in P_2} (y_i - \hat{y}_{P_2})^2,$$

where y_i is the observed target variable for an observation i, with a corresponding covariate x_i that determines which partition (P_1 or P_2) the observation is assigned to. The term $\hat{y}_{P_1}$ denotes the mean target across the observations in partition P_1, and $\hat{y}_{P_2}$ the mean in P_2. This splitting procedure continues until a pre-defined convergence criteria is met, for example, a minimum amount of observations in a partition. Usually, for reasons of computational complexity, each partition is assigned a constant prediction, with the mean being the optimal choice under the L_2-loss.

Due to their recursive construction, decision trees can model both non-linear effects and interactions between features, while remaining easy to interpret through these decision rules. However, they are also known to be highly sensitive to small changes in the data and prone to overfitting. Thus, individual trees are generally not used when prediction performance is the main priority. Nonetheless, they form the building blocks of several successful ensemble methods, such as random forests and boosting, which are discussed in the following.

4.2.3. Random Forests

Random forests are an ensemble learning method that uses bagging to train multiple decision trees to improve model stability and thus predictive performance. Ensemble learning refers to techniques that combine multiple individual models, often called base learners, to form a single, (stronger) composite model. Random forests date back to Breiman (2001a) and have since become a standard tool across many fields.

As discussed earlier, single decision trees are highly sensitive to small changes in the training data, meaning they exhibit high variance. Random forests address this by employing bagging, a technique where multiple samples from the original training dataset are drawn with replacement, i.e. bootstrapped, to train each tree on a different bootstrap sample. The predictions from all B trees are then averaged,

$$\hat{y}(\boldsymbol{x}_i) = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_b(\boldsymbol{x}_i),$$

where $\hat{y}_b(\boldsymbol{x})$ is the prediction made by the b-th tree based on the covariates $\boldsymbol{x}_i$, and $\hat{y}(\boldsymbol{x}_i)$ denotes the final (averaged) prediction made by the forest. This aggregation reduces variance and leads to a more robust model. Random forests introduce an additional layer of variety, by drawing a random subset of features that is considered at each split within each tree. This ensures that individual decision trees actually differ from one another, essentially de-correlating them to reduce the variance of the random forest. Without this subsetting, the trees may look very similar if there are strong predictors that dominate the splits, and the variance reduction from bagging would be limited. Since each tree is trained independently on a different bootstrap sample, the training can be parallelized.

Random forests can freely model complex non-linear relationships and interactions between features, without requiring these to be specified in advance. Contrary to individual decision trees, they are not sensitive to small changes in the dataset and are relatively robust to overfitting. However, due to the aggregation of often hundreds of trees, the final model is essentially a black box and cannot be interpreted. While random forests have many hyperparameters (e.g., the number of trees, the maximum depth of each tree, the minimum number of observations per leaf node), they tend to perform well even without any tuning (Fernández-Delgado et al., 2014).

Due to these reasons, random forests have become one of the most widely used machine learning methods, and are employed across various disciplines, including medicine (Jia et al., 2019), finance (Ballings et al., 2015), criminology (Kounadi et al., 2020), and psychology (Stachl et al., 2020). In recent years, in conflict research, they have become one of the most commonly used approaches for forecasting conflict (Colaresi and Mahmood, 2017; Bazzi et al., 2022; Rød et al., 2025). They also serve as the primary modelling component of the conflict early warning system and ensemble model ViEWS (Hegre et al., 2019, 2021a) and are integrated into several other early warning systems (Rød et al., 2024). In the contribution in Chapter 6, a random forest is among the four models evaluated, and achieves the best performance in forecasting conflict during the Syrian civil war. Additionally, the contributing article in Chapter 8, which analyses tweets in Ukraine, uses a random forest in the data cleaning stage to detect and filter out bots from the dataset.

4.2.4. (Gradient) Boosting

Boosting is another ensemble technique that combines multiple base learners into a single composite model. Unlike bagging, where each base learner is trained independently on bootstrapped samples, boosting algorithms iteratively fit new models by using information from previous iterations to continuously update and modify the training dataset. The original idea dates back to Schapire (1990), with AdaBoost (Freund and Schapire, 1997) being the first widely successful implementation.

The central idea behind boosting is to sequentially fit new base learners to the residuals of the current ensemble model. Hence, each new base learner specifically targets observations that the current ensemble struggles to predict accurately and thus allows the boosting algorithm to progressively refine predictions. The final model is a weighted sum of sequentially fitted base learners

$$\hat{y}(\boldsymbol{x}_i) = \sum_{b=1}^{B} \alpha_b \hat{y}_b(\boldsymbol{x_i}),$$

where $\hat{y}_b(\boldsymbol{x}_i)$ is the prediction made by the b-th base learner based on covariates $\boldsymbol{x}_i$, α_b denotes the corresponding weight assigned to that base learner's contribution, and $\hat{y}(\boldsymbol{x}_i)$ denotes the final prediction. The weights α are typically determined by each learner's ability to reduce prediction errors. Decision trees are most commonly chosen as base learners, as they are flexible and fast to construct, and usually lead to the best predictive performance of the boosted ensemble (Breiman, 1998; Hastie et al., 2009).

Gradient boosting, first introduced by Friedman (2001), views boosting as an iterative functional gradient descent routine, step by step "nudging" the model's predictions closer and closer to the observed data points. This allows for an efficient training procedure across various different loss functions and predictive tasks, which has made gradient boosting the preferred boosting approach. Popular and efficient implementations of gradient boosting include XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017).

Equivalently to random forests, gradient boosting methods can freely model complex non-linear relationships and interactions between features, but due to the ensembling also constitute black-boxes. Notably, they are often able to achieve higher predictive performances than random forests (Borisov et al., 2022). However, to perform well, they usually require hyperparameter tuning, as gradient boosting approaches are much more prone to overfitting. Furthermore, since the base

learners are trained sequentially rather than in parallel, the overall training procedure tends to take longer.

Gradient boosting generally outperforms other machine learning methods on tabular data and is thus often considered as state-of-the-art (Borisov et al., 2022; Grinsztajn et al., 2022). Recently, gradient boosting has also started to see increasing use in conflict forecasting applications (Vestby et al., 2022; Bazzi et al., 2022) and early-warning ensembles (Rød et al., 2024). In the contributing article in Chapter 6, XGBoost is among the four models evaluated, with predictive performance slightly below the random forest.

4.2.5. Other Machine Learning Approaches

In addition to the models discussed so far, naturally a wide range of other machine learning approaches exists. Examples include support vector machines (SVMs) and k-nearest neighbors (KNN) (James et al., 2013). However, in conflict forecasting, they have only seen limited application, as tree-based models such as random forests and gradient boosting generally achieve better predictive performance in these difficult tabular settings with large numbers of features.

Neural networks deserve explicit mention, as they have become increasingly important across a variety of fields, with the rise and improvement of deep learning methods over the past decade. Neural networks model complex relationships between covariates and target(s) by passing the data through multiple layers of interconnected nodes, called neurons, each applying a (non-linear) function on the neurons of the previous layer. Deep neural networks, which stack many such layers, can capture highly non-linear patterns and interactions in the data. A comprehensive introduction is, for example, provided in Goodfellow et al. (2016).

While deep learning has been widely successful in both computer vision and natural language processing (NLP), including the development of large language models (LLMs), its application to structured tabular data, as typically used for conflict forecasting, has been more limited. As studies show (Borisov et al., 2022; Grinsztajn et al., 2022), tree-based models still continue to outperform deep learning methods in these settings and, for this reason, they have likely not yet been integrated into conflict early-warning systems (Rød et al., 2024).

Nonetheless, deep learning methods play an important role when employing alternative data sources. For example, convolutional neural networks (CNNs) (Schmidhuber, 2015) and, more recently, vision transformers (Dosovitskiy et al., 2021) are used for analysing (satellite) images (Cong et al., 2022), while various transformer-based architectures (Vaswani et al., 2017) have become standard tools for analysing text data. Hence, deep learning methods are also increasingly applied in conflict research for such tasks (Won et al., 2017; Sticher et al., 2023). In the contribution in Chapter 8, a multilingual SentenceBERT model (Reimers and Gurevych, 2019) is used to infer the topics discussed in Ukrainian tweets before and during the Russian invasion.

4.3. Interpretable Machine Learning Techniques

While machine learning models achieve excellent predictive performance, most of them are considered black-boxes, as it is almost impossible to understand how these models arrive at their

predictions. In many applications however, interpretability is highly important, both to ensure model robustness and to potentially gain insights into the underlying processes and mechanisms.

Hence, in recent years, a variety of methods have been developed to make machine learning models and their predictions more interpretable. Most of these methods are model-agnostic, meaning they can be applied independently of the specific machine learning model to be interpreted. Following Molnar (2020), these so-called interpretable machine learning (IML) techniques can be divided into local and global methods. Local methods try to explain individual predictions, whereas global methods aim to explain the overall behaviour of a machine learning model across a dataset. In the following, one example of each is briefly introduced.

Individual Conditional Expectation (ICE) ICE plots represent a local interpretation technique. For a given observation, an ICE plot shows how the model's prediction changes when varying a single covariate, while keeping all others fixed. This is achieved by creating different versions of the same observation, replacing the covariate's original value with values from a pre-defined grid, and then making predictions for each modified observation. This procedure can be repeated for multiple (or all) observations to identify general effects and, for example, to detect interactions between covariates.

Permutation Feature Importance Permutation feature importance is a global interpretation method that measures the importance of a covariate, i.e., feature, by quantifying the increase in the model's prediction error when randomly permuting the covariate's values across a dataset. This permutation breaks any systematic association between covariate and target and thus eliminates any predictive power. More specifically, one first calculates the original error of the model

$$E^{\text{orig}} = \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, \hat{y}(\boldsymbol{x}_i)\right),$$

where y_i is the observed target, $\hat{y}(\boldsymbol{x}_i)$ is the prediction made by the model based on covariates $\boldsymbol{x}_i$ for observation i, and $L(.,.)$ denotes the pre-defined loss function. Then, for a covariate j, its values are randomly permuted across the dataset and the prediction error is recalculated,

$$E^{\text{perm},j} = \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, \hat{y}(\boldsymbol{x}_i^{\text{perm},(j)})\right),$$

where $\boldsymbol{x}_i^{\text{perm},(j)}$ denotes the covariate vector of an observation i after permuting covariate x_i^j. The permutation feature importance for covariate j is then computed by

$$\text{PFI}^j = E^{\text{perm},j} - E^{\text{orig}},$$

which measures the average drop in predictive performance after permutation, and thus the importance of that covariate for the model's predictive power. To achieve a more robust estimate, the calculation of $E^{perm,j}$ can be repeated multiple times and the results averaged. In practice, permutation feature importance is usually computed on the test set due to potential model overfitting.

While IML methods are useful, they also face limitations, particularly in high-dimensional settings with many correlated covariates. For example, importance scores become unstable and misleading

in the presence of high correlation between covariates. A detailed discussion on this is provided in Hooker et al. (2021). Hence, if interpretability is the primary goal, interpretable statistical models (see Chapter 3) are likely the better option.

Despite some of these limitations, IML methods have also started to see use in conflict research (see e.g., Rød et al., 2025, or Scholz et al., 2025). In the contributing article in Chapter 6, permutation feature importance is employed to analyse the importance of individual remote sensing variables for the random forest. A comprehensive overview on IML methods is provided in Molnar (2020).

5. Concluding Remarks

The increasing availability of fine-grained conflict data, alongside the emergence of new data sources, is reshaping the study of armed conflict. This dissertation explores how modern data science methods can leverage these developments to address longstanding questions in conflict research that were previously difficult or impossible to study. The contributing articles illustrate how both statistical modelling and machine learning, when combined with these datasets, can advance our understanding of conflict and its effects, and how they can be jointly used for predictive tasks.

5.1. Contributions

Part II focuses on how to utilize and integrate satellite and social media data, both new data sources for the field, into statistical and machine learning models. The contribution in Chapter 6 showcases that remote sensing datasets, derived from satellite imagery, can improve conflict forecasting performance in conflict-ridden countries, and thus serve as a replacement for the lack of official data sources in such countries. Specifically, the inclusion of remote sensing variables improves the models' ability to predict conflict onset during the Syrian civil war. Once all remote sensing datasets are included, the performance gap between statistical and machine learning models narrows substantially, suggesting that, given enough informative data is available, researchers may be able to forego the use of black-box machine learning models in favour of more interpretable alternatives.

The following contribution in Chapter 7 moves beyond remote sensing datasets and directly employs SAR satellite imagery to detect building destruction during war. Interferometric SAR (InSAR), a remote sensing algorithm that measures ground surface deformation between two radar images, is applied to generate pixel-wise stability scores. These scores are then evaluated over time with a non-parametric median regression and a robust estimate of the standard deviation of the resulting residuals, in order to identify destruction and its timing at the building level.

The contributing article in Chapter 8 utilizes social media data, specifically tweets from Ukraine, to analyse tweeting behaviour in Ukraine before and during the Russian invasion. First, the tweets undergo an extensive cleaning routine, including a random forest classifier to detect and filter out bot-generated tweets. Then, tweeting activity and language patterns are investigated via generalized additive mixed models (GAMMs), revealing a stark shift in language from Russian to Ukrainian following the outbreak of the war.

As also evident in the contribution in Chapter 6, past conflict is the best predictor of future conflict. Nonetheless, the diffusion of conflict, i.e., its spread across time and space, is not sufficiently accounted for in statistical models currently employed in the field. To address this, Part III develops a statistical model that captures the diffusion of conflict across space and time.

Specifically, the contribution in Chapter 9 proposes a novel generalized additive model with a smoothing basis over past conflict, which is constructed from a set of exponential decay functions with varying decay rates. The results show that armed conflict in Africa exhibits long-lasting and far-reaching dependencies that decay exponentially in both space and time, which can be captured and interpreted using the developed model.

5.2. Outlook

Looking ahead, there are several promising directions for future research. First, the data sources used in this thesis, satellite imagery and social media, are still relatively new to the field of conflict research and thus remain underutilized. As such, they hold considerable potential for addressing a broad set of questions and methodological improvements. For example, social media data could be integrated into early warning systems by monitoring online user behaviour through content and sentiment. In addition, building on the work in Chapter 8, social media studies could be expanded to infer geographic locations from the text of social media posts, allowing for regionally specific insights by matching post content to coordinates. Moreover, as satellite technology continues to improve, both the resolution and temporal availability of freely available satellite images are likely to increase. This will allow for a simpler and more reliable tracking of infrastructure damage and destruction, as well as the broader dynamics of conflicts and wars, in near-real time.

Second, future work could focus on developing more specialized models tailored to investigate specific mechanisms or particular conflicts in greater detail. For example, the diffusion model developed in the contributing article in Chapter 9 could be extended to incorporate actor-specific information, shedding light into how conflicts unfold between rebel groups and governments. Likewise, predictive models, part of early warning systems, could be adapted to better reflect the grid-based structure commonly adopted in forecasting scenarios, for instance, by employing convolutional neural networks (CNNs) or other machine learning architectures that explicitly model the spatial dependencies between the cells.

Third, an ongoing challenge in conflict research is the often substantial reporting delay that comes with the event data. These delays complicate real-time monitoring of armed conflicts, hinder the evaluation of peacekeeping efforts, and can distort forecasts, thereby reducing the reliability of early warning systems. Nowcasting methods, which aim to correct for such delays and have been successfully applied in other domains, for example during the COVID-19 pandemic (Schneble et al., 2021), could help to address this issue.

Finally, as conflict forecasts are increasingly informing real-world decision making, the quantification of uncertainty around these forecasts is becoming ever more important. This is also reflected in the most recent ViEWS conflict forecasting competition, which explicitly focused on evaluating the quality of uncertainty estimates (Hegre et al., 2024). As of the time of writing, none of the publicly available conflict early-warning systems provides uncertainty estimates alongside their forecasts yet.

Summarizing, this dissertation demonstrates how data science can address fundamental challenges and answer central questions in conflict research. By combining statistical modelling and machine learning with data sources such as satellite imagery and social media, it shows how conflict can be studied at fine-grained spatial and temporal levels, even in the absence of reliable official data sources. The contributions highlight both predictive and explanatory approaches and underscore

the importance of developing scalable and robust methods that account for the specific characteristics and limitations of conflict data. Advancing this line of work will require interdisciplinary collaboration, bringing together domain knowledge from political science and conflict research with methodological expertise from statistics and computer science.

References

Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., and Jetz, W. (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific data*, 5(1): 1–15.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4: 40 – 79.

Ballings, M., Van den Poel, D., Hespeels, N., and Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert systems with Applications*, 42(20): 7046–7056.

Bamler, R. and Hartl, P. (1998). Synthetic aperture radar interferometry. *Inverse problems*, 14(4): R1.

Bazzi, S., Blair, R. A., Blattman, C., Dube, O., Gudgeon, M., and Peck, R. (2022). The promise and pitfalls of conflict prediction: Evidence from Colombia and Indonesia. *Review of Economics and Statistics*, 104(4): 764–779.

Bazzi, S. and Blattman, C. (2014). Economic shocks and conflict: Evidence from commodity prices. *American Economic Journal: Macroeconomics*, 6(4): 1–38.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2022). Deep neural networks and tabular data: A survey. *IEEE transactions on neural networks and learning systems*.

Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics*, 26(3): 801–849.

Breiman, L. (2001a). Random forests. *Machine learning*, 45: 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3): 199–231.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). *Classification and Regression Trees*. CRC Press.

Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., and Smets, B. (2020). Copernicus global land cover layers—collection 2. *Remote Sensing*, 12(6): 1044.

Buhaug, H. (2010). Climate not to blame for african civil wars. *Proceedings of the National Academy of Sciences*, 107(38): 16477–16482.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Cheng, G. and Han, J. (2016). A survey on object detection in optical remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 117: 11–28.

Colaresi, M. and Mahmood, Z. (2017). Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research*, 54(2): 193–214.

Collier, P. et al. (2003). *Breaking the conflict trap: Civil war and development policy*, volume 41181. World Bank Publications.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., and Ermon, S. (2022). Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35: 197–211.

Cook, S. J. and Weidmann, N. B. (2022). Race to the bottom: Spatial aggregation and event data. *International Interactions*, 48(3): 471–491.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge university press.

Davidson, B. I., Wischerath, D., Racek, D., Parry, D. A., Godwin, E., Hinds, J., Van Der Linden, D., Roscoe, J. F., Ayravainen, L., and Cork, A. G. (2023). Platform-controlled social media APIs threaten open science. *Nature Human Behaviour*, 7(12): 2054–2057.

Davies, S., Engström, G., Pettersson, T., and Öberg, M. (2024). Organized violence 1989–2023, and the prevalence of organized crime groups. *Journal of Peace Research*, 61(4): 673–693.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Dowd, C., Justino, P., Kishi, R., and Marchais, G. (2020). Comparing 'new' and 'old' media for violence monitoring and crisis response: Evidence from Kenya. *Research & Politics*, 7(3): 2053168020937592.

Dunn, P. K., Smyth, G. K., et al. (2018). *Generalized linear models with examples in R*, volume 53. Springer.

Eck, K. (2012). In data we trust? A comparison of UCDP GED and ACLED conflict events datasets. *Cooperation and Conflict*, 47(1): 124–141.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2): 89–121.

Elvidge, C. D., Baugh, K., Zhizhin, M., Hsu, F. C., and Ghosh, T. (2021). VIIRS night-time lights. In *Remote Sensing of Night-time Light*, pages 6–25. Routledge.

Enright, K. (2022). An introduction to optical satellite imagery. Available at https://up42.com/blog/introduction-optical-satellite-imagery, Accessed: 01.05.2025.

ESA. (2025). The sentinel missions. Available at https://www.esa.int/Applications/Observing_the_Earth/Copernicus/The_Sentinel_missions, Accessed: 03.05.2025.

References

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., and Tutz, G. (2016). *Statistik: Der Weg zur Datenanalyse*. Springer-Verlag.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D. (2022). Regression models. In *Regression: Models, methods and applications*, pages 23–84. Springer.

FAO. (2025). Agricultural stress index system (ASIS). Available at http://www.fao.org/giews/earthobservation/, Accessed: 03.05.2025.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1): 3133–3181.

Fjelde, H. and Hultman, L. (2014). Weakening the enemy: A disaggregated study of violence against civilians in Africa. *Journal of Conflict Resolution*, 58(7): 1230–1257.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1): 119–139.

Friedl, M. and Sulla-Menashe, D. (2022). MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061 [Data set]. Available at https://doi.org/10.5067/MODIS/MCD12Q1.061, Accessed: 14.05.2025.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Fritz, C., Mehrl, M., Thurner, P. W., and Kauermann, G. (2022). The role of governmental weapons procurements in forecasting monthly fatalities in intrastate conflicts: A semiparametric hierarchical hurdle model. *International Interactions*, 48(4): 778–799.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., et al. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes, version 2.0. *Scientific data*, 2(1): 1–21.

Gates, S., Hegre, H., Nygård, H. M., and Strand, H. (2012). Development consequences of armed conflict. *World Development*, 40(9): 1713–1722.

Geissler, D., Bär, D., Pröllochs, N., and Feuerriegel, S. (2023). Russian propaganda on social media during the 2022 invasion of Ukraine. *EPJ Data Science*, 12(1): 35.

Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G., Ulfelder, J., and Woodward, M. (2010). A global model for forecasting political instability. *American journal of political science*, 54(1): 190–208.

Gómez, C., White, J. C., and Wulder, M. A. (2016). Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of photogrammetry and Remote Sensing*, 116: 55–72.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Gotlieb, N., Azhie, A., Sharma, D., Spann, A., Suo, N.-J., Tran, J., Orchanian-Cheff, A., Wang, B., Goldenberg, A., Chassé, M., et al. (2022). The promise of machine learning applications in solid organ transplantation. *NPJ digital medicine*, 5(1): 89.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35: 507–520.

Guisan, A., Edwards Jr, T. C., and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological modelling*, 157(2-3): 89–100.

Harff, B. and Gurr, T. R. (1998). Systematic early warning of humanitarian emergencies. *Journal of Peace Research*, 35(5): 551–579.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, volume 43. CRC Press.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hegre, H., Allansson, M., Basedau, M., Colaresi, M., Croicu, M., Fjelde, H., Hoyles, F., Hultman, L., Högbladh, S., Jansen, R., et al. (2019). Views: A political violence early-warning system. *Journal of peace research*, 56(2): 155–174.

Hegre, H., Bell, C., Colaresi, M., Croicu, M., Hoyles, F., Jansen, R., Leis, M. R., Lindqvist-McGowan, A., Randahl, D., Rød, E. G., et al. (2021a). Views2020: Revising and evaluating the views political violence early-warning system. *Journal of peace research*, 58(3): 599–611.

Hegre, H., Croicu, M., Eck, K., and Högbladh, S. (2020). Introducing the UCDP candidate events dataset. *Research & Politics*, 7(3): 2053168020935257.

Hegre, H., Metternich, N. W., Nygård, H. M., and Wucherpfennig, J. (2017). Introduction: Forecasting in peace research. *Journal of Peace Research*, 54(2): 113–124.

Hegre, H., Nygård, H. M., and Landsverk, P. (2021b). Can we predict armed conflict? How the first 9 years of published forecasts stand up to reality. *International Studies Quarterly*, 65(3): 660–668.

Hegre, H., Vesco, P., Colaresi, M., Vestby, J., Timlick, A., Kazmi, N. S., Becker, F., Binetti, M., Bodentien, T., Bohne, T., et al. (2024). The 2023/24 views prediction challenge: Predicting the number of fatalities in armed conflict, with uncertainty. *arXiv preprint arXiv:2407.11045*.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1): 55–67.

Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods*. John Wiley & Sons.

Hooker, G., Mentch, L., and Zhou, S. (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31: 1–16.

Jakubik, J., Yang, F., Blumenstiel, B., Scheurer, E., Sedona, R., Maurogiovanni, S., Bosmans, J., Dionelis, N., Marsocci, V., Kopp, N., et al. (2025). Terramind: Large-scale generative multimodality for earth observation. *arXiv preprint arXiv:2504.11171*.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An Introduction to Statistical Learning*, volume 112. Springer.

Jia, T.-Y., Xiong, J.-F., Li, X.-Y., Yu, W., Xu, Z.-Y., Cai, X.-W., Ma, J.-C., Ren, Y.-C., Larsson, R., Zhang, J., et al. (2019). Identifying EGFR mutations in lung adenocarcinoma by noninvasive imaging using radiomics features and random forest modeling. *European radiology*, 29: 4742–4750.

Kauermann, G., Küchenhoff, H., and Heumann, C. (2021). *Statistical Foundations, Reasoning and Inference*. Springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

King, G. and Zeng, L. (2001). Improving forecasts of state failure. *World Politics*, 53(4): 623–658.

Koenker, R. (2005). *Quantile Regression*, volume 38. Cambridge university press.

Kounadi, O., Ristea, A., Araujo, A., and Leitner, M. (2020). A systematic review on spatial crime forecasting. *Crime science*, 9: 1–22.

Kugler, F., Lee, S.-K., Hajnsek, I., and Papathanassiou, K. P. (2015). Forest height estimation by means of pol-insar data inversion: The role of the vertical wavenumber. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10): 5294–5311.

McCullagh, P. (2019). *Generalized Linear Models*. Routledge.

Molnar, C. (2020). *Interpretable Machine Learning*. Lulu. com.

Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I., and Papathanassiou, K. P. (2013). A tutorial on synthetic aperture radar. *IEEE Geoscience and remote sensing magazine*, 1(1): 6–43.

Mueller, H., Groeger, A., Hersh, J., Matranga, A., and Serrat, J. (2021). Monitoring war destruction from space using machine learning. *Proceedings of the national academy of sciences*, 118(23): e2025400118.

Mueller, H. and Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, 112(2): 358–375.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT press.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3): 370–384.

Pebesma, E. and Bivand, R. (2023). *Spatial Data Science: With Applications in R*. Chapman and Hall/CRC.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., et al. (2022). Forecasting: theory and practice. *International Journal of forecasting*, 38(3): 705–871.

Planet Labs. (2025). Planet pricing. Available at https://www.planet.com/pricing/, Accessed: 01.05.2025.

Raleigh, C., Linke, R., Hegre, H., and Karlsen, J. (2010). Introducing acled: An armed conflict location and event dataset. *Journal of peace research*, 47(5): 651–660.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Rød, E. G., Gåsste, T., and Hegre, H. (2024). A review and comparison of conflict early warning systems. *International Journal of Forecasting*, 40(1): 96–112.

Rød, E. G., Hegre, H., and Leis, M. (2025). Predicting armed conflict using protest data. *Journal of Peace Research*, 62(1): 3–20.

Rohner, D. (2024). Mediation, military, and money: The promises and pitfalls of outside interventions to end armed conflicts. *Journal of Economic Literature*, 62(1): 155–195.

Rufener, M.-C., Ofli, F., Fatehkia, M., and Weber, I. (2024). Estimation of internal displacement in Ukraine from satellite-based car detections. *Scientific Reports*, 14(1): 31638.

Ruths, D. and Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213): 1063–1064.

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, 5: 197–227.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61: 85–117.

Schneble, M., De Nicola, G., Kauermann, G., and Berger, U. (2021). Nowcasting fatal COVID-19 infections on a regional level in germany. *Biometrical Journal*, 63(3): 471–489.

Scholz, S., Weidmann, N. B., Steinert-Threlkeld, Z. C., Keremoğlu, E., and Goldlücke, B. (2025). Improving computer vision interpretability: Transparent two-level classification for complex scenes. *Political Analysis*, 33(2): 107–121.

Singer, J. D. (1973). The peace researcher and foreign policy prediction. *Peace Science Society (International)*, 21: 1–13.

Small, M. and Singer, J. D. (1982). Resort to arms: International and civil wars, 1816-1980. *(No Title)*.

Sobolev, A., Chen, M. K., Joo, J., and Steinert-Threlkeld, Z. C. (2020). News and geolocated social media accurately measure protest size variation. *American Political Science Review*, 114(4): 1343–1351.

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., et al. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, 117(30): 17680–17687.

Sticher, V., Wegner, J. D., and Pfeifle, B. (2023). Toward the remote monitoring of armed conflicts. *PNAS nexus*, 2(6): pgad181.

Sundberg, R. and Melander, E. (2013). Introducing the UCDP georeferenced event dataset. *Journal of peace research*, 50(4): 523–532.

Tatem, A. J. (2017). Worldpop, open data for spatial demography. *Scientific data*, 4(1): 1–4.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1): 267–288.

Tiecke, T. G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., Kilic, T., Murray, S., Blankespoor, B., Prydz, E. B., et al. (2017). Mapping the world population one building at a time. *arXiv preprint arXiv:1712.05839*.

Tredennick, A. T., Hooker, G., Ellner, S. P., and Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, 102(6): e03336.

USGS. (2018). Free, Open Landsat Data Unleashed the Power of Remote Sensing a Decade Ago. Available at https://www.usgs.gov/news/free-open-landsat-data-unleashed-power-remote-sensing-a-decade-ago, Accessed: 05.03.2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vesco, P., Hegre, H., Colaresi, M., Jansen, R. B., Lo, A., Reisch, G., and Weidmann, N. B. (2022). United they stand: Findings from an escalation prediction competition. *International Interactions*, 48(4): 860–896.

Vestby, J., Brandsch, J., Larsen, V. B., Landsverk, P., and Tollefsen, A. F. (2022). Predicting (de-) escalation of sub-national violence using gradient boosting: Does it work? *International Interactions*, 48(4): 841–859.

Von Uexkull, N., Croicu, M., Fjelde, H., and Buhaug, H. (2016). Civil conflict sensitivity to growing-season drought. *Proceedings of the National Academy of Sciences*, 113(44): 12391–12396.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380): 1146–1151.

Wan, Z., Hook, S., and Hulley, G. (2021). MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061 [Data set]. Available at https://doi.org/10.5067/MODIS/MOD11A1.061, Accessed: 14.05.2025.

Wasserman, L. (2013). *All of statistics: A concise course in statistical inference*. Springer Science & Business Media.

Won, D., Steinert-Threlkeld, Z. C., and Joo, J. (2017). Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 786–794.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. chapman and hall/CRC.

Wulder, M. A., Roy, D. P., Radeloff, V. C., Loveland, T. R., Anderson, M. C., Johnson, D. M., Healey, S., Zhu, Z., Scambos, T. A., Pahlevan, N., et al. (2022). Fifty years of Landsat science and impacts. *Remote Sensing of Environment*, 280: 113195.

X. (2025). X developer portal - products. Available at https://developer.x.com/en/portal/products, Accessed: 01.05.2025.

Zeitzoff, T. (2017). How social media is changing conflict. *Journal of Conflict Resolution*, 61(9): 1970–1991.

Zhukov, Y. M. (2012). Roads and the diffusion of insurgent violence: The logistics of conflict in russia's north caucasus. *Political geography*, 31(3): 144–156.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2): 301–320.

Part II.

Utilizing Novel Data Sources

6. Conflict forecasting using remote sensing data: An application to the Syrian civil war

Contributing article

Racek, D., Thurner, P. W., Davidson, B. I., Zhu, X. X., and Kauermann, G. (2024). Conflict forecasting using remote sensing data: An application to the Syrian civil war. *International Journal of Forecasting, 40(1):373–391.* https://doi.org/10.1016/j.ijforecast.2023.04.001.

Data and code

Available at https://osf.io/vmqct.

Copyright information

Supplementary material

Supplementary material is attached after the article.

Author contributions

The initial idea for using remote sensing data to predict conflict originated from Daniel Racek, Paul Thurner, Göran Kauermann and Xiao Xiang Zhu. Daniel Racek and Göran Kauermann designed the methodology of using various statistical and machine learning models to forecast conflict across the spatial grid. Daniel Racek processed the datasets, conducted the study and analysed the results. Daniel Racek and Brittany Davidson created the visualizations. Daniel Racek, Paul Thurner and Göran Kauermann wrote the first draft of the article. All authors were involved in improving and editing the article.

International Journal of Forecasting 40 (2024) 373–391

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Conflict forecasting using remote sensing data: An application to the Syrian civil war☆

Daniel Racek [a,*], Paul W. Thurner [b], Brittany I. Davidson [c], Xiao Xiang Zhu [d], Göran Kauermann [a]

[a] Institute of Statistics, Ludwig-Maximilians-University Munich, Germany
[b] Institute of Political Science, Ludwig-Maximilians-University Munich, Germany
[c] School of Management, University of Bath, United Kingdom
[d] School of Engineering and Design, Technical University of Munich, Germany

ARTICLE INFO

Keywords:
Remote sensing
Satellite imagery
Conflict prediction
Forecasting
Machine learning
Statistical modeling
Syria

ABSTRACT

Conflict research is increasingly influenced by modern computational and statistical techniques. Combined with recent advances in the collection and public availability of new data sources, this allows for more accurate forecasting models in ever more fine-grained spatial areas. This paper demonstrates the utilization of remote sensing data as a potential solution to the lack of official data sources for conflict forecasting in crisis-ridden countries. We evaluate and quantify remote sensing data's differentiated impact on forecasting accuracy across fine-grained spatial grid cells using the Syrian civil war as a use case. It can be shown that conflict, particularly its onset, can be forecasted more accurately by employing publicly available remote sensing datasets. These results are consistent across a range of established statistical and machine learning models, which raises the hope to get closer to reliable early-warning systems for conflict prediction.

1. Introduction

Conflict prediction has been considered the number one task of peace research for decades (Singer, 1973). According to Hegre, Metternich, Nygård, and Wucherpfennig (2017), early work on conflict prediction was inspired by the works of Richardson (1960), Sorokin (1962), Wright (1965) and influenced by the Correlates of War Project (Small, Singer, & Bennett, 1982), which in 1963 started systematically collecting quantitative data on war, adhering to scientific principles. During this time, the development of early-warning systems for conflict was already one of the central goals of conflict research. Shifting to the 1970s and early 1980s, interest in this space declined, where explicit prediction studies were an exception in the published literature (Hegre et al., 2017). This subsequently changed in the decades thereafter. Although there were early efforts to move away from country–year datasets (Schrodt & Gerner, 2000) such as the Correlates of War Project, in the following decades most of the prediction studies predominantly focused on large-scale country-level events, such as civil wars (Gleditsch & Ward, 2013; Goldstone et al., 2010; Harff & Gurr, 1998; King & Zeng, 2001). Only within the last decade, an increasing number of studies have moved to more fine-grained subnational levels (Bazzi et al., 2022; Hegre et al., 2019; Koren & Bagozzi, 2017), as more and more disaggregated global conflict datasets have become available (Raleigh, Linke, Hegre, & Karlsen, 2010; Sundberg & Melander, 2013). Most recently, a forecasting competition organized by ViEWS (Vesco et al., 2022) showcased the

☆ **Funding Information:** This work is supported by the Helmholtz Association, Germany under the joint research school "Munich School for Data Science - MUDS".
* Correspondence to: Institut für Statistik, Ludwigstr. 33, 80539 München, Germany.
E-mail address: daniel.racek@stat.uni-muenchen.de (D. Racek).

https://doi.org/10.1016/j.ijforecast.2023.04.001
0169-2070/

current advances in the field in both methodology and data sources.

One of the main difficulties in subnational analyses is often the lack of (reliable) structural data sources, such as population density or economic indicators commonly associated with conflict (Blattman & Miguel, 2010; Jerven, 2013). Hence, subnational studies are either conducted using the (low-level) administrative zones of those countries that have sufficient data available (see Bazzi et al., 2022), or are using the well-known 0.5 × 0.5 decimal degree PRIO grid cells, corresponding to an area of roughly 55 × 55 km (at the Equator; Tollefsen, Strand, & Buhaug, 2012). The former are highly country-dependent and thus cannot be easily compared. Administrative zones often vary greatly in size, can be irregular in shape and may change over time, which renders spatial inference more difficult (Wood & Sullivan, 2015). The latter specification via PRIO grid cells, however, is quite coarse, with a cell size of roughly 55 × 55 km, which similarly limits spatial inference. Furthermore, defining cells in decimal degrees means that their size differs by a substantial margin across the world. For example, in Africa alone, the width of the PRIO grid cells differs by up to 11 km based on this definition. But even when using one of those two spatial structures, sufficient data availability for developing and/or crisis-ridden countries are rare, which makes reliable forecasting of conflict in those countries, in which it is needed the most, particularly difficult.

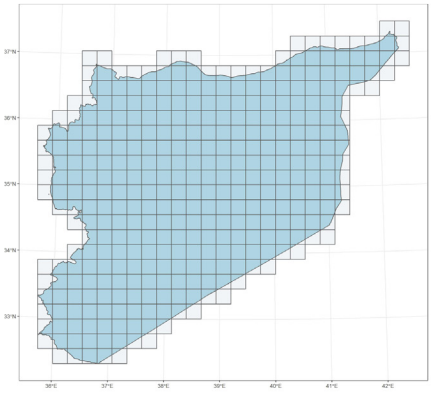
Recently, new and emerging data sources, such as news (Attinà, Carammia, & Iacus, 2022; Mueller & Rauh, 2018), social media (Zeitzoff, 2017), and remote sensing data (Avtar et al., 2021), have increasingly gained attention to solve these problems. This work examines the capabilities of remote sensing data for the task of conflict prediction. Remote sensing data are acquired by applying complex prediction pipelines on sets of high-resolution satellite images. This results in highly fine-grained datasets previously unheard of. Notably, these datasets typically have global coverage, which creates a number of opportunities for conflict research. For instance, as we show in this work, this allows for their use anywhere across the world in custom-defined spatial areas of any size or shape. In recent years, a number of new, high-quality, and high-resolution remote sensing datasets have been made publically available, such as improved global landcover maps (Buchhorn et al., 2020) and vegetation indicators (FAO, 2022). This has become possible due to long-term records of satellite imagery through satellite systems such as AVHRR and MODIS (Pedelty et al., 2007), Landsat Loveland and Dwyer (2012), and Sentinel (Berger, Moreno, Johannessen, Levelt, & Hanssen, 2012), in combination with improvements in classification techniques through, for example, deep learning (Ball, Anderson, & Chan Sr, 2017).

Our work utilizes (novel) remote sensing datasets in order to forecast conflict in self-defined, fine-grained, and regular-sized cells across Syria and tests as well as quantifies their effectiveness for this task. Syria experienced (and to some extent still experiences) one of the largest and deadliest civil wars of the past century, with more than 392,000 recorded fatalities by the end of 2020 (Pettersson et al., 2021). According to scholars, the uprising and the subsequent civil war started with mass protests in the city of Dara'a (Leenders & Heydemann, 2012, p. 142) in March 2011, which quickly escalated due to repression and the use of heavy force exerted by government security forces (Leenders & Heydemann, 2012, p. 149) and ultimately led to war. Although there was a significant decline in violence in 2020, as of today, the war is still ongoing (Human Rights Watch, 2022; Pettersson et al., 2021).

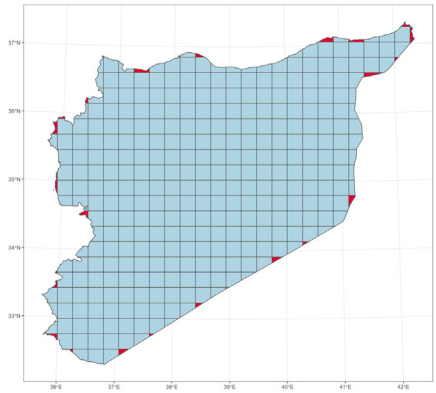
Given Syria's long history of conflict, it presents itself as an ideal use case for examining the potentials of remote sensing data, because the availability of other data sources is sparse to non-existent. To the best of our knowledge, one can likely only obtain the social demographics for the 14 Syrian governorates based on the 2004 census, as well as location polygons of selected ethnic groups. This problem of data limitation, which exists for many developing and crisis-ridden countries, can be alleviated by drawing on remote sensing data sources with global coverage, as motivated in this paper.

In this work, we systematically test the effectiveness of various remote sensing datasets for spatial forecasts of armed conflict across Syria and quantify the change in performance using each data source for this task. Only recently, remote sensing data have been identified as an essential addition in the development of early warning systems (Avtar et al., 2021). Hence, we extend this notion by systematically analyzing their effectiveness for forecasting. Additionally, we further extend current work in the field, as the use of remote sensing data allows us to conduct our forecasting in custom-defined cells. This means that we are not constrained in having to use traditional administrative zones or the PRIO grid cells in our analyses. Instead, we manually construct cells that are more fine-grained than those employed in other studies. These cells are regular and fixed in size across Syria, as they are defined in the Universal Transverse Mercator (UTM) coordinate system. It is worth noting that through this definition, our cells are indeed country-independent. Moreover, all of our employed data sources have global coverage. Hence, applications, extensions, and comparisons to other countries or even continents can easily be undertaken.

Specifically, for our analysis, we rasterize Syria into 25 × 25 km cells and match these cells with various remote sensing datasets. Then, for each cell, we construct aggregated remote sensing variables and use those alongside other traditional predictors to forecast the monthly occurrence of armed conflict. We do this through a one-step-ahead recursive window forecast using a range of established statistical and machine learning models. By repeatedly re-running our models with different specifications, in which we alter the set of included variables, we are able to quantify the effectiveness of each remote sensing dataset with respect to a classical literature-inspired baseline specification. This allows us to evaluate the gain of using remote sensing data for conflict prediction without being reliant on a specific model type. We provide details on our forecasting procedure and chosen models in our methodology section and an in-depth discussion of model selection in our discussion.



(a) The original (uncropped) cells are depicted in grey. The relevant Syrian area is marked in blue.



(b) The 34 removed cells with an area of less than 150 km^2 are depicted in red. The others, which are used in this study, are depicted in blue.

Fig. 1. Illustration of the rasterization and cropping process to construct the 322 Syrian cells used as the main observation units in the study.

We are able to show that by adding remote sensing variables to our baseline, we can consistently improve the overall forecasting performance of our models. By further differentiating conflict into onset and persistence, a distinction frequently made and discussed in the literature (Bazzi et al., 2022; Blattman & Miguel, 2010; Fearon & Laitin, 2003), we show that most of the overall performance increase stems from the former. In other words, utilizing remote sensing data primarily helps to predict new conflicts in areas not suffering from conflicts before. According to our results, population, landcover, and crop data are the most important remote sensing-based predictors of conflict. Our full specification, including variables from each dataset, performs the most consistently well across all types of models. Finally, we want to highlight the generalizability and ease of reproducibility of our study, with which we are answering a recent call by Vesco et al. (2022). Our definition of cells is country-independent and all of our employed data sources are freely available and have global coverage.

The rest of the paper is organized as follows. First, we describe all data sources utilized and how they are processed, followed by a thorough description of our methodological approaches and evaluation criteria. Next, we report the results of our study, before providing an in-depth discussion and a final conclusion.

2. Data

2.1. Constructing the dataset

For our analysis, we rasterize Syria into fine-grained evenly sized 25 × 25 km (625 km^2) grid cells, in which we forecast the monthly occurrence of armed conflict from 2011 to 2020.[1] We chose this time period because 2011 marks the year in which the Syrian civil war broke out, and 2020 is the last year for which the Uppsala Conflict Data Program (UCDP) Georeferenced Event Dataset (GED) provides conflict data. The rasterization takes place in accordance with the UTM coordinate system (zone 36N), a projected coordinate system, in order to avoid cell size distortions. Such distortions would otherwise occur when using any geographic coordinate system (e.g. the World Geodetic System (WGS) 1984 with latitude/longitude coordinates). Additionally, we crop any cells crossing the Syrian border, because for this case study we are only interested in conflict taking place within Syria. Last but not least, we remove any cells with an area of less than 150 km^2 after cropping (roughly a quarter of the full cell size), as those cells almost never experience conflict and hence might skew the results. Nonetheless, we additionally report our results without this filtering in Appendix I. This process results in a total of 356 cells, of which 34 are removed after cropping. The whole process is illustrated in Fig. 1. The resulting 322 cells are our primary observations of interest in this study.

For the time period between 2011 and 2020, we match all subsequently described (remote sensing) datasets based on time and geolocation to our cells of interest and motivate their potential use based on findings from the conflict literature. We emphasize that all data used in this study are freely available and can be directly accessed and downloaded from the web addresses provided in Table A.6 in Appendix A. All our code and the processed datasets can be found on the Open Science Framework (OSF) here.

For our remote sensing data sources, perfect spatial matching is not always possible, as we are matching a raster of cells (consisting of remote sensing data) to our raster of Syrian cells. Hence, in some instances, the cells of the former will overlap with two or more Syrian cells. To solve this issue, we relatively distribute their contribution based on the percentage of covered area of the respective Syrian cells. An illustrative example of this procedure is shown in Fig. 2.

[1] We exclude the Quneitra Governorate, which mainly consists of the Israeli-occupied Golan Heights and the United Nations Disengagement Observer Force (UNDOF) buffer zone.

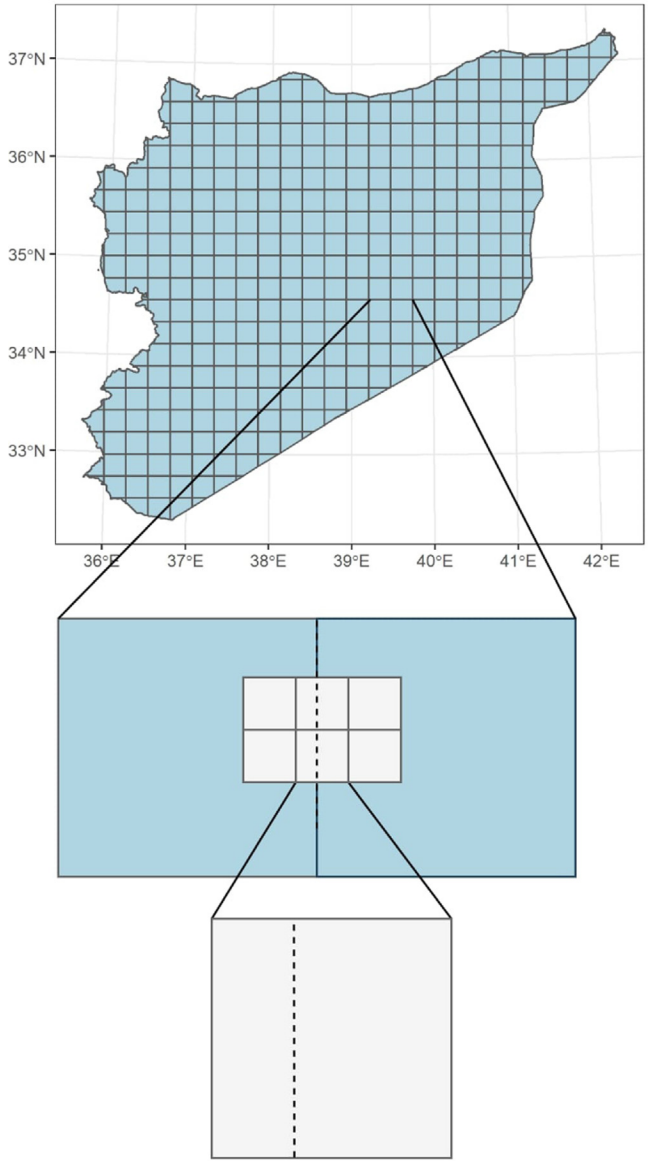


Fig. 2. Illustration of the matching process for overlapping remote sensing data grid cells. The top shows all the resulting Syrian grid cells (in blue) after the rasterization process described in Section 2.1. In the middle, we zoom in on two of those constructed cells and plot six remote sensing data cells (in grey; e.g. consisting of population numbers) onto them. These remote sensing cells (and their data values) are matched to the Syrian grid cells in the matching routine. The four cells on the sides (two on the left, and two on the right) are entirely covered by a single Syrian grid cell and thus matched in their entirety to the respective cell. The two remote sensing cells in the middle are partially covered by both of the Syrian grid cells. Hence, a relative distribution of their data values is necessary. In the bottom, we zoom in on one of those remote sensing data cells. From the dashed line we can infer that one-third of the cell is covered by the left Syrian grid cell, and the remaining two-thirds by the right cell. Hence, in this instance, we would allocate one-third of the data value (e.g. the amount of population) to the left cell and two-thirds to the right cell. The same strategy is applied for all remote sensing datasets.

2.2. Conflict & ethnic data

Data on armed conflict were drawn from the widely known UCDP Georeferenced Event Dataset (GED) (Sundberg & Melander, 2013). The dataset reports events of organized violence, resulting in at least one estimated direct death through armed force, across the world from 1989 to 2020. The data were systematically collected and coded by experienced researchers using national and international news reports, as well as data from NGOs and international organizations. Each event is (among others) assigned a specific date, place, type of violence, and estimated number of resulting fatalities. This study focuses on forecasting battle-related fatalities, i.e. deaths resulting from either state-based or non-state conflict between organized parties. It additionally uses (lagged) one-sided violence, i.e. violence against civilians, as an explanatory variable in order to account for preceding escalatory processes. For a small portion of events, the exact location and/or time point is unknown. We discard these events

in our analysis.[2] As a result, around 15% of the Syrian events from 2011 onwards are discarded. Naturally, this is a limitation in our study. All remaining events are matched based on location and time to the respective cell in the respective month of the study period. Hence, for each cell we have aggregated monthly information on the prevalence of conflict from the beginning of 2011 to the end of 2020.

As noted by scholars (Abosedra, Fakih, & Haimoun, 2021; Ismail, 2011), ethnicities play a central role in the Syrian civil war. Hence, we gather information on the location of "politically relevant ethnic groups" from the Geo-referencing Ethnic Power Relations (GeoEPR) 2021 dataset (Vogt et al., 2015), which is part of the Geographical Research On War, Unified Platform ($GROW^{up}$) (Girardin, Hunziker, Cederman, Bormann, & Vogt, 2015). It assigns every ethnic group to settlement patterns and provides polygons of their location globally from 1946 to 2020. For Syria, the identified (and largest) ethnic groups are the Sunni Arabs, Alawis, Christians, Kurds, and Druze. For each of those groups we create an indicator that describes whether at least 5% of the respective cell area is covered by the respective settlement polygon.

2.3. Remote sensing data

Population data come from the WorldPop project (Tatem, 2017). The project estimates population numbers across the world in 100 m resolution grid cells, with the help of census data and detailed geospatial datasets through a semi-automated dasymetric modeling technique using random forests. In order to ensure that these estimates are as close to reported real-world population numbers, we chose the estimates adjusted to match national UN numbers. Furthermore, because the Syrian population estimates are based on the 2004 census, and the civil war reportedly led to large-scale migration across the country (Kelley, Mohtadi, Cane, Seager, & Kushnir, 2015), we only employ the population numbers stemming from the year 2010 in this study. Additional information on, for example, age and sex structures are only available on a country-wide level and thus are not considered. We aggregate these estimated population numbers for each of the 322 Syrian cells to obtain total amounts. In accordance with earlier studies (Raleigh & Hegre, 2009), we expect areas with higher population numbers to be more likely to experience conflict and thus this information to be highly relevant.

Landcover information is drawn from the Copernicus Global landcover map collection 3 (Buchhorn et al., 2020), which classifies the entire world into 23 different landcover classes at 100 m resolution through an elaborate prediction pipeline (supervised classification, expert rules, temporal cleaning via break detection, etc.) on a yearly basis from 2015 to 2019. We employ some of the "level 1" classes, such as cropland, forest, and permanent water, for which the authors report an average accuracy of 80.6%. Data from each year of the landcover map are matched to the cells accordingly, and average shares for each class (per cell) are derived. These landcover classes contain structural information that might be relevant for the prediction of conflict, for instance, the amount of urban area (related to population, see above; and/or economic activity, see below), crop area (see below), bare area (e.g. desert), tree-covered area (with potential hideouts), or the existence of rivers. The latter might be related to strategically important locations, as rivers might allow for the transportation of weapons and food, and can be used for energy production.

Topography data are collected from (Amatulli et al., 2018). The authors calculate a variety of elevation-based topographic variables such as slope, roughness, and terrain ruggedness for the entire globe using the digital elevation model products of the 250 m Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) (Danielson & Gesch, 2011), which was released in 2010 by the United States Geological Survey and the National Geospatial-Intelligence Agency. They aggregate each variable to a resolution of 1, 5, 10, 50, and 100 km using different spatial aggregation methods. In this study, we employ several topographic variables from the most fine-grained 1 km resolution dataset and derive the respective median of each variable (as also used by the authors for the original aggregate calculation) for each cell. Topography, specifically ruggedness or rough terrain, has been associated with conflict (Collier & Hoeffler, 2004; Fearon & Laitin, 2003), as such terrain arguably provides protection and opportunities to hide for rebels.

Satellite-observed nighttime lights are drawn from the Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS) version 4 nighttime stable lights, which discards ephemeral events and background noise, for the years 2010 to 2013 (Baugh, Elvidge, Ghosh, & Ziskin, 2010) with a 1 km resolution. In earlier studies, nighttime illumination has been shown to be a good indicator of built infrastructure, and thus of economic activity, on a country-wide level (see Elvidge, Hsu, Baugh, & Ghosh, 2014). A more recent study has demonstrated that nighttime lights are also a good predictor of economic wealth at local (within-country) levels (Weidmann & Schutte, 2017). We derive the total amount of nighttime lights for each of our cells and, similar to other studies (Bazzi et al., 2022; Weidmann & Schutte, 2017), calculate a logged per capita value. Generally, countries with persistent conflict are associated with lower per capita gross domestic product (GDP) (Collier, 2004; Pinstrup-Andersen & Shimokawa, 2008). In our specific setting, nighttime lights might point to industry-heavy and wealthier areas, which might again constitute strategically important locations in conflicts and thus improve predictive performance.

Crop production statistics are drawn from MapSPAM (Yu et al., 2020). MapSPAM estimates detailed patterns

[2] This includes all events for which the UCDP variable where_prec is larger than two (i.e. only the second-order administrative division for the location is known) and all events for which date_prec is larger than four (i.e. a day range which is longer than 30 days is reported). All remaining events have a geo-precision of <= 25 km and a reported time period of <= 30 days and thus can be unambiguously assigned to the most likely month/cell combination. Note that there is still some uncertainty left using this assignment, as it is not always guaranteed to be correct.

of crop statistics for 42 different crops in 10 km grid cells across the world. This is achieved through an estimation procedure that combines information on crops (from the lowest available administrative units), land-cover classes, and climate and soil conditions—all three derived from satellite imagery. We employ the most recent (2010) statistics and calculate the total amount of production for all types of crops for each of our cells. According to reports (Eng & Martinez, 2014), rebel areas with agricultural crops were specifically targeted by the Syrian army as a form of punishment by the government. More generally, it is common that conflict parties deliberately destroy infrastructure and resources for food production (Messer & Cohen, 2015) or try to seize cropland in order to secure and guarantee access to food for sustenance (Koren & Bagozzi, 2017).

Daytime temperature recordings are collected from the Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Land Surface Temperature/Emissivity Daily (MOD11A1) version 6.1 (Wan, Hook, & Hulley, 2021). It retrieves daily temperature levels at 1 km resolution across the world from 2000 onwards, using the MODIS thermal infrared channel received by satellite sensors, and is validated by accurate ground-based measurements. In this work, we calculate the average monthly temperature for each cell over the entire study period. Drought (and thus to an extent temperature) has been discussed as a potential contributing factor to the outbreak of the Syrian civil war (Kelley et al., 2015) and has been related to an increase in violence against civilians (Bagozzi, Koren, & Mukherjee, 2017) as well as conflict in general (Von Uexkull, Croicu, Fjelde, & Buhaug, 2016), as it threatens food security. Moreover, we can see drought as a form of local income shock, which can trigger violent mobilization in the case of existing grievances such as ethnic political cleavage (Buhaug, Croicu, Fjelde, & Uexkull, 2021).

Monthly precipitation data come from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS). CHIRPS estimates rainfall maps at 5 km resolution going all the way back to 1981 (Funk et al., 2015) using climatology models, satellite imagery, and local station data. We derive the total amount of precipitation for each cell in a given month over the study period. Like temperature, precipitation is closely related to drought.

Last but not least, we use monthly data on vegetation health from the Food and Agriculture Organization (FAO) of the UN (FAO, 2022). More specifically, we draw on the Vegetation Health Index (VHI) developed by Kogan (1997), which is a composite indicator constructed from the Vegetation Condition Index (VCI) and the Temperature Condition Index (TCI), both derived from Advanced Very High Resolution Radiometer (AVHRR) satellite imagery. The VHI has been used in numerous studies to identify droughts (see for example Kogan, Yang, Wei, Zhiyuan, & Xianfeng, 2005 and Rojas, Vrieling, & Rembold, 2011) and is reported for 10-day periods from 1984 onwards. We derive the monthly average VHI for each cell over the entire study period. We provide an overview on the spatial and (employed) temporal resolution of each remote sensing data source in Table A.7 in Appendix A.

3. Methodology & models

3.1. Variables & model specifications

For our forecasting analysis, we define our target variable as a binary measure, indicating whether there was at least one direct fatality as a result of armed force through either state-based or non-state conflict (battles) in a given cell in a given month. With 322 grid cells, from the beginning of our study period in 2011 to the end of 2020 (=120 months), this adds up to a total of 38,640 observations, of which 6698 (∼17.3%) experienced conflict according to our definition.

Our explanatory variables (predictors) are constructed from the data sources listed in Section 2. All time-varying explanatory variables are lagged by one month or year, in order to reflect a real-world setting. For example, we only have temperature data available for the current month, and not for the month ahead for which we want to forecast conflict. We construct our "zero-model" using only temporal and spatial covariates. The zero-model accounts for all effects relating to the spatial and temporal structure of the civil war and thus should already capture some of the main effects. We obtain our baseline specification by extending the zero-model with a set of covariates capturing both civilian and battle-related fatalities of the past 12 months, in addition to ethnic indicators. The former are generally included in most well-performing forecasting routines (see for example Bazzi et al., 2022; Fritz, Mehrl, Thurner, & Kauermann, 2022; Hegre et al., 2019), whereas the latter are specifically included because of the Syrian use case, as explained above. Hence, our baseline model constitutes a specification a researcher would typically employ when not including any type of special or novel dataset and thus should allow for a fair performance comparison against more complex specifications.

In order to test the effectiveness of different remote sensing data sources for our task at hand, we extend our baseline specification by separately adding variables from each source. Doing this for all of the remote sensing datasets reported in Section 2.3, we obtain a total of eight additional specifications. Finally, we construct a full specification, which adds variables from all of the remote sensing datasets jointly to the baseline. An overview over all 11 tested specifications is provided in Table 1. A detailed table with the exact variables used in each specification is reported in Table C.8 in Appendix C. All specifications are run and their performance evaluated with the forecasting procedure described in the subsequent section.

3.2. Forecasting setup

Forecasting is conducted through a monthly one-step-ahead recursive (expanding) window classification, in which for each month t, we forecast conflict in $t+1$. This means our models are trained on all historical data from the first month t_0 up to the current month t. This is a commonly employed evaluation strategy for time series data, as explained e.g. in Petropoulos et al. (2022). We leave multiple-step-ahead forecasts as a potential future work (forecasting conflict for $t+2$, $t+3$, etc., while

Table 1
Model specifications.

#	Specification	Remote sensing data source
1	Zero-Model	–
2	Baseline	–
3	Baseline + Population	Tatem (2017)
4	Baseline + Landcover Classes	Buchhorn et al. (2020)
5	Baseline + Nighttime Lights	Baugh et al. (2010)
6	Baseline + Topography	Amatulli et al. (2018)
7	Baseline + Vegetation Health	FAO (2022)
8	Baseline + Crops	Yu et al. (2020)
9	Baseline + Precipitation	Funk et al. (2015)
10	Baseline + Temperature	Wan et al. (2021)
11	Baseline + All	All of the above

Notes: Definition of the different specifications tested in this study, with a reference to the respective remote sensing data source used.

being in month t), as this would require training separate models for each step ahead and thus substantially increase computational complexity. With a time period from 2011 to 2020, we have $t = 1, 2, \ldots, 120$. In order to ensure we have enough data to train our models, we skip the first year and start our forecasting procedure in $t = 12$. Because most of the employed models require the specification of hyperparameters, hyperparameter tuning is carried out every 12 steps of the forecast, by training the models on data from t_0 up to $t - 1$ and forecasting for t, optimizing the AUROC (the area under the receiver operator characteristic curve; see Section 3.4 for an exact definition). We decided against tuning our models at each step in order to reduce computational complexity. Moreover, the goal of this work is to analyze and highlight the capabilities of remote sensing data for conflict prediction, rather than optimizing performance scores to the very last digit. As a result, we have a total of nine tuning runs ($108/12 = 9$) per model per specification over the entire study period.

Specifically, the procedure at each step $t = 12, 13, \ldots, 119$ is as follows:

1. Assign all observations from t_0 up to t to the training dataset. Assign all observations in $t + 1$ to the test dataset.
2. If t *mod* $12 = 0$: conduct hyperparameter tuning by repeatedly training the model on data from t_0 up to $t-1$ and evaluating the performance for t. Save the best performing hyperparameters for the model.[3]
3. Load the last hyperparameter specification for the model.
4. Train the model with those hyperparameters on the entire training dataset.
5. Predict (forecast) conflict for all observations in the test dataset.

This procedure is repeated for each of the specifications reported in Table 1 and each of the models listed hereafter.

[3] We define *mod* as the modulo operation, which returns the remainder of the division.

3.3. Models, packages, & hyperparameters

In our study, we employ a range of established statistical and machine learning models in and outside the field. For an in-depth discussion of the selection process, see Section 5. In the following, we list each of the chosen models and name the packages employed to conduct the analysis. Moreover, we report the selected hyperparameters. Hyperparameter optimization is carried out for those parameters that are generally understood to be most vital to the performance and set to default values for the remaining ones, in an attempt to reduce computational complexity. This limitation is necessary, as we are running 11 specifications with nine tuning runs each per model, resulting in a total of 99 tuning runs per model.

The following models are employed in our study:

1. Least absolute shrinkage and selection operator (LASSO): We make use of the R package *glmnet* to fit the logistic LASSO regression. We optimize the only hyperparameter λ, specifying the weight of the penalty.
2. Generalized additive model (GAM): We use the R package *mgcv* for fitting, add selection penalties to our model, and make use of both thin plate splines and P-Splines in our models. See Table C.8 in Appendix C for a complete model description. The selection penalties are optimized during training. Additional hyperparameters, such as the number of knots, are not tuned and are left at their default values.
3. Random forests (RF): We employ the highly optimized R implementation *ranger*. Since RFs rarely overfit when increasing the number of fitted trees, we set this number to be sufficiently large, at 500. We also tested larger numbers of trees, but they did not lead to any relevant changes in performance. We optimize the maximum depth of the trees and the minimum node size. The remaining hyperparameters are left at their default values.
4. Gradient boosting (GB): We employ the commonly used extreme gradient boosting (XGBoost) algorithm and its R implementation *xgboost*. We optimize the number of rounds, the learning rate (eta), and the minimum child weight, whereas the remaining hyperparameters are left at their default values.

See Appendix B for more information about the models, including references for more details.

3.4. Performance evaluation

Our main evaluation criteria are the area under the receiver operator characteristic (ROC) curve (AUROC) and the area under the precision–recall curve (AUPRC), which are described subsequently.

The ROC curve plots the true positive rate (TPR), also known as the recall and defined as the ratio of correctly identified positives ($\frac{TP}{P}$), against the false positive rate (FPR), defined as the ratio of false positives to negatives ($\frac{FP}{N}$). The curve describes the tradeoff between the two

when choosing different classification thresholds for a trained model. This means that the AUROC is between 0 and 1, where 0.5 describes a random classifier and 1 describes a perfect classifier. The advantage of the AUROC is its invariance to the classification threshold. That is, it alleviates the difficult decision of choosing a threshold or thresholds for the performance analysis. Hence, it is often used in general classification scenarios and more specifically in the conflict literature (Bazzi et al., 2022; Hegre et al., 2019; Hegre, Nygård, & Landsverk, 2021).

In imbalanced classification settings, the AUROC can sometimes be misleading, as the focus does not lie in the prediction of the minority (positive) class (see Cranmer and Desmarais (2017) for a discussion on this topic in empirical political research). In our study, we face such an issue, as only ∼16% of our observations experience conflict and thus are assigned to the positive class. In such cases, the area under the precision–recall curve (AUPRC), which again does not require thresholding, is typically used in addition. The precision–recall curve describes the tradeoff between precision, defined as the ratio of true positives over all positive predictions ($\frac{TP}{TP+FP}$), and recall when using different classification thresholds. The AUPRC is similarly between 0 and 1, where 1 describes a classifier that is perfectly able to identify the positive class, and 0 the opposite.

4. Results

Our overall results for the entire study period for our four models and 11 different specifications are reported in Table 2 and Table 3, respectively. The former reports the results using the AUROC, the latter using the AUPRC. In both tables, for each model and given each specification, we report the absolute number of the respective evaluation criteria. Additionally, in brackets, we report the relative difference (in percentage) compared to our baseline specification (using the same model), to describe the performance gain (or loss) of the each specification with respect to our baseline. The ROC curves and PR curves are reported in Appendix I. Bootstrapped confidence intervals for the performance of three of our specifications (zero-model, baseline, and full specification) are also available there.

We start the analysis of our results with the AUROC, i.e. Table 2. Our zero-model specification, which only includes spatial and temporal variables, is arguably already performing well, as the AUROC is between 0.77 (LASSO) and 0.923 (RF). This can be explained by two factors. First, a significant part of Syria is covered by desert (95 out of 322 cells are covered by more than 80% "bare" area according to our landcover map; see Appendix I for an illustration) with little to no inhabitants (mean population of 2224 vs. 82,845 in the remaining cells), in which naturally almost no conflict took place over the study period (1.7% vs. 26.7% of observations). This can be largely captured through our spatial variables. Second, civil wars are generally characterized by location-specific battle lines that might slowly change over time (Raleigh & Hegre, 2009), which can be (partly) captured through a combination of both spatial and temporal variables. We can also see that both GB and RF, which can freely model non-linear effects, are much better at capturing the two aforementioned patterns, compared to our GAM, which can only partially model them, and LASSO, which can only model linear effects.

Moving to our baseline specification, which adds both lagged fatality information as well as ethnic indicators, we can see performance increases compared to the zero-model specification for all models. These increases are rather small for our non-linear models (RF and GB) and significantly larger for the other two (GAM and LASSO), as also confirmed by our bootstrapped confidence intervals (see Appendix I). This means that both RF and GB can model the spatial and temporal structure of the civil war so well that adding additional information on past fatalities and ethnicities only leads to minor improvements in overall performance, whereas both GAM and LASSO profit much more from the addition. In our baseline specification, all models perform well (AUROC $\geq$ 0.91) and the differences between the models are not substantial.

Next, we look into the performance gains when including remote sensing data, our main question of interest in this study. We can see (small) performance increases across the board (compared to our baseline) for the individual specifications using landcover classes, population, and crops, whereas the results for the other remote sensing data sources are more mixed and model-dependent. For both nighttime lights and temperature, we can see (minor) performance gains for GAM, RF, and GB, whereas the LASSO performance is on par with the baseline. In other cases, for instance for topography, the inclusion leads to either negligible performance increases or even small decreases, depending on which model performance we look at. For topography, the highest increase is achieved by GB with only +0.26%, and we can spot a decrease for RF (−0.21%). Patterns such as this can occur if the dataset itself does not contribute a lot of relevant information to our task at hand, i.e. forecasting conflict. Then, the inclusion can possibly lead to overfitting and performance will fall off, as our models will pick up random signals (noise) from the respective dataset during training. Although both GB and GAM show minor performance gains for topography, we can arguably still conclude that our chosen topography variables contain little to no relevant information to predict conflict, at least without combining it with additional remote sensing information (from other data sources). Similar arguments can be made for vegetation health and precipitation. Finally, our full specification (Baseline + All), including variables from all remote sensing data sources, performs consistently better than our baseline (0.37%–1.75%) but (for GB and RF) slightly worse than some of the individual specifications, again hinting at the fact that some remote sensing data sources (or the combination of them) are unnecessarily included. Overall, we can conclude that the inclusion of remote sensing data (marginally) increases our predictive performance in terms of the AUROC.

As noted above, the AUROC can be misleading, particularly if a researcher is interested in identifying the minority class, the prevalence of conflict in our setting.

Table 2
AUROC performance (all observations).

Specification	AUROC for model			
	GAM	LASSO	RF	GB
Zero-Model	0.850 (−6.68%)	0.767 (−16.37%)	0.923 (−0.62%)	0.917 (−0.7%)
Baseline	0.910	0.917	0.929	0.923
Baseline + Landcover Classes	0.912 (+0.23%)	0.922 (+0.53%)	0.933 (+0.43%)	0.927 (+0.39%)
Baseline + Population	0.915 (+0.49%)	0.919 (+0.26%)	0.933 (+0.51%)	0.929 (+0.67%)
Baseline + Nighttime Lights	0.919 (+0.99%)	0.917 (+0%)	0.931 (+0.22%)	0.926 (+0.25%)
Baseline + Topography	0.911 (+0.06%)	0.917 (+0%)	0.927 (−0.21%)	0.926 (+0.26%)
Baseline + Vegetation Health	0.912 (+0.19%)	0.917 (−0.02%)	0.929 (+0.04%)	0.926 (+0.35%)
Baseline + Crops	0.916 (+0.63%)	0.919 (+0.23%)	**0.933 (+0.46%)**	**0.930 (+0.74%)**
Baseline + Precipitation	0.916 (+0.58%)	0.917 (+0.02%)	0.928 (−0.05%)	0.924 (+0.12%)
Baseline + Temperature	0.917 (+0.73%)	0.917 (+0%)	0.929 (+0.02%)	0.925 (+0.19%)
Baseline + All	**0.926 (+1.75%)**	**0.923 (+0.73%)**	0.932 (+0.37%)	0.929 (+0.59%)

Notes: Average area under the receicer operator characteristics curve (AUROC) performance for one-step ahead forecasts over the entire forecasting horizon of the different model specifications and types. For an explanation of the AUPRC, see Section 3.4. Each row reports the performance of one specification (details in Section 3.2), each column of one type of model (details in Section 3.3). In brackets we report the relative performance difference to our baseline specification (of the same model), except for the baseline specification itself. The best performing specification for each model is highlighted in bold.

Table 3
AUPRC performance (all observations).

Specification	AUPRC for model			
	GAM	LASSO	RF	GB
Zero-Model	0.657 (−16.84%)	0.499 (−35.76%)	0.784 (−3.11%)	0.779 (−2.23%)
Baseline	0.790	0.777	0.810	0.796
Baseline + Landcover Classes	0.794 (+0.49%)	0.786 (+1.11%)	0.821 (+1.43%)	0.805 (+1.12%)
Baseline + Population	0.796 (+0.68%)	0.781 (+0.5%)	**0.822 (+1.5%)**	0.809 (+1.64%)
Baseline + Nighttime Lights	0.798 (+1%)	0.777 (+0%)	0.815 (+0.64%)	0.802 (+0.74%)
Baseline + Topography	0.792 (+0.24%)	0.774 (−0.46%)	0.802 (-1%)	0.802 (+0.65%)
Baseline + Vegetation Health	0.794 (+0.51%)	0.773 (−0.52%)	0.815 (+0.72%)	0.807 (+1.35%)
Baseline + Crops	0.797 (+0.9%)	0.781 (+0.47%)	0.822 (+1.49%)	**0.813 (+2.08%)**
Baseline + Precipitation	0.797 (+0.88%)	0.774 (−0.46%)	0.813 (+0.47%)	0.799 (+0.34%)
Baseline + Temperature	0.797 (+0.85%)	0.777 (−0.02%)	0.814 (+0.5%)	0.804 (+0.95%)
Baseline + All	**0.805 (+1.91%)**	**0.789 (+1.51%)**	0.817 (+0.95%)	0.812 (+1.93%)

Notes: Average area under the precision–recall curve (AUPRC) performance for one-step ahead forecasts over the entire forecasting horizon of the different model specifications and types. For an explanation of the AUPRC, see Section 3.4. Each row reports the performance of one specification (details in Section 3.2), each column of one type of model (details in Section 3.3). In brackets we report the relative performance difference to our baseline specification (of the same model), except for the baseline specification itself. The best performing specification for each model is highlighted in bold.

Hence, Table 3 reports the AUPRC of the different specifications for our forecasting task. First of all, we can see that the difference between the zero-model and the baseline is significantly larger than before. This makes sense in that, for our positive-class (conflict) observations, knowing and modeling the location of the Syrian Desert will reduce the number of false positives (FPs) and hence increase the precision, but it will not increase the number of true positives (TPs) and thus the recall. For the latter, knowing past conflict through information on lagged fatalities will most certainly have a positive effect, hence the larger performance difference between the zero-model and the baseline. Notably, this difference is not as large for both RF and GB, as both models can more easily capture the dynamics of the civil war through non-linear combinations of both spatial and temporal variables.

Moving to the remote sensing specifications, we see similar results as above. We report increases for landcover classes, population, crops, and nighttime lights. Notably, the relative increases are around double in percentage points compared to earlier and range up to 2.08% (GB with crops). For vegetation health, precipitation, and temperature we similarly report performance gains, with the exception for LASSO, whereas topography is more mixed, as it again shows decreases for both LASSO and RF. Our full specification performs consistently well (0.95%–1.93%) and is only marginally outperformed by some individual specifications for RF and GB.

By setting a probability threshold, we can analyze the performance increase of our full specification compared to the baseline in individual observation numbers. For reasons of simplicity, we do not tune this threshold and instead set it to the standard value of 0.5. Doing this, out of 34,776 conflict observations in the forecasting test sample, the GAM is able to predict 31,266 correctly (+89 compared to the baseline), and LASSO, RF, and GB correctly predict 31,147 (+20), 31,452 (+53), and 31,330 (+78), respectively. Note that this does not necessarily mean that we are able to correctly forecast the same instances as with our baseline. Moreover, we can very likely achieve better performance by tuning the probability threshold.

Next, we explore where some of these performance increases stem from. By differentiating our observations

Table 4
AUPRC performance for conflict onset observations.

Specification	AUPRC for model			
	GAM	LASSO	RF	GB
Zero-Model	0.241 (−26.79%)	0.161 (−48.49%)	0.357 (−1.49%)	0.323 (−1.9%)
Baseline	0.329	0.312	0.362	0.330
Baseline + Landcover Classes	0.336 (+1.94%)	0.324 (+4%)	0.384 (+6.03%)	0.341 (+3.61%)
Baseline + Population	0.337 (+2.5%)	0.318 (+2.11%)	0.38 (+4.88%)	0.352 (+6.84%)
Baseline + Nighttime Lights	0.331 (+0.39%)	0.312 (+0.02%)	0.37 (+2.07%)	0.336 (+1.82%)
Baseline + Topography	0.335 (+1.59%)	0.308 (−1.13%)	0.365 (+0.7%)	0.346 (+5.1%)
Baseline + Vegetation Health	0.336 (+1.9%)	0.309 (−0.99%)	0.364 (+0.53%)	0.338 (+2.59%)
Baseline + Crops	0.336 (+2.05%)	0.317 (+1.77%)	0.378 (+4.37%)	**0.359 (+8.83%)**
Baseline + Precipitation	0.338 (+2.61%)	0.308 (−1.06%)	0.361 (−0.33%)	0.334 (+1.28%)
Baseline + Temperature	0.334 (+1.55%)	0.312 (+0%)	0.362 (+0.01%)	0.342 (+3.66%)
Baseline + All	**0.347 (+5.42%)**	**0.327 (+4.9%)**	**0.386 (+6.66%)**	0.356 (+7.91%)

Notes: Average area under the precision–recall curve (AUPRC) performance for one-step ahead forecasts over the entire forecasting horizon of the different model specifications and types. The sample is limited to all conflict onset observations as described in the text. For an explanation of the AUPRC, see Section 3.4. Each row reports the performance of one specification (details in Section 3.2), each column of one type of model (details in Section 3.3). In brackets we report the relative performance difference to our baseline specification (of the same model), except for the baseline itself. The best performing specification for each model is highlighted in bold.

into two categories, we can analyze how well we are able to predict conflict onset vs. conflict persistence. We define our conflict onset observations as observations with no conflict in the current month (t). Hence, analyzing the AUPRC and thus the (minority) conflict class in the next month ($t + 1$) means that we are analyzing how well we are able to predict the outbreak of conflict from one month to the next. This category includes 28,141 observations, of which 1988 experience such an outbreak. We define our conflict persistence observations as those observations that experience conflict in the current month (t). Similarly, by analyzing the AUPRC, we are analyzing how well we are able to predict the persistence of conflict from one month to the next. This category comprises the remaining 6635 observations, of which 4652 experience persistent conflict according to our definition. Based on this distinction, we recalculate the AUPRC from Table 3 for all specifications and report the results in Table 4 (onset) and Table 5 (persistence). By comparing the two tables, we can immediately spot the performance difference in forecasting conflict onset vs. persistence. Predicting "new" conflict is generally more difficult (AUPRC $\sim$ 0.35) than predicting the continuance of it (AUPRC $\sim$ 0.9).

By comparing the performance for conflict onset across our specifications (Table 4), we can see the clear positive impact of including remote sensing data sources. For example, the inclusion of landcover classes leads to performance increases of 1.94% to 6.03%. We can see mixed results for the individual specifications regarding topography, vegetation health and precipitation, and (substantial) performance increases for landcover classes, population, nighttime lights, crops, and temperature. The combination of all remote sensing data sources leads to the highest increases with respect to the baseline for the GAM (5.42%), LASSO (4.9%), and RF (6.66%), and to the second-highest increase for GB (7.91%). Hence, when it comes to the prediction of conflict onset, the inclusion of remote sensing data increases the performance considerably, and a combination of different data sources seems to works well.

Moving to Table 5 and thus the results for conflict persistence, the performance gains when adding remote sensing data to our baseline are much more moderate. Similar data sources seem to perform well and not so well. The full model leads to consistent performance increases, but they are much smaller compared to conflict onset (0.62%–1.61%).

5. Discussion

In this work, we set out to test the effectiveness of various remote sensing datasets for conflict prediction. A number of key findings can be inferred from our results. First and foremost, our results confirm that remote sensing data help to increase overall predictive performance according to both the AUROC (up to 1.75% for the full specification) and AUPRC (1.93%). The overall increases seem rather small at first glance, but this was to be expected due to three reasons. First, our baseline performed well from the outset (AUROC $\geq$ 0.91; AUPRC $\geq$ 0.78), which is a common finding in the literature that is reinforced for our Syrian case study (e.g. see Bazzi et al., 2022; Hegre et al., 2019. Hence, large increases in performance are typically difficult to achieve and much less expected. Second, our marginal performance gains are in line with (and sometimes above) those reported in studies across the literature, all of which extend a baseline specification consisting of lagged fatality information by additional predictors (AUROC increase of $\sim$1.6% in Bazzi et al., 2022; $\sim$1.2% increase in AUROC in Hegre et al., 2019; $\sim$0.8% decrease in MSE (=performance increase) in Mueller & Rauh, 2022). Third, our baseline is actually a richer model than those employed as a baseline in the cited studies, as we additionally include spatial, temporal, and ethnic variables. Therefore, our results being consistent with current literature, despite employing this richer baseline, demonstrates the additional value of remote sensing data impressively. Fourth, even when removing contextual information such as lagged fatalities and ethnicities from the model (=zero-model), remote sensing

Table 5
AUPRC performance for conflict persistence observations.

Specification	AUPRC for Model			
	GAM	LASSO	RF	GB
Zero-Model	0.869 (−4.39%)	0.825 (−7.16%)	0.894 (−1.97%)	0.886 (−1.89%)
Baseline	0.909	0.889	0.912	0.903
Baseline + Landcover Classes	0.911 (+0.3%)	0.895 (+0.67%)	0.923 (+1.19%)	0.912 (+0.96%)
Baseline + Population	0.912 (+0.34%)	0.891 (+0.24%)	0.924 (+1.31%)	0.914 (+1.25%)
Baseline + Nighttime Lights	0.912 (+0.41%)	0.889 (+0%)	0.917 (+0.56%)	0.909 (+0.69%)
Baseline + Topography	0.910 (+0.09%)	0.886 (−0.29%)	0.904 (−0.84%)	0.906 (+0.34%)
Baseline + Vegetation Health	0.911 (+0.24%)	0.886 (−0.33%)	0.919 (+0.82%)	0.915 (+1.31%)
Baseline + Crops	0.913 (+0.42%)	0.891 (+0.22%)	**0.924 (+1.32%)**	0.918 (+1.58%)
Baseline + Precipitation	0.912 (+0.38%)	0.886 (−0.28%)	0.917 (+0.59%)	0.907 (+0.41%)
Baseline + Temperature	0.912 (+0.33%)	0.889 (−0.02%)	0.917 (+0.6%)	0.911 (+0.9%)
Baseline + All	**0.915 (+0.64%)**	**0.898 (+1.05%)**	0.918 (+0.62%)	**0.918 (+1.61%)**

Notes: Average area under the precision–recall curve (AUPRC) performance for one-step ahead forecasts over the entire forecasting horizon of the different model specifications and types. The sample is limited to all conflict persistence observations as described in the text. For an explanation of the AUPRC, see Section 3.4. Each row reports the performance of one specification (details in Section 3.2), each column of one type of model (details in Section 3.3). In brackets we report the relative performance difference to our baseline specification (of the same model), except for the baseline itself. The best performing specification for each model is highlighted in bold.

data provide clear performance gains when they are included (see Appendix D). Overall, we can conclude that remote sensing data indeed provide an additional source of information relevant for the prediction of conflict.

Although we did not initially set out to differentiate between the onset and persistence of conflict, during our performance analysis we identified that remote sensing data are particularly important for correctly predicting the onset of conflict. Generally, predicting conflict onset is considered a much more difficult task in the conflict literature, as similar studies (Hegre et al., 2021; Mueller & Rauh, 2022), as well as our results (lower AUPRC), show. Providing additional information through remote sensing data sources turns out to be particularly important for this challenging task. Notably, our full specification, which includes all remote sensing data sources, performs the best with one exception (GB; but in second place here), which supports our argument. By inspecting the onset results more closely (see Appendix F), we can identify that the prediction task (expectedly) becomes more and more difficult as the number of months since last conflict increases (lower AUPRC). At the same time, the relative performance increase of the full specification becomes larger and larger. Hence, according to our results, the importance of remote sensing data increases, as the time that has passed since last conflict becomes longer.

This leads to our third observation. Depending on the model and setting (onset vs. persistence), particular specifications perform better or worse, respectively. There are a few implications to draw from this finding. First, these results point to the fact that training and using individual models for onset and persistence, respectively (possibly with different data sources or variables included), might improve forecasting performance. To the best of our knowledge, this has not been considered in the literature yet. Second, including all remote sensing data sources into the models might not necessarily be the best choice. For our non-linear models (RF and GB), individual specifications (with only one remote sensing data source included) at times outperform the full specification (with all of them included). Notably, these patterns are not consistent across models. This makes sense in that regard, insofar as different models vary in their ability to extract relevant information for the conflict prediction task. Both RF and GB are in theory able to extract (highly) non-linear relationships between the target variable (conflict) and our explanatory variables, including possible interactions, whereas both the GAM and LASSO are only partially able to or are unable to, respectively. Hence, for the former two models, the same variables offer more possibly relevant information (e.g. reflected in the higher performance of the baseline), which makes it more likely that some of the remote sensing data sources are rendered redundant, and thus that their inclusion leads to overfitting and a performance decrease. As a consequence, researchers not only need to be careful what variables or data sources they consider, but also need to take into account the model they intend to use when making any decisions on variable inclusions. Future works could consider automating this data-driven process to achieve the best possible performances. For example, one could pursue a forward selection process, in which data sources and/or specific variables are continually added to the model in a systematic fashion (starting from the baseline) as long as the performances are increasing. A researcher could even go so far as to perform this forward selection and thus adjust the included variables for each step of the study period (i.e. every time a new model is trained) in order to achieve optimal results across the entire forecasting horizon. Note that such a selection approach would need to be treated in a similar fashion as hyperparameter tuning; i.e. we need to guarantee a true out-of-sample performance evaluation.

Next, we investigated the performance of our models over time (see Appendix E for the results and a more thorough discussion). We conclude that both the GAM and to some extent LASSO profit considerably from the inclusion of remote sensing data early into the study period, where less training data are available. Closer to the end of the study period, the AUPRC performance starts to drop for all models and specifications, as conflict events are thinning out and the task of correctly identifying conflict becomes

considerably more difficult. Nonetheless, the relative performance gains from including remote sensing data are largely steady throughout, strengthening the confidence of our findings.

Moreover, we want to contribute to the ongoing discussion in the literature on the tradeoff between explanation and prediction (Hegre et al., 2017) by highlighting that our model performances do not substantially differ. Although there is a clear ranking in terms of model performance (RF > GB > GAM > LASSO), once we arrive at our baseline specification, the performance difference in both AUROC (up to 1.3%) and AUPRC (4.2%) between the models for any specification is much lower than one might expect. Notably, for our full specification, these differences further decrease. Hence, a researcher could easily fall back on using one of the inherently interpretable models, such as LASSO or the GAM, without giving up or forgoing substantial performance gains. Note that in-depth analyses in this setting remain a difficult endeavor, even with interpretable models, as we are re-training our models for each step of the study period (e.g. here, 108 different GAMs for one specification only). Moreover, identifying causal chains of effects additionally requires distinct variable setups and theoretical considerations. Nonetheless, the small performance gap reported here seems promising.

This brings us to our sixth point of discussion: our selection of models. Drawing from the latest ViEWS conflict prediction competition (Vesco et al., 2022), and similar studies such as (Bazzi et al., 2022; Hegre et al., 2019), we decided on a versatile set of commonly employed models, ranging from classical statistical to common machine learning models, in order to ensure that our results are consistent across a variety of different models. All chosen models (GAM, LASSO, RF, and GB) have proven themselves to perform well in the conflict forecasting domain (see the cited studies) as well as across a variety of other fields (Bastin et al., 2019; Chaudhary, Richardson, Schoeman, & Costello, 2021; Fabbri et al., 2020; Fife & D'Onofrio, 2022; Greener, Kandathil, Moffat, & Jones, 2022; Rustam et al., 2020; Schroeders, Schmidt, & Gnambs, 2022; Xie & Zhu, 2020). Our set of chosen models consists of two "simpler" models with the GAM and LASSO, which have a limited capacity to model non-linear and interaction effects but remain easier to analyze and interpret. On the other hand, with RF and GB, we chose two models that can freely model non-linear and interaction effects, but are consequently much more difficult to analyze and interpret (and hence oftentimes considered black-box models), with GB additionally being more prone to overfitting. We refrained from including neural networks, as they require large amounts of data to train and are generally outperformed by classical machine learning models on tabular datasets such as the one here (Borisov et al., 2021). Notably, redefining our observations in the form of images and implicitly taking into account the spatial structure of each cell and their surroundings—and thus taking advantage of the high resolution the remote sensing datasets offer, and moving away from the tabular structure—through convolutional neural networks (CNNs) might be a viable future path. Moreover, we decided against including an ensemble of different individual models, as for example pursued in Bazzi et al. (2022) and in the ViEWS forecasting competition (Vesco et al., 2022), since such an ensemble is simply a weighted combination of individual models. Hence, performance increases in the individual models (as reported here across the board) are very likely similarly reflected in the performance of the ensemble.

Last but not least, we want to briefly reflect on the performance and thus importance of different remote sensing datasets. According to our individual results, landcover classes, population, and crop data provide the highest and most consistent overall performance increases, whereas the remaining datasets seem less important. Nonetheless, once we look at the results for conflict onset, most, if not all, of the remote sensing datasets seem to provide relevant information for the prediction of conflict. This is confirmed by the fact that our full specification performed the best in three out of four models. Moreover, by looking at the variables selected by LASSO, we can see that in our full specification, each remote sensing dataset is selected in more than half of the model fits (see Table G.14 in Appendix G), with landcover classes, population, and crop data being selected most often. Finally, by inspecting the feature importance scores for RF (full specification, see Fig. H.3 in Appendix H), we can corroborate this pattern. While our lagged battle-related features seem to be most important for the prediction, both landcover classes and population are not too far behind. On average, each remote sensing dataset contributes to the model performance according to these importance scores. Hence, we conclude that each remote sensing data source seems to provide relevant information for the prediction of conflict, but some of them are more important than others.

6. Conclusion

We tested the effectiveness and capabilities of remote sensing data for conflict prediction in the context of the Syrian civil war. Using remote sensing data enabled us to conduct our study in self-defined, fine-grained, and evenly sized spatial cells across Syria. Our results confirmed that including a variety of remote sensing datasets consistently improved forecasting performance compared to a rich baseline independent of the chosen prediction model. As our analysis showed, a large portion of this performance gain came from correctly identifying the onset of conflict. We conclude that remote sensing data can and indeed should be used to forecast conflict in countries with a lack of reliable official data sources.

Future work could try to take advantage of the fine-grained spatial structure of the remote sensing datasets through specific modeling techniques. Moreover, as more and more high-quality datasets are published, identifying causal effects for conflict on a subnational level might prove to be possible. Finally, evaluating other emerging data sources such as news or social media, and finding ways to combine disparate and emerging datasets into a joint model, may further improve forecasting performance. This provides an interesting set of data sources, methods, and approaches to further advance conflict research globally.

CRediT authorship contribution statement

D.R., P.W.T., X.X.Z. and G.K. conceived the research. P.W.T., B.I.D., X.X.Z. and G.K. supervised the research. D.R., P.W.T. and G.K. designed the methodology. D.R. collected & processed the data, conducted the study and analyzed the results. D.R. and B.I.D. created the visualizations. D.R., P.W.T. and G.K. wrote the first draft. D.R., P.W.T., B.I.D. and G.K. edited and all authors approved the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Data sources

See Tables A.6 and A.7.

Appendix B. Model descriptions

Here, we provide concise descriptions of the models employed in this study. For details we refer the reader to the respective publications.

1. Least absolute shrinkage and selection operator (LASSO): Using LASSO (Tibshirani, 1996), we fit a simple (logistic) regression model in conjunction with an L_1-penalty on the coefficients. This shrinks the model coefficients compared to a standard regression. Additionally, some of the coefficients (those for "less important" variables) are set to exactly 0, which resembles a feature selection. Generally, the penalty reduces the generalization error of the fitted model. Using LASSO, we can only capture linear effects of the included covariates. Hence, any non-linear effects or interactions of covariates need to be explicitly included in the model through transformations, as done in standard regression models.
2. Generalized additive model (GAM): A GAM (Hastie, 2017) is a generalized linear model (in our case a logistic regression) with an additional set of linearly included unknown smooth functions of (some of the) explanatory variables. The set of smooth functions is chosen by the user, in terms of both the included variables and the types of functions. Commonly chosen smooth functions are, for example, thin plate regression splines (Wood, 2003), cubic regression splines (Durrleman & Simon, 1989), and P-splines (Eilers & Marx, 1996). By including these smooth terms into the model, the GAM is able to model non-linear effects of the chosen variables. Similar to LASSO, interactions need to be explicitly included in the model.
3. Random forests (RF): RF (Breiman, 2001) is an ensemble consisting of multiple (decision) trees. The overall prediction is the average of the individual predictions over all trees. RFs typically employ bagging and a random selection of the features in order

Table A.6
Data sources with web addresses for download.

Data	Source	Downloaded from:
Conflict	Sundberg and Melander (2013)	https://ucdp.uu.se/downloads/
Ethnicity	Vogt et al. (2015)	https://icr.ethz.ch/data/epr/geoepr/
Population	Tatem (2017)	https://hub.worldpop.org/project/categories?id=3
Landcover Classes	Buchhorn et al. (2020)	https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_Landcover_100m_Proba-V-C3_Global
Nighttime Lights	Baugh et al. (2010)	https://developers.google.com/earth-engine/datasets/catalog/NOAA_DMSP-OLS_NIGHTTIME_LIGHTS
Topography	Amatulli et al. (2018)	http://www.earthenv.org/topography
Vegetation Health	FAO (2022)	https://data.apps.fao.org/map/catalog/srv/eng/catalog.search#/metadata/84e27651-0bb4-4a26-8b4a-2b10bbccb7e0
Crops	Yu et al. (2020)	https://www.mapspam.info/data/
Precipitation	Funk et al. (2015)	https://developers.google.com/earth-engine/datasets/catalog/UCSB-CHG_CHIRPS_DAILY
Temperature	Wan et al. (2021)	https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MOD11A1

Table A.7
Remote sensing data sources temporal & spatial resolution.

Data	Available temporal resolution	Temporal resolution & time period employed	Spatial resolution
Population	Yearly	Fixed (2010)	100 m
Landcover Classes	Yearly	Yearly from 2015–2019, fixed before & after	100 m
Nighttime Lights	Yearly	Yearly from 2010–2013	1 km
Topography	None	Fixed	1 km
Vegetation Health	10 days	Monthly from 2010–2020	1 km
Crops	10 years	Fixed (2010)	10 km
Precipitation	Daily	Monthly from 2010–2020	5 km
Temperature	Daily	Monthly from 2010–2020	1 km

to reduce the correlation between the trees and thus the variance of the ensemble. RF is able to freely model non-linear effects and the interactions between the variables without the requirement of including any of them explicitly.

4. Gradient boosting (GB): In boosting, we construct an ensemble of weak prediction models, typically (decision) trees, and iteratively apply the model to modified versions of the training data. After each iteration, misclassified inputs are assigned higher weights such that they are focused on in the next training iteration. GB (Friedman, 2001) mimics this process by viewing the boosting algorithm as iterative functional gradient descent and "nudging" the model prediction function step by step, closer to the real data points. An efficient and scalable implementation of this technique is XGBoost (Chen & Guestrin, 2016), which is employed in this study. GB can freely model non-linear effects and the interactions between the variables without including any of them explicitly.

Appendix C. Specifications

See Table C.8.

Appendix D. Additional model specification results

The following two tables expand on our main results by reporting the performance of an additional specification that extends the zero-model with all remote sensing data sources (Zero-Model + All). Hence, the results of Table D.9 correspond to Table 2 and Table D.10 corresponds to Table 3 in the main text. Notably, we only compare the performance between these two specifications. The results show that even without including information on lagged fatalities and ethnicities, remote sensing data increase predictive performance in both AUROC and AUPRC. The performance increases are particularly large for the "simpler" models (GAM and LASSO) and smaller for the other two (RF and GB). Arguably, both RF and GB already learn most of the relevant remote sensing information implicitly through a (non-linear) combination of spatial and temporal variables (e.g. more populous areas experience more conflict during certain time periods). Nonetheless, even without additional contextual information (lagged fatalities, ethnicities), we see consistent performance increases when adding remote sensing data to the respective model across all four models.

Appendix E. Performance over time

Here, we provide additional insights into the performance of our different models and specifications over the study period. Table E.11 splits up the AUROC performance reported in Table 2 in the main text into separate scores for each year for both the baseline and the full specification. Table E.12 does the same for the AUPRC performance reported in Table 3.

Table E.11 shows that both the GAM and LASSO (and to a smaller extent, RF) considerably benefit from the inclusion of remote sensing data in the first year of the analysis period, as indicated by the large performance gains in AUROC. Apparently, the "simpler" models particularly struggle with the lack of historical information when only using baseline features, but most of this performance gap can be made up for by the inclusion of remote sensing data. In the remaining years, the performance increases from baseline to full specification are mostly steady (with some smaller fluctuations). Notably, we see a small drop in performance from 2016 onwards. We attribute this drop to the changing circumstances (Russian support in form of airstrikes, pushing back of the Islamic State by both the Kurds and the Assad regime, and Turkish offensives in northern Syria) that are not immediately picked up by the models. Nonetheless, we want to highlight that our performance gains from remote sensing data are (mostly) consistent even throughout this period.

Similarly, Table E.12 shows considerable AUPRC performance gains for the GAM with the inclusion of remote sensing in the first year. Again, we can identify a drop in performance in 2016, from which the performance continues to fall off. While we can attribute some of this performance drop to the changing circumstances described above, after 2018 especially, the task of correctly identifying conflict becomes considerably more difficult, as the civil war moves more towards the border regions next to Turkey and Iraq, and fewer cells experience conflict. In 2018, only 13% of the observations suffer from conflict (see the second column in the table), compared to 25.2% in 2015. In 2020 we are down to only 9.3%. The increased difficulty in the prediction task that results from this is distinctly reflected across all models and specifications, with a considerably lower AUPRC of around 0.55. Nonetheless, on average, the performance increase from the use of remote sensing continues to hold even throughout these years.

Appendix F. Detailed onset performance

The following table offers a more detailed view on the performance results for conflict onset, by differentiating observations by the number (#) of months since the last conflict took place in the respective cell. Overall, we can see that the prediction task becomes more and more difficult as the number of months increase (i.e. AUPRC decreases as we move down the rows), across all models. At the same time, the relative performance increase of the full specification increases substantially. Even when leaving out the first row (conflict persistence), the performance increase rises by 7 (GB) to 16 (RF) percentage points when moving down to the last row (months since last conflict $\geq$ 6). We can observe this pattern consistently across all models. Notably, for GB we can spot a performance decrease with respect to the baseline twice (for 3 and 5 months). We attribute this effect to potential overfitting, since the number of observations is quite low (1259 and 773 vs. 23,202 for $\geq$ 6 months), GB is particularly prone to overfitting and we cannot observe the same pattern for any of the other models. Hence, overall, we can conclude that remote sensing data become more and more important for the prediction of conflict as the time since last conflict increases (see Table F.13).

Table C.8
Specifications with covariates.

#	Specification	Included covariates
1	Zero-Model	- Time Trend: integer GAM: Included using a P-Spline - Monthly dummies (11): binary - Cell size in km^2: numeric - Latitude: numeric GAM: Inclusion see Longitude - Longitude: numeric GAM: Included with Latitude using a thin plate spline - Distance to capital in km: numeric
2	Baseline	- All from specification 1) - Alawi dummy (share $>$ 0.05): binary - Christians dummy (share > 0.05): binary LASSO: Included as interaction with Time Trend GAM: Included as interaction with Time Trend using a P-Spline - Druze dummy (share > 0.05): binary LASSO: Included as interaction with Time Trend GAM: Included as interaction with Time Trend using a P-Spline - Kurds dummy (share > 0.05): binary LASSO: Included as interaction with Time Trend GAM: Included as interaction with Time Trend using a P-Spline - Sunni Arabs dummy (share > 0.05): binary LASSO: Included as interaction with Time Trend GAM: Included as interaction with Time Trend using a P-Spline - # of fatalities through battle last month: integer - # of fatalities through battle last 12 months: integer - # of month since last fatality through battle: integer - # of civilian fatalities last month: integer - # of civilian fatalities last 12 months: integer - # of month since last civilian fatality: integer
3	Baseline + Population	- All from specification 2) - Total amount of population (logged): numeric
4	Baseline + Landcover Classes	- All from specification 2 - Share crop area: numeric - Share bare area: numeric - Share built area: numeric - Share grass & shrub area: numeric - Permanent water dummy (share $>$ 0.01): binary - Tree-covered area dummy (share $>$ 0.01): binary
5	Baseline + Nighttime Lights	- All from specification 2) - Total amount of stable lights per person (logged): numeric
6	Baseline + Topography	- All from specification 2) - Elevation (median): numeric - Slope (median): numeric - Vector ruggedness measure (median): numeric
7	Baseline + Vegetation Health	- All from specification 2) - Vegetation health index (average): numeric
8	Baseline + Crops	- All from specification 2) - Total amount of food crops (logged): numeric - Total amount of non-food crops (logged): numeric
9	Baseline + Precipitation	- All from specification 2) - Total amount of precipitation: numeric
10	Baseline + Temperature	- All from specification 2) - Day temperature (average): numeric
11	Baseline + All	- All variables

Notes: If not noted otherwise, all variables are included linearly without any interaction in both LASSO and the GAM.

Appendix G. Lasso model selection

See Table G.14.

Appendix H. Feature importance

See Fig. H.3.

Appendix I. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2023.04.001. There, we 1) provide further information on the Syrian desert cells, 2) provide the ROC curves & PR curves, 3) report the bootstrapped performance and 4) report performance results without the cell filtering.

Table D.9
AUROC performance (all observations), expansion.

Specification	AUROC for model			
	GAM	LASSO	RF	GB
Zero-Model	0.850	0.767	0.923	0.917
Zero-Model + All	0.892 (+5.04%)	0.876 (+14.21%)	0.926 (+0.34%)	0.925 (+0.92%)

Notes: Expansion of the performance results of Table 2 in the main results section. Here, we compare the results of our zero-model, with a specification that adds all remote sensing data sources to the zero-model. Cell filtering takes place. Relative performance differences are calculated with respect to the zero-model.

Table D.10
AUPRC performance (all observations), expansion.

Specification	AUPRC for model			
	GAM	LASSO	RF	GB
Zero-Model	0.657	0.499	0.784	0.779
Zero-Model + All	0.725 (+10.38%)	0.665 (+33.18%)	0.786 (+0.19%)	0.793 (+1.91%)

Notes: Expansion of the performance results of Table 3 in the main results section. Here, we compare the results of our zero-model, with a specification that adds all remote sensing data sources to the zero-model. Cell filtering takes place. Relative performance differences are calculated with respect to the zero-model.

Table E.11
AUROC Performance Yearly (All Observations).

Year	Share Pos.	GAM		LASSO		RF		GB	
		Baseline	Baseline + All	Baseline	Baseline + All	Baseline	Baseline + All	Baseline	Baseline + All
2012	0.182	0.818	0.923 (+12.83%)	0.924	0.952 (+2.95%)	0.931	0.948 (+1.83%)	0.942	0.951 (+0.93%)
2013	0.250	0.945	0.947 (+0.19%)	0.947	0.951 (+0.39%)	0.950	0.953 (+0.35%)	0.948	0.95 (+0.26%)
2014	0.258	0.929	0.933 (+0.42%)	0.920	0.920 (+0%)	0.933	0.938 (+0.47%)	0.929	0.93 (+0.05%)
2015	0.252	0.942	0.943 (+0.11%)	0.941	0.944 (+0.38%)	0.945	0.948 (+0.31%)	0.940	0.944 (+0.38%)
2016	0.236	0.914	0.912 (−0.19%)	0.912	0.914 (+0.19%)	0.914	0.916 (+0.25%)	0.897	0.913 (+1.81%)
2017	0.209	0.899	0.901 (+0.24%)	0.889	0.901 (+1.31%)	0.904	0.905 (+0.11%)	0.898	0.903 (+0.59%)
2018	0.130	0.912	0.911 (−0.12%)	0.910	0.915 (+0.54%)	0.911	0.917 (+0.66%)	0.900	0.912 (+1.38%)
2019	0.108	0.915	0.918 (+0.33%)	0.903	0.891 (−1.31%)	0.916	0.922 (+0.63%)	0.910	0.905 (−0.53%)
2020	0.093	0.895	0.899 (+0.38%)	0.891	0.898 (+0.77%)	0.902	0.902 (+0.01%)	0.900	0.895 (−0.56%)

Notes: Detailed AUROC performance results for all observations differentiated by the forecasting year for the two main specifications (Baseline, Baseline + All). The year 2011 is not included, as it is only used for training. The second column (Share Pos.) reports the share of observations experiencing conflict in the respective year. As in the main text, the relative performance differences (%) are calculated with respect to the baseline specification of the same model type.

Table E.12
AUPRC performance yearly (all observations).

Year	Share Pos.	GAM		LASSO		RF		GB	
		Baseline	Baseline + All	Baseline	Baseline + All	Baseline	Baseline + All	Baseline	Baseline + All
2012	0.182	0.703	0.82 (+16.59%)	0.850	0.872 (+2.64%)	0.859	0.872 (+1.55%)	0.861	0.868 (+0.89%)
2013	0.250	0.885	0.888 (+0.34%)	0.889	0.891 (+0.26%)	0.887	0.900 (+1.53%)	0.889	0.894 (+0.59%)
2014	0.258	0.839	0.848 (+1.01%)	0.826	0.826 (+0%)	0.854	0.860 (+0.71%)	0.841	0.843 (+0.18%)
2015	0.252	0.864	0.861 (−0.24%)	0.862	0.863 (+0.11%)	0.870	0.876 (+0.75%)	0.858	0.867 (+1.1%)
2016	0.236	0.803	0.799 (−0.47%)	0.808	0.811 (+0.45%)	0.807	0.810 (+0.34%)	0.767	0.809 (+5.48%)
2017	0.209	0.742	0.744 (+0.26%)	0.730	0.747 (+2.35%)	0.747	0.751 (+0.56%)	0.728	0.746 (+2.38%)
2018	0.130	0.703	0.696 (−0.93%)	0.693	0.711 (+2.63%)	0.721	0.730 (+1.35%)	0.696	0.713 (+2.53%)
2019	0.108	0.695	0.700 (+0.68%)	0.692	0.673 (−2.72%)	0.707	0.713 (+0.91%)	0.691	0.694 (+0.46%)
2020	0.093	0.525	0.534 (+1.59%)	0.520	0.531 (+2.28%)	0.568	0.570 (+0.39%)	0.557	0.559 (+0.36%)

Notes: Detailed AUPRC performance results for all observations differentiated by the forecasting year for the two main specifications (Baseline, Baseline + All). The year 2011 is not included, as it is only used for training. The second column (Share Pos.) reports the share of observations experiencing conflict in the respective year. As in the main text, the relative performance differences (%) are calculated with respect to the baseline specification of the same model type.

Table F.13
AUPRC performance since last conflict (all observations).

# Months last conflict	GAM		LASSO		RF		GB	
	Baseline	Baseline + All	Baseline	Baseline + All	Baseline	Baseline + All	Baseline	Baseline + All
1	0.909	0.915 (+0.64%)	0.889	0.898 (+1.05%)	0.912	0.918 (+0.62%)	0.903	0.918 (+1.61%)
2	0.513	0.532 (+3.73%)	0.486	0.503 (+3.49%)	0.535	0.576 (+7.55%)	0.502	0.550 (+9.51%)
3	0.378	0.398 (+5.38%)	0.360	0.380 (+5.38%)	0.378	0.397 (+5.18%)	0.377	0.354 (−5.94%)
4	0.276	0.277 (+0.43%)	0.268	0.276 (+3%)	0.297	0.333 (+12.16%)	0.242	0.261 (+7.86%)
5	0.208	0.212 (+2.28%)	0.181	0.212 (+17.47%)	0.210	0.247 (+17.37%)	0.203	0.183 (−10.05%)
≥6	0.094	0.105 (+12.07%)	0.076	0.091 (+19.27%)	0.088	0.103 (+16.16%)	0.084	0.098 (+16.64%)

Notes: Detailed AUPRC performance results for all observations differentiated by the number (#) of months since last conflict for the two main specifications (Baseline, Baseline + All). The results in the first row (1 month since last conflict) are equivalent to the results reported in Table 5 in the main text, as they represent conflict persistence. The five remaining rows offer a more detailed view on the performance for conflict onset. As in the main text, the relative performance differences (%) are calculated with respect to the baseline specification of the same model type.

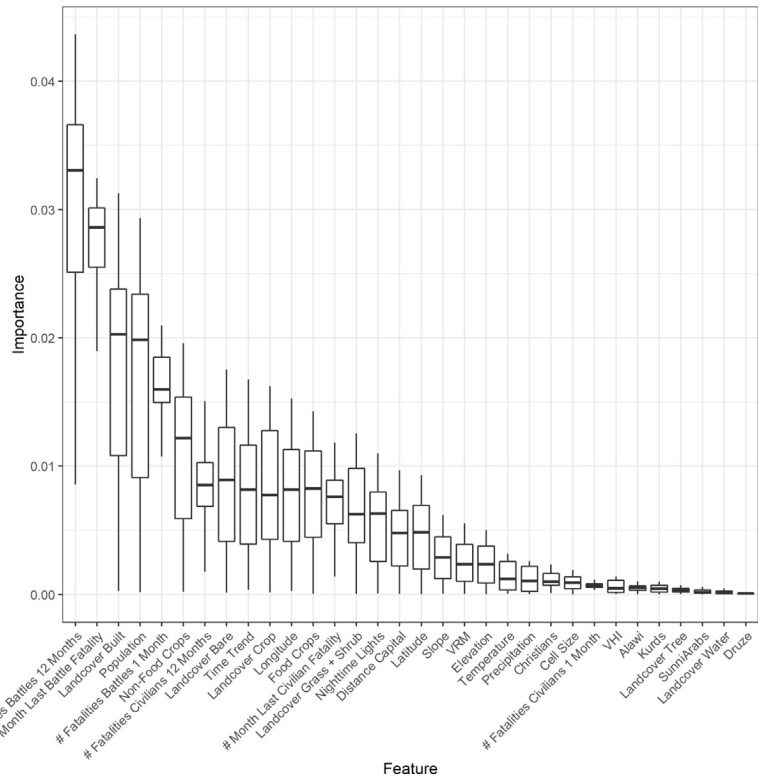


Fig. H.3. Boxplot of the permutation feature importance scores for the full specification (Baseline + All) using RF. We derived the importance scores for each of the 108 models (trained over the entire forecasting horizon) by calculating the average increase in classification error on the out-of-bag data sample for all trees. The features are ordered with respect to their median importance. We excluded the monthly dummies from our feature set, as they require a more elaborate permutation strategy. The boxplots for all other specifications are available on the OSF.

Table G.14
Remote sensing dataset inclusions by LASSO.

Remote sensing dataset	% Included
Landcover classes	77.8%
Population	76.9%
Nighttime Lights	55.6%
Topography	60.2%
Vegetation Health	66.7%
Crops	73.1%
Precipitation	73.1%
Temperature	50.9%

Notes: % of times a remote sensing dataset was included by LASSO in the full specification over the entire forecasting horizon (108 model fits). Each time one variable of a remote sensing dataset is included, we count this as an inclusion of the respective dataset.

References

Abosedra, S., Fakih, A., & Haimoun, N. (2021). *Ethnic divisions and the onset of civil wars in Syria*: *Technical Report*, GLO Discussion Paper.

Amatulli, G., Domisch, S., Tuanmu, M. N., Parmentier, B., Ranipeta, A., Malczyk, J., et al. (2018). A suite of global cross-scale topographic variables for environmental and biodiversity modeling. *Scientific Data*, *5*, 1–15. http://dx.doi.org/10.1038/sdata.2018.40.

Attinà, F., Carammia, M., & Iacus, S. M. (2022). Forecasting change in conflict fatalities with dynamic elastic net. arXiv preprint arXiv:2205.14073.

Avtar, R., Kouser, A., Kumar, A., Singh, D., Misra, P., Gupta, A., et al. (2021). Remote sensing for international peace and security: Its role and implications. *Remote Sensing*, *13*(439).

Bagozzi, B. E., Koren, O., & Mukherjee, B. (2017). Droughts land appropriation, and rebel violence in the developing world. *The Journal of Politics*, *79*, 1057–1072.

Ball, J. E., Anderson, D. T., & Chan Sr, C. S. (2017). Comprehensive survey of deep learning in remote sensing: Theories tools, and challenges for the community. *Journal of Applied Remote Sensing*, *11*, Article 042609.

Bastin, J. F., Finegold, Y., Garcia, C., Mollicone, D., Rezende, M., Routh, D., et al. (2019). The global tree restoration potential. *Science*, *365*, 76–79.

Baugh, K., Elvidge, C. D., Ghosh, T., & Ziskin, D. (2010). Development of a 2009 stable lights product using DMSP-OLS data. *Proceedings of the Asia-Pacific Advanced Network*, *30*(114).

Bazzi, S., Blair, R. A., Blattman, C., Dube, O., Gudgeon, M., & Peck, R. (2022). The promise and pitfalls of conflict prediction: Evidence from Colombia and Indonesia. *The Review of Economics and Statistics*, *104*, 764–779.

Berger, M., Moreno, J., Johannessen, J. A., Levelt, P. F., & Hanssen, R. F. (2012). ESA's sentinel missions in support of Earth system science. *Remote Sensing of Environment*, *120*, 84–90.

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

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2021). Deep neural networks and tabular data: A survey. arXiv preprint arXiv:2110.01889.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Buchhorn, M., Smets, B., Bertels, L., Roo, B. De., Lesiv, M., Tsendbazar, N. E., et al. (2020). Copernicus global land service: land cover 100 m: version 3 globe 2015-2019. http://dx.doi.org/10.5281/Zenodo.3939050.

Buhaug, H., Croicu, M., Fjelde, H., & Uexkull, N. von. (2021). A conditional model of local income shock and civil conflict. *The Journal of Politics*, *83*, 354–366.

Chaudhary, C., Richardson, A. J., Schoeman, D. S., & Costello, M. J. (2021). Global warming is causing a more pronounced dip in marine species richness around the equator. *Proceedings of the National Academy of Sciences*, *118*, Article e2015094118.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining* (pp. 785–794).

Collier, P. (2004). *Development and conflict* (pp. 1–12). Centre for the Study of African Economies.

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

Cranmer, S. J., & Desmarais, B. A. (2017). What can we learn from predictive modeling? *Political Analysis*, *25*, 145–166.

Danielson, J. J., & Gesch, D. B. (2011). *Global multi-resolution terrain elevation data 2010 (GMTED2010)*. US Department of the Interior, US Geological Survey Washington, DC, USA.

Durrleman, S., & Simon, R. (1989). Flexible regression models with cubic splines. *Statistics in Medicine*, *8*, 551–561.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, *11*, 89–121.

Elvidge, C. D., Hsu, F. C., Baugh, K. E., & Ghosh, T. (2014). National trends in satellite-observed lighting. In *Global Urban Monitoring and Assessment Through Earth Observation, vol. 23* (pp. 97–118).

Eng, B., & Martinez, J. (2014). Starvation, submission and survival the Syrian War through the prism of food. Middle East Report 44.

Fabbri, C., Kasper, S., Kautzky, A., Zohar, J., Souery, D., Montgomery, S., et al. (2020). A polygenic predictor of treatment-resistant depression using whole exome sequencing and genome-wide genotyping. *Translational Psychiatry*, *10*, 1–12.

FAO (2022). Agricultural stress index system (ASIS). dataset identifier: 84e27651-0bb4-4a26-8b4a-2b10bbccb7e0 http://www.fao.org/giews/earthobservation/. (Accessed 15 March 2022).

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

Fife, D. A., & D'Onofrio, J. (2022). Common, uncommon, and novel applications of random forest in psychological research. *Behavior Research Methods*.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 1189–1232.

Fritz, C., Mehrl, M., Thurner, P. W., & Kauermann, G. (2022). The role of governmental weapons procurements in forecasting monthly fatalities in intrastate conflicts: a semiparametric hierarchical hurdle model. *International Interactions*, 1–22.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., et al. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes version 2.0. *Scientific Data*, *2*, 1–21. http://dx.doi.org/10.1038/sdata.2015.66.

Girardin, L., Hunziker, P., Cederman, L. E., Bormann, N. C., & Vogt, M. (2015). Growup-geographical research on war, unified platform. ETH Zurich. http://growup.ethz.ch.

Gleditsch, K. S., & Ward, M. D. (2013). Forecasting is difficult especially about the future: using contentious issues to forecast interstate disputes. *Journal of Peace Research*, *50*, 17–31.

Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G., et al. (2010). A global model for forecasting political instability. *American Journal of Political Science*, *54*, 190–208.

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*, 40–55.

Harff, B., & Gurr, T. R. (1998). Systematic early warning of humanitarian emergencies. *Journal of Peace Research*, *35*, 551–579.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S. Routledge* (pp. 249–307).

Hegre, H., Allansson, M., Basedau, M., Colaresi, M., Croicu, M., Fjelde, H., et al. (2019). Views: A political violence early-warning system. *Journal of Peace Research*, *56*, 155–174.

Hegre, H., Metternich, N. W., Nygård, H. M., & Wucherpfennig, J. (2017). Introduction: forecasting in peace research.

Hegre, H., Nygård, H. M., & Landsverk, P. (2021). Can we predict armed conflict? How the first 9 years of published forecasts stand up to reality. *International Studies Quarterly*, *65*, 660–668.

Human Rights Watch (2022). Syria: Events of 2021. https://www.hrw.org/world-report/2022/country-chapters/syria. (Accessed 11 April 2022).

Ismail, S. (2011). The Syrian uprising: Imagining and performing the nation. *Studies in Ethnicity and Nationalism*, *11*, 538–549.

Jerven, M. (2013). *Poor numbers: How we are misled by african development statistics and what to do about it*. Cornell University Press.

Kelley, C. P., Mohtadi, S., Cane, M. A., Seager, R., & Kushnir, Y. (2015). Climate change in the Fertile Crescent and implications of the recent Syrian drought. *Proceedings of the National Academy of Sciences*, *112*, 3241–3246.

King, G., & Zeng, L. (2001). Improving forecasts of state failure. *World Politics*, *53*, 623–658.

Kogan, F. N. (1997). Global drought watch from space. *Bulletin of the American Meteorological Society*, *78*, 621–636.

Kogan, F., Yang, B., Wei, G., Zhiyuan, P., & Xianfeng, J. (2005). Modelling corn production in China using AVHRR-based vegetation health indices. *International Journal of Remote Sensing*, *26*, 2325–2336.

Koren, O., & Bagozzi, B. E. (2017). Living off the land: The connection between cropland food security, and violence against civilians. *Journal of Peace Research*, *54*, 351–364.

Leenders, R., & Heydemann, S. (2012). Popular mobilization in Syria: Opportunity and threat and the social networks of the early risers. *Mediterranean Politics*, *17*, 139–159.

Loveland, T. R., & Dwyer, J. L. (2012). Landsat: Building a strong future. *Remote Sensing of Environment*, *122*, 22–29.

Messer, E., & Cohen, M. J. (2015). Breaking the links between conflict and hunger redux. *World Medical & Health Policy*, *7*, 211–233.

Mueller, H., & Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, *112*, 358–375.

Mueller, H., & Rauh, C. (2022). Using past violence and current news to predict changes in violence. *International Interactions*, *48*, 579–596.

Pedelty, J., Devadiga, S., Masuoka, E., Brown, M., Pinzon, J., Tucker, C., et al. (2007). Generating a long-term land data record from the AVHRR and MODIS instruments. In *2007 IEEE International geoscience and remote sensing symposium* (pp. 1021–1025). IEEE.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., et al. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, *38*, 705–871.

Pettersson, T., Davies, S., Deniz, A., Engström, G., Hawach, N., Högbladh, S., et al. (2021). Organized violence 1989–2020 with a special emphasis on Syria. *Journal of Peace Research*, *58*, 809–825.

Pinstrup-Andersen, P., & Shimokawa, S. (2008). Do poverty and poor health and nutrition increase the risk of armed conflict onset? *Food Policy*, *33*, 513–520.

Raleigh, C., & Hegre, H. (2009). Population size concentration, and civil war. A Geographically Disaggregated Analysis. *Political Geography*, *28*, 224–238.

Raleigh, C., Linke, A., Hegre, H., & Karlsen, J. (2010). Introducing ACLED: An armed conflict location and event dataset: Special data feature. *Journal of Peace Research*, *47*, 651–660.

Richardson, L. F. (1960). *Statistics of deadly quarrels. vol. 10*. Boxwood Press.

Rojas, O., Vrieling, A., & Rembold, F. (2011). Assessing drought probability for agricultural areas in Africa with coarse resolution remote sensing imagery. *Remote Sensing of Environment*, *115*, 343–352.

Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B. W., Aslam, W., et al. (2020). Covid-19 future forecasting using supervised machine learning models. *IEEE Access*, *8*, 101489–101499.

Schrodt, P. A., & Gerner, D. J. (2000). Cluster-based early warning indicators for political change in the contemporary Levant. *American Political Science Review*, *94*, 803–817.

Schroeders, U., Schmidt, C., & Gnambs, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*, *82*, 29–56.

Singer, J. D. (1973). The peace researcher and foreign policy prediction. *Peace Science Society (International)*, *21*, 1–13.

Small, M., Singer, J. D., & Bennett, R. (1982). *Resort to arms: International and civil wars 1816-1980*. SAGE Publications, Incorporated.

Sorokin, P. A. (1962). *Social and Cultural Dynamics: Fluctuation of social relationships, war, and revolution. volume 3*. Bedminster Press.

Sundberg, R., & Melander, E. (2013). Introducing the UCDP georeferenced event dataset Version 21.1. *Journal of Peace Research*, *50*, 523–532.

Tatem, A. J. (2017). Worldpop open data for spatial demography. *Scientific Data*, *4*, 1–4. http://dx.doi.org/10.5258/SOTON/WP00660.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*, 267–288.

Tollefsen, A. F., Strand, H., & Buhaug, H. (2012). PRIO-GRID: A unified spatial data structure. *Journal of Peace Research*, *49*, 363–374.

Vesco, P., Hegre, H., Colaresi, M., Jansen, R. B., Lo, A., Reisch, G., et al. (2022). United they stand: Findings from an escalation prediction competition. *International Interactions*, *48*, 860–896.

Vogt, M., Bormann, N. C., Rüegger, S., Cederman, L. E., Hunziker, P., & Girardin, L. (2015). Integrating data on ethnicity geography, and conflict: The ethnic power relations data set family, version 2021. *Journal of Conflict Resolution*, *59*, 1327–1342. http://dx.doi.org/10.1177/0022002715591215.

Von Uexkull, N., Croicu, M., Fjelde, H., & Buhaug, H. (2016). Civil conflict sensitivity to growing-season drought. *Proceedings of the National Academy of Sciences*, *113*, 12391–12396.

Wan, Z., Hook, S., & Hulley, G. (2021). MOD11A1 MODIS/terra land surface temperature/emissivity daily L3 global 1km SIN Grid, Version 6.1. http://dx.doi.org/10.5067/MODIS/MOD11A1.061, NASA EOSDIS Land Processes DAAC, 2015.

Weidmann, N. B., & Schutte, S. (2017). Using night light emissions for the prediction of local wealth. *Journal of Peace Research*, *54*, 125–140.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *65*, 95–114.

Wood, R. M., & Sullivan, C. (2015). Doing harm by doing good? The negative externalities of humanitarian aid provision during civil conflict. *The Journal of Politics*, *77*, 736–748.

Wright, Q. (1965). *A study of war* (2nd edition). Chicago, IL: University of Chicago Press.

Xie, J., & Zhu, Y. (2020). Association between ambient temperature and COVID-19 infection in 122 cities from China. *Science of the Total Environment*, *724*, Article 138201.

Yu, Q., You, L., Wood-Sichra, U., Ru, Y., Joglekar, A. K., Fritz, S., et al. (2020). A cultivated planet in 2010–Part 2: The global gridded agricultural-production maps. *Earth System Science Data*, *12*, 3545–3572. http://dx.doi.org/10.5194/essd-12-3545-2020.

Zeitzoff, T. (2017). How social media is changing conflict. *Journal of Conflict Resolution*, *61*, 1970–1991.

S1. Syrian Desert Cells

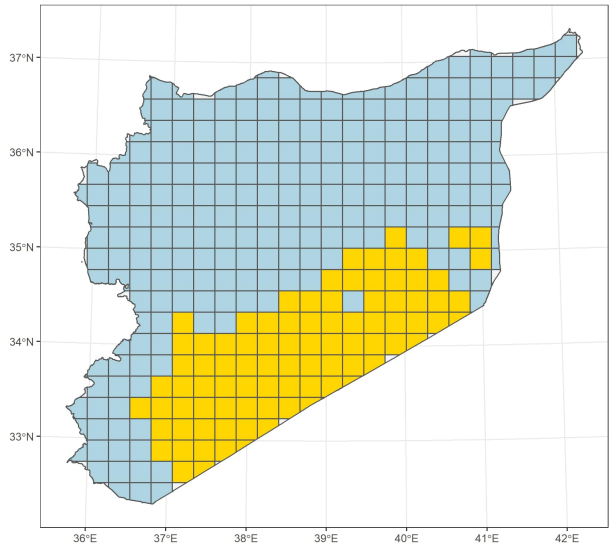
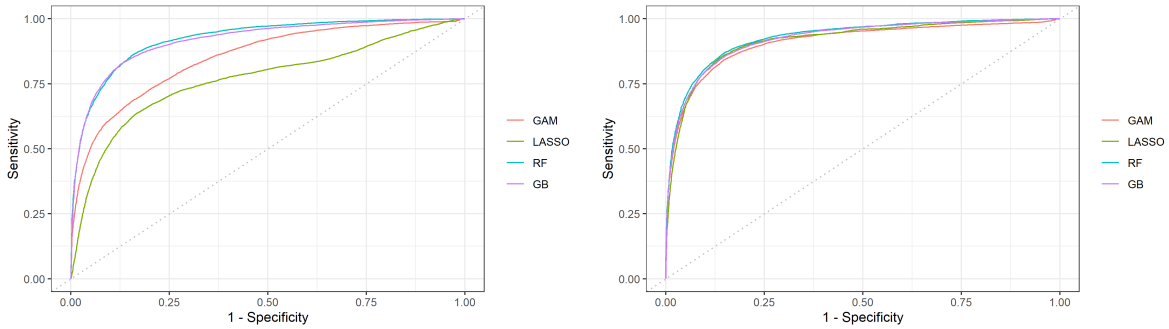
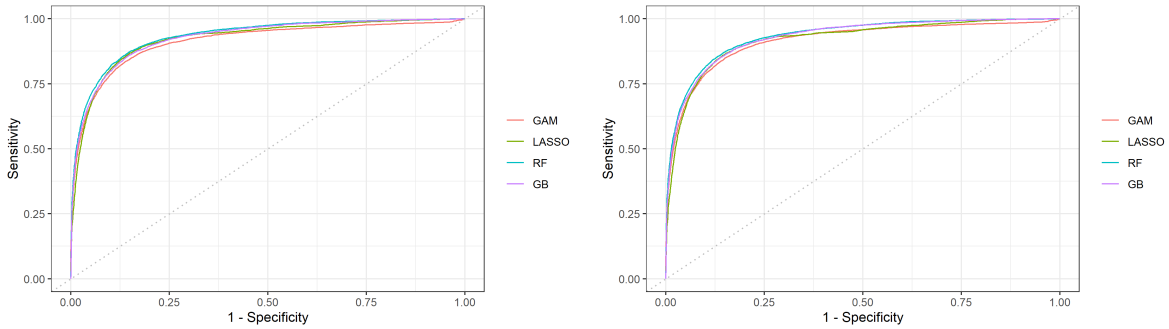


Fig. S1: Illustration of the grid cells mostly covered by the Syrian Desert. The Figure shows all the Syrian grid cells derived according to the rasterization process described in section 2.1. The 34 removed cells with an area of less than 150 km^2 at the Syrian border are shown in grey and not considered here. The remaining cells are either categorized into "desert" (yellow) or "normal" (blue) cells respectively. We consider a cell as covered by desert, if the landcover classes indicate a "bare" area of more than 80% of the total area. This holds true for 95 of the remaining 322 cells. The Figure shows that most of those cells are clustered together at the southern Syrian boarder (which aligns with other maps depicting the Syrian Desert). Hence, this area can be easily captured by models through spatial coordinates.

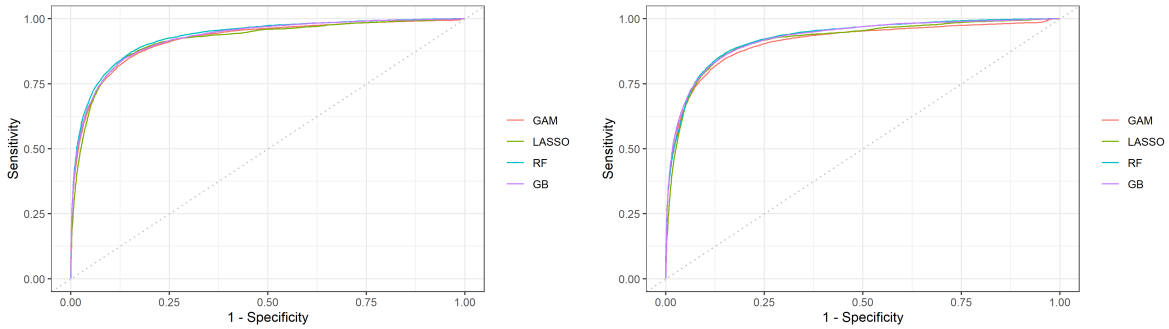
S2. Curves



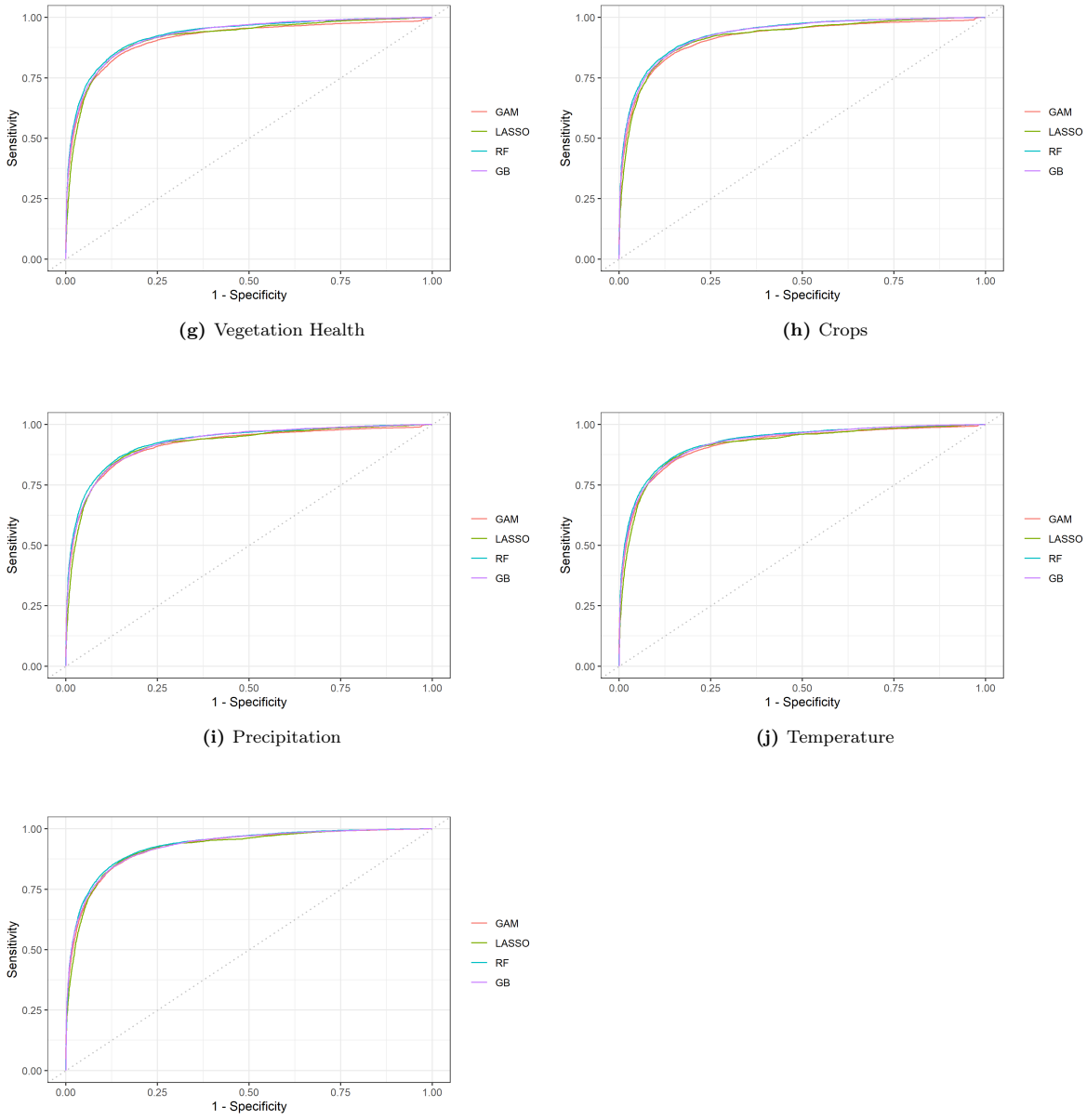
(a) Zero-Model **(b)** Baseline



(c) Landcover classes **(d)** Population



(e) Nighttime Lights **(f)** Topography



(g) Vegetation Health

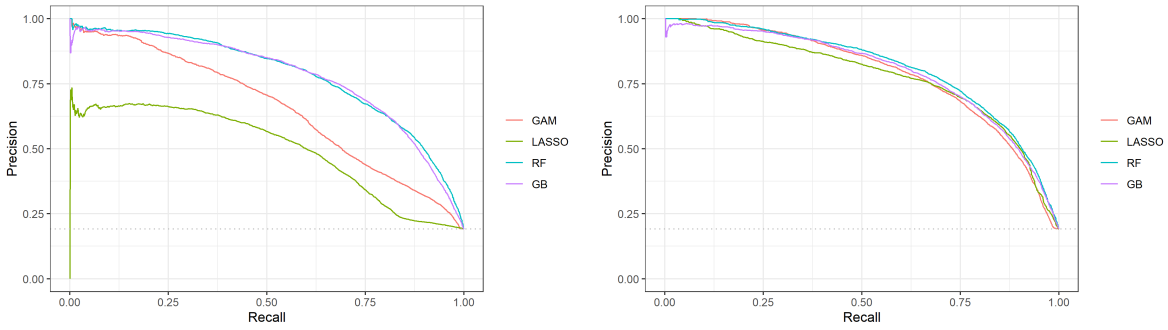
(h) Crops

(i) Precipitation

(j) Temperature

(k) All

Fig. S2: Receiver operating characteristic curves corresponding to the main results in Table 2. Each subplot corresponds to one specification, with the curves for all four models. The larger the area under a curve, the better a model performs according to the AUROC.



(a) Zero-Model

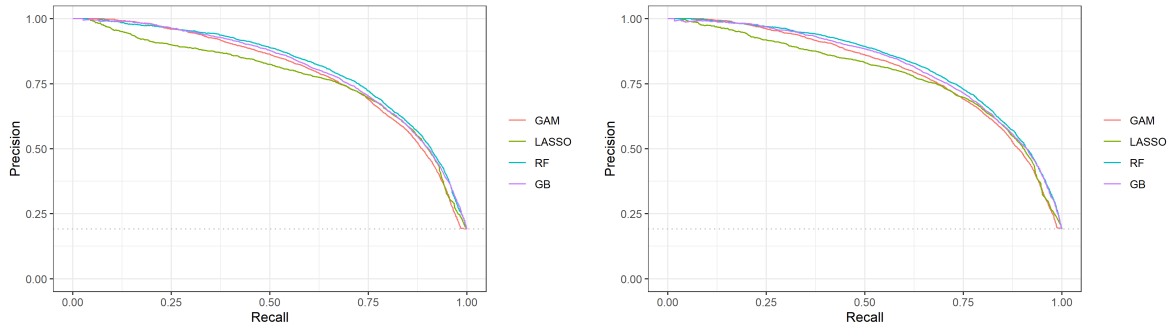
(b) Baseline

(c) Landcover classes

(d) Population

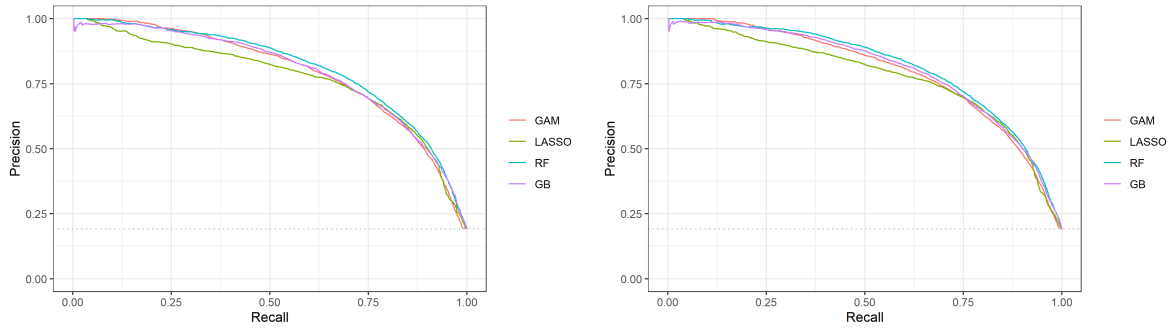
(e) Nighttime Lights

(f) Topography



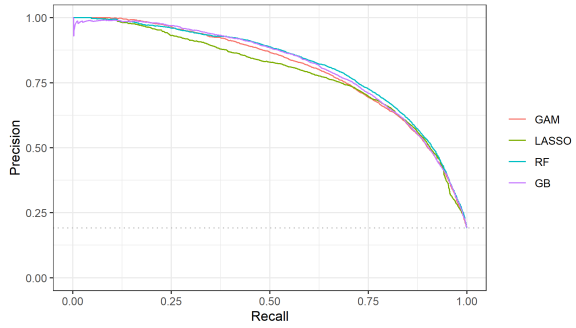
(g) Vegetation Health

(h) Crops



(i) Precipitation

(j) Temperature



(k) All

Fig. S3: Precision recall curves corresponding to the main results in Table 3. Each subplot corresponds to one specification, with the curves for all four models. The larger the area under a curve, the better a model performs according to the AUPRC.

S3. Bootstrapped Performance Results

The following three tables report the two-sided 95% confidence intervals of our performance results in the main text for three selected specifications (zero-model, baseline, full specification) using bootstrapping. We restricted the bootstrap to our test dataset, as a full bootstrap (train & test) would drastically increase the computational complexity. Each bootstrap iteration would then require re-running our entire forecasting pipeline including the hyperparameter tuning. Instead, we evaluate each model of the forecasting pipeline on $B = 25$ bootstrapped samples, drawing observations with replacement from the respective test dataset for each month (i.e. if we are in $t = 100$ and want to evaluate the forecasting performance for $t + 1$, we are only drawing test observations with $t = 101$), and derive the performance and relative difference for each of the samples. Table S1 provides confidence intervals for the results in Table 2, Table S2 for Table 3 and Table S3 for Table 4.

Table S1: Bootstrapped AUROC Performance (All Observations), 95% Confidence Interval

Specification	AUROC for Model			
	GAM	LASSO	RF	GB
Baseline	[0.9095; 0.9111]	[0.9162; 0.9175]	[0.9280; 0.9293]	[0.9225; 0.9239]
Zero-Model	[0.8483; 0.8507]	[0.7650; 0.7685]	[0.9223; 0.9235]	[0.9160; 0.9174]
Baseline + All	[0.9255; 0.9270]	[0.9227; 0.9242]	[0.9315; 0.9326]	[0.9280; 0.9292]
% Difference Zero-Model	[-6.781; -6.574]	[-16.556; -16.185]	[-0.662; -0.578]	[-0.758; -0.650]
% Difference Baseline + All	[+1.696; +1.802]	[+0.690; +0.766]	[+0.342; +0.389]	[+0.536; +0.634]

Notes: Bootstrapped performance two-sided confidence intervals (95%) pertaining to Table 2 in the main results section. The %-difference is calculated with respect to the Baseline specification.

Table S2: Bootstrapped AUPRC Performance (All Observations), 95% Confidence Interval

Specification	AUPRC for Model			
	GAM	LASSO	RF	GB
Baseline	[0.7886; 0.7917]	[0.7757; 0.7788]	[0.8084; 0.8109]	[0.7948; 0.7978]
Zero-Model	[0.6551; 0.6591]	[0.4968; 0.5018]	[0.7827; 0.7863]	[0.7772; 0.7801]
Baseline + All	[0.8038; 0.8066]	[0.7875; 0.7904]	[0.8161; 0.8186]	[0.8101; 0.8132]
% Difference Zero-Model	[-17.029; -16.647]	[-36.058; -35.457]	[-3.252; -2.962]	[-2.368; -2.084]
% Difference Baseline + All	[+1.837; +1.977]	[+1.414; +1.598]	[+0.908; +0.994]	[+1.779; +2.072]

Notes: Bootstrapped performance two-sided confidence intervals (95%) pertaining to Table 3 in the main results section. The %-difference is calculated with respect to the Baseline specification.

Table S3: Bootstrapped AUPRC Performance for Conflict Onset Observations, 95% Confidence Interval

Specification	AUPRC for Model			
	GAM	LASSO	RF	GB
Baseline	[0.3246; 0.3340]	[0.3068; 0.3168]	[0.3570; 0.3675]	[0.3242; 0.3349]
Zero-Model	[0.2379; 0.2442]	[0.1583; 0.1629]	[0.3521; 0.3617]	[0.3176; 0.3289]
Baseline + All	[0.3417; 0.3525]	[0.3215; 0.3326]	[0.3808; 0.3921]	[0.3503; 0.3608]
% Difference Zero-Model	[-27.490; -26.092]	[-49.259; -47.715]	[-2.203; -0.782]	[-2.850; -0.958]
% Difference Baseline + All	[+4.913; +5.923]	[+4.392; +5.401]	[+6.178; +7.142]	[+6.995; +8.823]

Notes: Bootstrapped performance two-sided confidence intervals (95%) pertaining to Table 4 in the main results section. The %-difference is calculated with respect to the Baseline specification.

S4. Results without Cell Filtering *(will be moved to supplementary material)*

The following two tables report our results for the entire study period (all observations), without the filtering of the 34 small cells as described in section 2.1. Hence, the results of Table S4 correspond to Table 2 and Table S5 correspond to Table 3 in the main text. As we can see, the overall patterns are very similar to the ones observed before and we can similarly conclude that remote sensing data increases forecasting performance. Notably, these increases are smaller than in our chosen sample for both LASSO and GB, larger for RF and similar for the GAM. Possible performance drops are expected due to two reasons. First, for those 34 cells the main variable of importance is the cell size (included in all specifications), as the cells are so small, that conflict almost never takes place. Hence, including additional predictors such as variables constructed from remote sensing data will provide little to no additional information, which in turn will decrease average performance gains for our remote sensing specifications. Second, these small cells pose an additional problem in model training, as they substantially increase the risk of overfitting. Some of these cells are not larger than 10 km^2 (compared to 625 km^2 for full cells), thus they sometimes exhibit unusual explanatory variable values. Hence, in those rare instances where conflict indeed takes place in one of those cells, the models might incorrectly attribute this effect to one of the remote sensing explanatory variables. Arguably, in such a setting, models that can capture non-linear effects and at the same time have a low risk of overfitting should thus perform best. Indeed, this is confirmed by the fact that the performance (gain) of RF is not negatively affected by the inclusion of those cells, as RFs can model non-linear effects and are much less prone to overfitting than for example GB.

Table S4: AUROC Performance (All Observations), Non-filtered Cells

Specification	AUROC for Model			
	GAM	LASSO	RF	GB
Zero-Model	0.854 (-6.52%)	0.771 (-16.58%)	0.928 (-0.03%)	0.926 (-0.1%)
Baseline	0.914 (+0%)	0.924 (+0%)	0.929 (+0%)	0.927 (+0%)
Baseline + Landcover Classes	0.915 (+0.09%)	0.924 (+0%)	0.936 (+0.83%)	0.93 (+0.38%)
Baseline + Population	0.915 (+0.19%)	0.927 (+0.32%)	0.936 (+0.77%)	0.933 (+0.67%)
Baseline + Nighttime Lights	0.914 (+0.07%)	0.924 (+0%)	0.935 (+0.64%)	0.93 (+0.41%)
Baseline + Topography	0.911 (-0.25%)	0.923 (-0.09%)	0.933 (+0.48%)	0.929 (+0.23%)
Baseline + Vegetation Health	0.915 (+0.12%)	0.924 (+0.01%)	0.929 (+0.06%)	0.93 (+0.38%)
Baseline + Crops	0.917 (+0.36%)	0.926 (+0.23%)	0.936 (+0.79%)	0.931 (+0.43%)
Baseline + Precipitation	0.917 (+0.35%)	0.923 (-0.08%)	0.932 (+0.33%)	0.93 (+0.35%)
Baseline + Temperature	0.916 (+0.3%)	0.924 (-0.01%)	0.933 (+0.48%)	0.927 (+0.07%)
Baseline + All	0.922 (+0.94%)	0.927 (+0.36%)	0.936 (+0.85%)	0.926 (-0.05%)

Notes: Average area under the receicer operator characteristics curve (AUROC) performance for one-step ahead forecasts over the entire forecasting horizon of the different model specifications and types. The results reported are for all cells, without the filtering described in section 2.1. For further details we refer to the corresponding Table without the cell filtering (Table 2) in the main results section.

Table S5: AUPRC Performance (All Observations), Non-filtered Cells

Specification	AUPRC for Model			
	GAM	LASSO	RF	GB
Zero-Model	0.649 (-17.59%)	0.494 (-36.87%)	0.784 (-1.76%)	0.783 (-0.74%)
Baseline	0.787	0.782	0.798	0.789 (+0%)
Baseline + Landcover Classes	0.789 (+0.24%)	0.776 (-0.74%)	0.817 (+2.48%)	0.799 (+1.34%)
Baseline + Population	0.788 (+0.09%)	0.784 (+0.18%)	0.815 (+2.24%)	0.802 (+1.61%)
Baseline + Nighttime Lights	0.788 (+0.08%)	0.782 (+0.03%)	0.811 (+1.72%)	0.798 (+1.19%)
Baseline + Topography	0.78 (-0.84%)	0.776 (-0.75%)	0.809 (+1.45%)	0.798 (+1.15%)
Baseline + Vegetation Health	0.788 (+0.1%)	0.783 (+0.08%)	0.805 (+0.91%)	0.801 (+1.58%)
Baseline + Crops	0.79 (+0.37%)	0.782 (+0.01%)	0.816 (+2.33%)	0.803 (+1.85%)
Baseline + Precipitation	0.789 (+0.29%)	0.777 (-0.66%)	0.805 (+0.92%)	0.799 (+1.26%)
Baseline + Temperature	0.787 (-0.01%)	0.782 (-0.04%)	0.811 (+1.64%)	0.794 (+0.63%)
Baseline + All	0.796 (+1.12%)	0.783 (+0.1%)	0.815 (+2.17%)	0.795 (+0.78%)

Notes: Average area under the precision-recall curve (AUPRC) performance for one-step ahead forecasts over the entire forecasting horizon of the different model specifications and types. The results reported are for all cells, without the filtering described in section 2.1. For further details we refer to the corresponding Table without the cell filtering (Table 3) in the main results section.

7. Unsupervised Detection of Building Destruction during War from Publicly Available Radar Satellite Imagery

Contributing article

Racek, D., Zhang, Q., Thurner, P., Zhu, X. X., and Kauermann, G. (2025). Unsupervised Detection of Building Destruction during War from Publicly Available Radar Satellite Imagery. *Center for Open Science, (No. 86t3g_v2).* https://doi.org/10.31219/osf.io/86t3g_v2.

Data and code

Available at https://github.com/Daniel-Rac/Detecting_destruction_from_public_satellite_images.

Supplementary material

Supplementary material is available online.

Author contributions

The original idea to use publicly available satellite images to detect destruction in conflict zones originated from Daniel Racek, Paul Thurner, Xiao Xiang Zhu and Göran Kauermann. Xiao Xiang Zhu and Qi Zhang suggested to use synthetic-aperture radar (SAR) images from Sentinel-1 for the detection. Qi Zhang provided the detection algorithm and obtained the interferometric coherence scores. Daniel Racek and Göran Kauermann designed the statistical methodology to differentiate destruction from random noise. Daniel Racek conducted the remainder of the study, analysed the results and created the visualizations. Daniel Racek and Qi Zhang wrote the first draft of the article. All authors were involved in improving and editing the article.

UNSUPERVISED DETECTION OF BUILDING DESTRUCTION DURING WAR FROM PUBLICLY AVAILABLE RADAR SATELLITE IMAGERY

Daniel Racek[*]
Institute of Statistics, LMU Munich
München, Germany

Qi Zhang
School of Engineering and Design, TU Munich
München, Germany

Paul W. Thurner
Institute of Political Science, LMU Munich
München, Germany

Xiao Xiang Zhu
School of Engineering and Design, TU Munich
Munich Center for Machine Learning
München, Germany

Göran Kauermann
Institute of Statistics, LMU Munich
München, Germany

ABSTRACT

Automated detection of building destruction in conflict zones is crucial for human rights monitoring, humanitarian response, and academic research. However, existing approaches 1) rely on proprietary satellite imagery, both expensive and not accessible at wartime, 2) require manually labeled training data, usually not available in war-affected regions, or 3) use optical imagery, regularly obstructed by cloud cover. This study addresses these challenges by introducing an unsupervised method to detect destruction at the building level using freely and globally available Sentinel-1 synthetic aperture radar (SAR) images from the European Space Agency (ESA). By statistically assessing interferometric coherence changes over time, unlike existing approaches, our method enables the detection of destruction from a single satellite image, allowing for near real-time destruction assessments every 12 days. We provide a continuous, statistically grounded probability measure for the likelihood of destruction at both the building and pixel level, thereby quantifying the level of uncertainty of the detection. Using ground truth data and reported sequences of events, we validate our approach both quantitatively and qualitatively, across three case studies in Beirut, Mariupol, and Gaza, demonstrating its ability to accurately identify the spatial patterns and timing of destruction events. Using open-access data, our method offers a scalable, global, and cost-effective solution for monitoring building destruction in conflict zones.

Keywords Destruction · Conflict · Remote Sensing · Satellites

Significance Statement

Understanding the extent and timing of building destruction during conflicts is crucial for improving crisis response and advancing our understanding of conflicts. Yet, current approaches are often inaccessible due to their reliance on proprietary satellite data and ground truth labels unavailable in conflict zones. Combining remote sensing techniques with robust statistics, our study introduces an unsupervised algorithm that uses freely available Sentinel-1 radar imagery to detect destruction with uncertainty estimates. Tested across three real-world case studies, our method is able to reconstruct the chronology of destruction events. By leveraging open data, we democratize access to critical tools for conflict monitoring and assessment.

[*]Corresponding author: daniel.racek@lmu.de

1 Introduction

The ability to detect and assess damage and destruction of buildings in conflict zones is important for monitoring human rights, facilitating humanitarian aid, guiding reconstruction efforts, and more generally academic research on armed conflict. Traditionally, data on destruction has come from ground reports or manually inspected satellite images, which is often resource-intensive, prone to bias, and limited in scope. This has led to a recent shift towards automated remote sensing solutions, often using machine or deep learning learning techniques in combination with high-resolution satellite imagery.

Despite substantial progress, existing methods face significant limitations. Many rely on proprietary high-resolution imagery (e.g., 30 cm), which is expensive, difficult to scale, and usually not made available during wartime [Hou et al., 2024, Kahl and Chen, 2024]. Approaches using publicly available medium-resolution images (e.g., 10–20m), on the other hand, face challenges in reliably identifying the destruction of individual buildings, hence recently introduced techniques rely on multiple images taken over a longer period of time for detection [Hou et al., 2024, Dietrich et al., 2025]. As a result, they cannot precisely determine when destruction occurs or provide near real-time assessments of destruction patterns. Another limitation of most methods is that they are supervised [Cheng et al., 2024, Hou et al., 2024, Dietrich et al., 2025], requiring labeled ground truth data for training, which is, at least initially, unavailable in war-affected regions. Finally, most existing approaches rely on optical satellite images, which are frequently obstructed by cloud cover, further limiting the timely detection of destruction.

To address these limitations, in this work, we propose an unsupervised solution to detect building destruction caused by armed conflict and war, using publicly available imagery from the European Space Agency (ESA), specifically synthetic-aperture radar (SAR) images from Sentinel-1, in 12-day time periods, available globally since 2016. Unlike optical satellite imagery, which relies on clear weather and daylight for optimal image quality, SAR technology can operate day and night under all weather conditions [Moreira et al., 2013, Cheng et al., 2024], making it particularly suitable for conflict zones. Although the spatial resolution of Sentinel-1 imagery is comparably low (20m after processing; see also Fig.1), our approach is able to identify destruction every 12 days at the building level.

We employ Interferometric SAR (InSAR) to measure the stability of an area between two SAR images acquired at different points in time [Bamler and Hartl, 1998, Yagüe-Martínez et al., 2016]. We repeatedly calculate these interferometric coherence scores of temporally adjacent images over extended periods of time. Based on a statistical assessment, using non-parametric median regressions and outlier-robust estimation techniques, this allows us to differentiate destruction from random background noise for each time period, one of the central challenges for change detection techniques [Shafique et al., 2022]. Generally, in the remote sensing literature, change detection refers to a set of methods that aim to identify changes in the earth's surface, with broad applications in fields such as urban development [Plank, 2014], agriculture [Zhang et al., 2020], and land cover monitoring [Khan et al., 2017], using

Fig. 1: Comparison of satellite images of the Beirut harbor, July 2020. (A) is a proprietary high-resolution (30cm) optical image from Maxar WorldView-3. Due to the harbor explosion, the image was made freely available by Maxar [Maxar Technologies, 2024]. (B) is a publicly available optical image with medium-resolution (10m) from Sentinel-2A, using bands 4, 3 and 2. (C) is a publicly available multi-looked SAR image with medium-resolution (20m after processing; see Methods for details) from Sentinel-1. It is one of the images used in our detection. For visualization of the two-dimensional image of complex samples with a real and imaginary part, we use a γ_0 greyscale visualization of the VV polarization.

various types of satellite images with varying levels of resolution. For a recent overview of the literature, we refer the reader to Osmanoğlu et al. [2016], Shi et al. [2020], Cheng et al. [2024]. For an introduction to SAR imagery and its applications, we refer to Bamler and Hartl [1998], Moreira et al. [2013]. In Fig.1, we provide a comparison of different satellite images, including a Sentinel-1 image.

Our approach is fully unsupervised and hence does not require any labeled training data, making it applicable in scenarios where ground truth data are sparse or even entirely unavailable. This is particularly relevant in the early stages of wars and in conflicts that receive less media attention, where little information on damage and destruction is available [Dietrich et al., 2025]. Contrary to destruction caused by natural disasters, conflict-related scenarios typically exhibit extreme class imbalance, as destruction is limited to comparatively few buildings [Mueller et al., 2021, Hou et al., 2024]. Notably, our approach remains robust to this imbalance. Moreover, it provides a continuous, statistically grounded probability measure for the likelihood of destruction at both the building and pixel level. This stands in contrast to most existing approaches, which generally lack a measure of uncertainty [Sticher et al., 2023] and focus exclusively on classifying either pixels or buildings [Cheng et al., 2024].

Focusing on three case studies - Beirut, Mariupol, and Gaza - each with different destruction dynamics, we demonstrate both quantitatively as well as qualitatively that our method can reliably detect the destruction of buildings and also determine the timing of the destruction, using ground truth data and reported sequences of events for validation. Summarizing, we propose a novel unsupervised approach for detecting building destruction using publicly available Sentinel-1 SAR images. Our results highlight the potential of using freely available satellite imagery to detect destroyed buildings during armed conflict and war at scale, as our approach can be transferred to any other place or region in the world.

2 Results

2.1 Beirut

The Beirut harbor explosion on August 4, 2020, constitutes our first case study. Although not caused by armed conflict, the explosion's effects resemble those observed in conflict-related destruction, with extensive damage concentrated within a densely populated area. This event is particularly practical for analyzing our method, as destruction is limited to a single day, providing a clear temporal boundary for our detection. Furthermore, ground truth data for buildings located in the harbor that were fully destroyed by the explosion are available and field-validated.

Fig.2A visualizes p-values of destruction of all 10m$\times$10m building pixels over 12-day time periods in the area around the explosion, using our detection algorithm. The first time period marks the 12 days before the explosion, the second covers the time of the explosion, and the third is directly after the explosion. Lower p-values indicate a higher likelihood and more evidence that a building was destroyed. We provide the same maps for all time periods in Supp.1. The location of the explosion is denoted by the red dot. Buildings at the harbor annotated as fully destroyed by the explosion are marked by a red border. As evident from the figure, most of the annotated buildings show high evidence of destruction according to our method in the time period of the explosion. Additionally, almost in a perfect radius around the explosion, we identify additional buildings as destroyed. As we move further away from the explosion site, p-values increase, representing likely lower levels of damage to these buildings. These findings are in line with previous research [Pilger et al., 2021, Al-Hajj et al., 2021], which has identified destruction and substantial damage to buildings farther from the explosion site. In both of the other time periods, as expected, there is limited evidence of destruction (high p-values).

Fig.2D provides kernel density estimates of the distribution of p-values of all building pixels across different time periods. The explosion is clearly evident from the spike in low p-values during that period (density in green). Furthermore, we can observe that our approach almost exclusively assigns low p-values for the explosion, and not for any other time period. Note that there is an asymmetric spike in p-values in the periods after the explosion due to the use of first differences in our detection.

To evaluate our approach in a strict classification setting, we assign the 1,754 pixels of annotated buildings to our positive destruction class, as for these we have definitive evidence that the corresponding buildings were fully destroyed. However, there are different options on how to define the set of pixels of the negative class, i.e. those that were neither damaged nor destroyed, as ground truth information for the remaining buildings in the explosion time period is unfortunately not available. This is a general problem when designing and evaluating change detection algorithms [Shi et al., 2020, Shafique et al., 2022] and exacerbated in conflict scenarios [Mueller et al., 2021, Dietrich et al., 2025], thus highlights one of the advantages of our unsupervised approach, which does not need to be trained.

The simplest option is to use all building pixels from all other time periods as our set of data points in the negative class. However, this results in a total of 994,651 data points and thus in an extremely unbalanced sample (0.16% pos. class), for

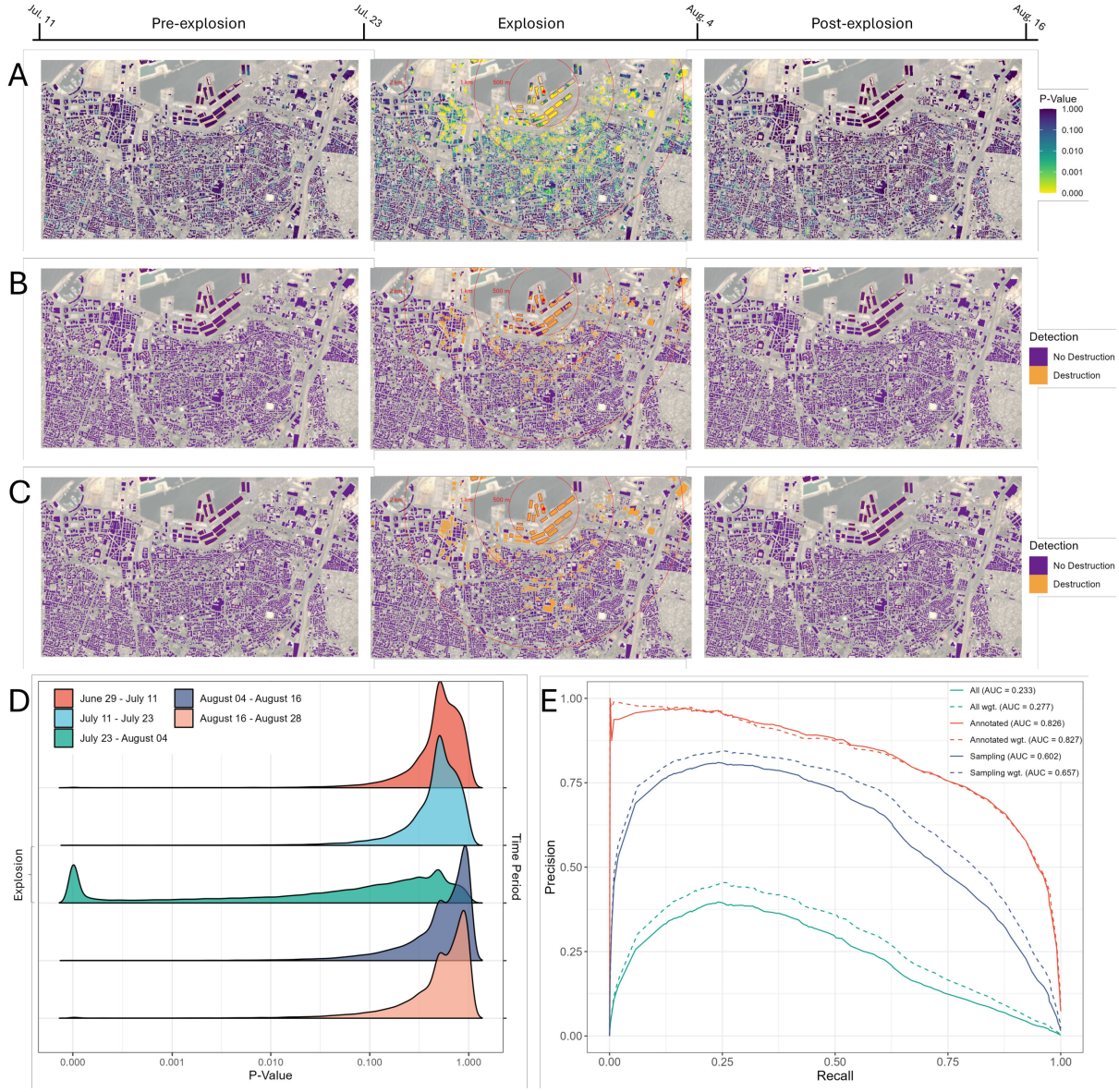


Fig. 2: (A) P-values of destruction of all 10m×10m building pixels in Beirut over 12-day periods from July 11 to July 23 (left), July 23 to August 4 (middle), August 4 to August 16 (right; all 2020). Lower p-values indicate a higher likelihood that part of a building was destroyed. The harbor explosion on August 4 is denoted by the red dot in the middle image, with radii of the blast wave with varying distances (also in red). Buildings located directly next to the sea are missing some pixels due to the processing of the images. The background of each image is an optical Sentinel-2A image (freely available) from July 24, 2020 based on bands 4, 3 and 2. (B) The same building pixels over the same time periods as in A, categorized into destruction and no destruction for an F1-score-optimizing probability threshold. (C) Classification of entire buildings into destruction and no destruction for an F1-score-optimizing probability threshold after combining pixel-wise p-values. (D) Kernel density estimates of the distribution of all building pixel p-values over 12-day time periods before, during and after the explosion (all 2020). (E) Precision-recall curves using the building pixels of the annotated buildings only (red), using random sampling for the negative class (blue), using all building pixels over all time periods (green). The dashed lines denote the curves when pixels are weighted by their building coverage.

which performance scores are skewed by individual outliers that naturally occur, e.g., due to building construction works. This imbalance also motivates our subsequent use of precision-recall (PR) curves for evaluation [Davis and Goadrich, 2006]. Independent of the definition of our evaluation dataset, the corresponding receiver operating characteristic (ROC) curves exhibit almost perfect areas under the curves (AUCs) of > 0.985 (see Supp.5) and thus are not well suited for evaluation. The issue of imbalance in the context of detecting building destruction is extensively discussed in Mueller et al. [2021], and thus we refer the reader for more information to the work of these authors. Hence, instead, in Fig.2E we visualize the precision-recall (PR) curves and the corresponding areas under the curve (AUPRCs) when defining varying probability thresholds as cut-offs for categorizing pixels into destruction vs. no destruction, for various definitions of the evaluation dataset.

Another possibility for evaluation is to only use the pixels of the annotated buildings across all other time periods (32,571 data points; 5.39% pos. class). However, this arguably leads to overoptimistic performance scores (red PR curves). We find that an F1-optimizing probability threshold in this setting to classify many building pixels in other time periods incorrectly as destroyed (see Supp.3). Hence, instead, we opted for randomly sampling building pixels across all other time periods, while retaining the imbalance at a reasonable level (155,100 data points; 1% pos. class). In order to reduce noise, we repeat this process 100 times and average the results. Finally, because many pixels only partially cover buildings, as the resolution of most buildings is more fine-grained than our interferometric coherence scores, we additionally weight each pixel by its relative share of building coverage. The corresponding average PR curve is visualized in Fig.2E in dashed-blue and has a AUPRC of 0.657 (SD = 0.010). The F1-optimizing probability threshold, results in an F1 score of 0.663 (SD = 0.004), with an associated precision of 0.694 (SD = 0.01) and a recall of 0.634 (SD < 0.001).

In Fig.2B we visualize the corresponding classification of all building pixels over the same three time periods as in Fig.1A. We provide the maps for all time periods in Supp.2. As evident, using this threshold, we correctly classify most pixels of the annotated buildings as destroyed. Note, not all pixels will always provide the same level of evidence for destruction (e.g., buildings might only be partially damaged or destroyed). Hence, it is reasonable to classify entire buildings as destroyed, even when only some pixels indicate destruction.

To classify buildings, we combine the pixel-wise p-values of each building by constructing the harmonic mean p-value (HMP) [Wilson, 2019], often used for meta-analyses [Errington et al., 2021]. Equivalently to our pixel-wise classification, we randomly sample buildings from other time periods for evaluation (2,200 data points; 1% pos. class), and repeat this process 100 times. The corresponding PR curve has an AUPRC of 0.905 (SD = 0.058). The F1-optimizing probability threshold results in an F1 score of 0.905 (SD = 0.032), with an associated precision of 0.861 (SD = 0.059) and a recall of 0.955 (SD < 0.001). In Fig.2C, we again visualize this classification. Only a single building (top right in the middle image) is incorrectly classified as not destroyed, while the amount of false positives in the remaining time periods is highly limited. We provide maps for the remaining time periods in Supp.4, and both ROC and PR curves in Supp.5.

In the explosion time period, we classify a total of 361 buildings as destroyed, which are 5.222% of all buildings in our analysis region. In all other time periods, on average, we incorrectly classify only 11.450 (SD = 13.407) as destroyed, a share of 0.166% (SD = 0.194) over all buildings. Notably, the latter is very likely to be at least partially driven by reconstruction efforts, which are similarly identified by our detection algorithm, as these constitute changes in a building's structure. Before the explosion, we classify on average 9.250 (SD = 13.551) as destroyed, whereas this increases to 12.916 (SD = 13.701) after the explosion. We provide further evidence for this theory in Supp.7. We report results for the other two evaluation strategies in Supp.6.

In Supp.8, we present a sensitivity analysis to evaluate the impact of the number of time periods, that is, satellite images, used for the detection. The results show that destruction can be identified immediately within the time period in which it occurs, with as few as eight images observed prior.

2.2 Mariupol

For our second case study, we investigate building destruction during the course of the Russian invasion of Ukraine. Our area of interest is the center of Mariupol, Zhovtnevyi district, for which UNOSAT has compiled ground truth data for a subset of buildings, manually labeled through high-resolution satellite images from Maxar and not field validated. Unlike in Beirut, destruction in Mariupol unfolded over several weeks, allowing us to test our method's ability to detect destruction over an extended period of time. Note, due to the limited and partially validated nature of the ground truth data, performance evaluation should be interpreted with caution.

Fig.3A visualizes p-values of destruction of all 10m$\times$10m building pixels over 12-day time periods from the start of the invasion to the fall of Mariupol. During the first 4 days of the invasion, the center of Mariupol remained mostly unscathed. Over the course of the following weeks, we can observe how the patterns of destruction move

from north-west towards the south-east of the center district, as the Russian army destroys most of the city through bombardments [Ellyatt, 2022]. As evident, our approach is capable of tracking these dynamics over the course of the invasion. We provide the same maps for all time periods in Supp.9.

For evaluation, we rely on a limited number of ground truth labels from May 12, 2022. As these only contain buildings that are destroyed, equivalently to the previous application, we sample data points for our negative class by drawing on pixels and buildings for time periods before the invasion. We derive our summarized destruction classification, by categorizing a pixel or building as destroyed when it is marked by our algorithm as destroyed in any of the 12-day time periods from February 28 to May 11.

We report both pixel- and building-level performance scores in Table 1. For each, we present the F1 score, recall, and precision using both the optimal probability threshold derived from the Beirut use case and the F1-maximizing threshold for Mariupol. AUROC and AUPRC are broadly comparable to those observed in the Beirut case study for both pixels and buildings, with the building AUPRC being the notable exception. Unlike in Beirut, where the building AUPRC was exceptionally high, in Mariupol, it is much closer to pixel-level performance. This difference is likely caused by the fact that the large buildings in the Beirut harbor were particularly easy to classify correctly. Although calibration further improves classification performance (+2.06% for pixels and +17.02% for buildings; both F1 scores), our approach performs well overall. For the same reason as previously discussed, building-level calibration also results in a larger performance improvement.

Based on the optimal classification threshold, we estimate that 2437 (22.22%) out of 10964 buildings were completely destroyed in Zhovtnevyi district, with likely many more damaged. We visualize a map for this classification with the corresponding ground truth labels in Fig.3B. Individual classification maps for each time period are visualized in Supp.10. Both PR and ROC curves are provided in Supp.11.

Table 1: Mariupol performance

	Pixels	Buildings
Annotations	421	93
AUROC	0.987 (<0.001)	0.992 (<0.001)
AUPRC	0.550 (0.020)	0.650 (0.035)
F1	0.534 (0.011)	0.558 (0.011)
	0.545 (0.014)	0.653 (0.017)
Recall	0.470 (<0.001)	0.430 (<0.001)
	0.587 (<0.001)	0.624 (<0.001)
Precision	0.619 (0.030)	0.795 (0.045)
	0.509 (0.024)	0.687 (0.038)

Note: Performance scores for Mariupol, Zhovtnevyi district, for both pixels and buildings. The annotations denote the number of pixels resp. buildings in the positive destruction class. The data points of the negative class are sampled from periods before the invasion (up to a 99:1 distribution). This process is repeated 100 times. The mean performance scores across these repetitions are reported for each measure with the standard deviation in brackets. For F1, recall and precision, the first line denotes the corresponding performance score when using the optimal probability threshold from the Beirut case study, the second line denotes the score when using the F1-optimizing threshold.

2.3 Gaza

The third case study analyzes the destruction of buildings during the ongoing Israel–Hamas war in Gaza, which began on October 7, 2023. Here, our focus lies on tracking the dynamics of the war over several months. In Fig.4 we present p-values of destruction for building pixels from September 18 to December 11, 2023. Key events are displayed in the bottom time bar. The figure highlights that the dynamics of the war are reflected in patterns of destruction. Following the initial attack on October 7, widespread airstrikes across the Gaza Strip are clearly visible in the second map. Subsequent calls for the evacuation of northern Gaza (north of the drawn evacuation border) correspond to destruction being mostly limited to these areas, as seen in maps 3 and 4. From the fifth map onward, the Israeli ground offensive becomes evident, with destruction closely tracking troop movements, initially in Gaza City (map 5), and later in Shuja'iyya and Khan Yunis (final map). We provide the maps for the remaining time periods in Supp.12.

UNOSAT regularly publishes updated damage statistics for Gaza (see Methods). Using the optimal building classification threshold from our previous case study, we compare their composite destruction estimates with ours in Table 2. Our estimates of the share of destroyed buildings in Gaza follow a trend similar to those of UNOSAT. As expected, due to the unbalanced sampling strategy and a resulting lower classification threshold to reduce false positives, our estimates are more conservative than those of UNOSAT. In the time periods before the war, we only incorrectly classify

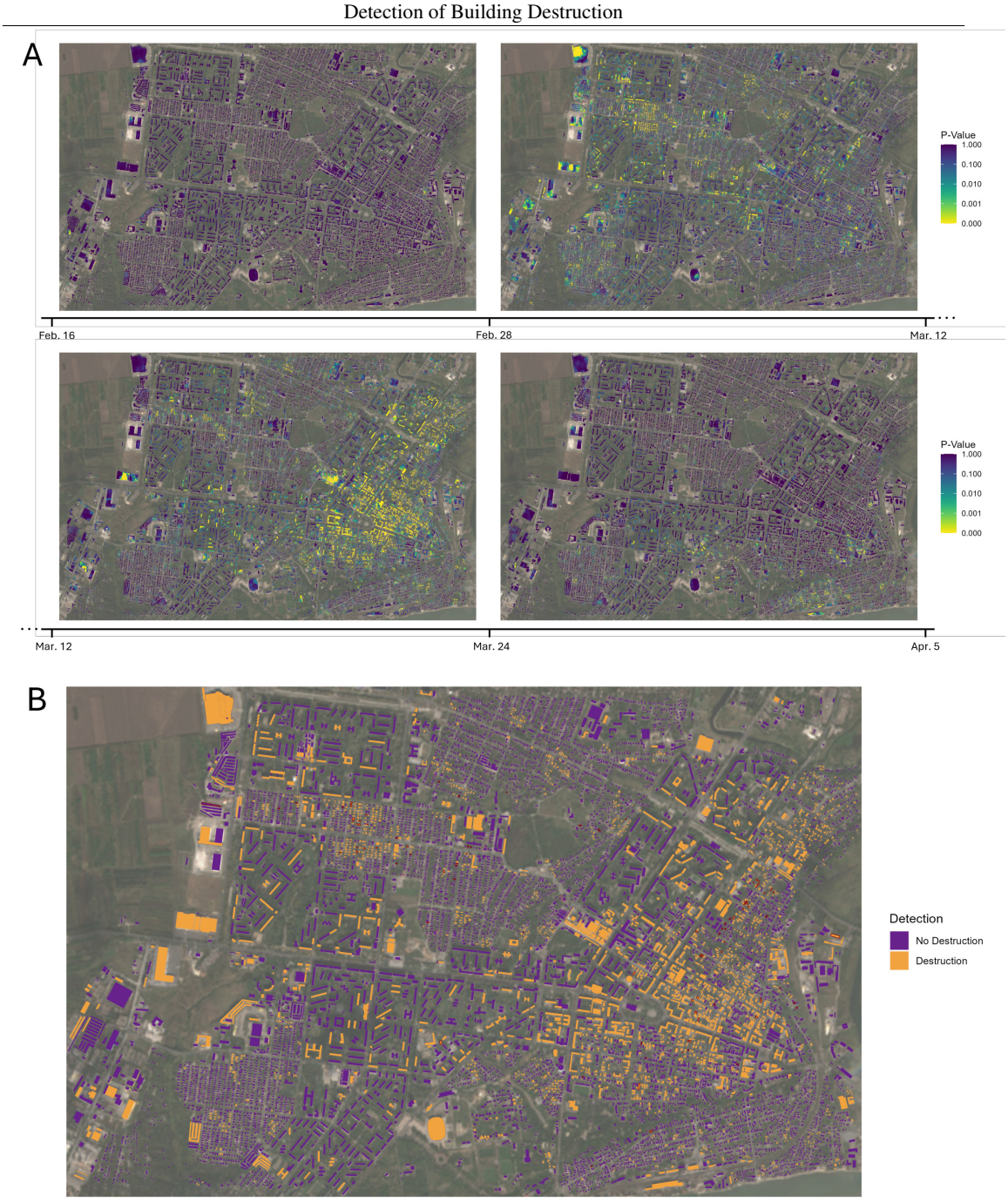


Fig. 3: (A) P-values of destruction of all 10m×10m building pixels in the center of Mariupol (Zhovtnevyi district) over 12-day periods clockwise from February 16 to February 28 (top left), February 28 to March 12 (top right), March 12 to March 24 (bottom left), March 24 to April 5 (bottom right; all 2022). Lower p-values indicate a higher likelihood that part of a building was destroyed. The Russian invasion of Ukraine started on February 24, 2022. According to reports, the siege of Mariupol intensified after March 2 [Gunter, 2022]. The background of each image is an optical Sentinel-2A image (freely available) from February 10, 2021 based on bands 4, 3 and 2. (B) Classification of entire buildings into destruction and no destruction for an F1-score-optimizing probability threshold after combining pixel-wise p-values. The classification is summarized across all 12-day time periods from February 28 to May 11, which aligns most closely with the ground truth labels from May 12. On May 16, the fall of Mariupol was officially declared by the Ukrainian army [Hopkins and Santora, 2022]. Buildings labeled by UNOSAT as fully destroyed are marked by a red border. Note, this labeling is incomplete. Many more buildings were likely destroyed.

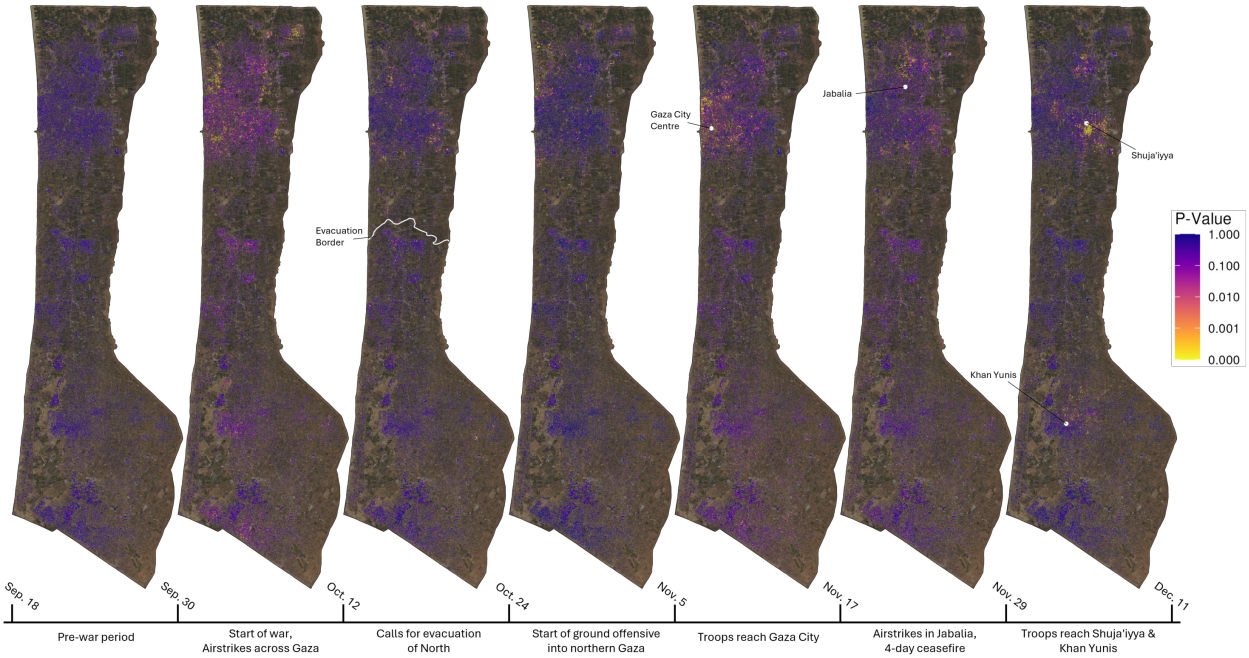


Fig. 4: P-values of destruction of all 10m×10m building pixels in Gaza over 12-day periods from September 18, to December 11, 2023. Lower p-values indicate a higher likelihood that part of a building was destroyed. The background of each image is an optical Sentinel-2A image (publicly available) from May 5, 2023 based on bands 4, 3 and 2. The timeline at the bottom denotes key events taking place between image acquisition dates.

on average 0.021% (SD = 0.010) buildings as destroyed. We provide corresponding classification maps across all time periods in Supp.13.

Table 2: Gaza cumulative building destruction estimates

UNOSAT		Ours	
Date	% destroyed	Date	% destroyed
October 10, 2023	0.84%	October 12, 2023	1.57%
November 11, 2023	2.99%	November 5, 2023	2.56%
November 26, 2023	4.44%	November 29, 2023	4.34%
January 7, 2024	10.31%	January 4, 2024	7.89%
February 29, 2024	14.25%	March 4, 2024	10.31%
April 1, 2024	15.43%	March 28, 2024	10.69%

Note: Comparison of the cumulative share of destroyed buildings estimated by UNOSAT (see Methods) compared to our estimates, based on the optimal classification threshold obtained from Mariupol. Includes all buildings across Gaza over the course of the Israel-Hamas war. The date corresponds to the day of the latest satellite image used for the destruction estimates. As these do not perfectly align due to image acquisition, we use the closest possible match.

3 Discussion

In this work, we have presented an unsupervised method for detecting building destruction in conflict zones using freely available Sentinel-1 SAR imagery, and applied it across three case studies, Beirut, Mariupol, and Gaza. We demonstrate that our approach is not only able to identify the destruction of buildings, but, unlike recently introduced techniques [Hou et al., 2024, Dietrich et al., 2025], also determine when it occurs. The unsupervised nature of our algorithm

eliminates the need for labeled training data, which is often unavailable in conflict regions, and thus also allows for an application in scenarios where timely information on destruction is of importance.

Our reliance on publicly available satellite data from the European Space Agency (ESA) makes our method easily accessible and scalable. Sentinel-1 imagery is available globally since 2016 and is completely free of use to both researchers and the public. Although our approach based on lower resolution imagery (20m) will generally not match the performance of methods using proprietary high-resolution images (e.g., 20cm), our results demonstrate that it is sufficient to detect patterns of destruction at the building level, without any financial costs for these images. In contrast, acquiring up-to-date high-resolution imagery (20–50cm) costs approximately $25–50 per km^2 per image [Planet Labs, 2025, LAND INFO, 2025]. For the entire Gaza Strip, which spans roughly 365km^2, a single day of high-resolution coverage would cost between $9,125 and $18,250, with an analysis similar to ours (28 images) costing up to $511,000. It is important to note that high-resolution images are usually not made available for regions at war [Hou et al., 2024]. High-resolution approaches also face challenges in processing the vast amounts of data required for large-scale applications. Our method, on the other hand, is computationally efficient and can be deployed at scale. Additionally, this means it can also serve as a valuable tool for initial analysis to guide decisions on where and when costly high-resolution imagery should be acquired for more detailed investigations.

An important feature of our algorithm is its use of p-values to quantify the evidence for destruction. This statistical framework provides a measure of uncertainty for the detection, and allows users to adjust thresholds based on specific priorities. For example, lower thresholds might be employed for exploratory analyses of broad damage patterns, at the cost of a higher rate of false positives. Beyond threshold adjustment, p-values offer users additional flexibility and granularity in their analyses, compared to a strict binary classification. Finally, the use of p-values enables a seamless and statistically sound transition between pixel- and building-level analyses, addressing a common limitation in the literature where these approaches are typically treated as mutually exclusive [Cheng et al., 2024].

However, limitations remain. The comparably low resolution of Sentinel-1 imagery prevents the reliable detection of lower levels of damage to buildings. Additionally, while our method performs well without ground truth data, as demonstrated, performance improvements in classification can be achieved through calibration with ground truth labels specific to the corresponding use case.

As also discussed in prior research [Mueller et al., 2021], the automated large-scale detection of destruction during war and conflict allows for the analysis of questions that were previously difficult or impossible to answer. For example, they can provide information on when, where, and which regions and types of buildings are targeted and destroyed, including schools, hospitals, and other critical infrastructure. While this supports academic research, such information is particularly valuable for governments and international organizations to help guide humanitarian aid and plan post-conflict reconstruction. By offering an objective assessment of destruction, they may also help mitigate potential biases inherent in manually conducted destruction reporting. A logical next step could be the development and deployment of an interactive online dashboard that continuously updates and visualizes destruction patterns across the world in near real-time. Such a tool would further enhance transparency and accessibility, ensuring that information on building destruction is available to a broad range of stakeholders.

4 Materials and Methods

4.1 Areas of Interest

Our areas of interest (AOIs) for analysis of both Beirut and Mariupol are based on available ground truth data (see section Destruction Labels). The considered bounding boxes (x_{min}, x_{max}, y_{min}, y_{max}) in the corresponding Universal Transverse Mercator (UTM) zones are for Beirut, UTM zone 36N with a bounding box of 730447.1, 735069, 3751937, 3754567, and for Maripol, UTM zone 37N with a bounding box of 385635.7, 391574.4, 5215443, 5219254. For Gaza, we consider the whole Gaza Strip, and use UTM zone 36N with a bounding box of 615957.4, 648847.4, 3455159, 3496499.

4.2 Satellite Data

We utilize Sentinel-1 Single Look Complex (SLC) Synthetic Aperture Radar (SAR) satellite images for our three case studies Beirut, Mariupol, and Gaza. Sentinel-1A and 1B were launched on April 3, 2014 and April 25, 2016, respectively, to a sharing sun-synchronous orbit at 693 km altitude with a repeat cycle of 12 days for a single satellite and 6 days for the two-satellite constellation. The Sentinel-1 satellites each carry a C-band SAR instrument with a central frequency at 5.405 GHz providing acquisitions in all weather and time conditions [ESA, 2025]. The SAR sensors aboard can collect images in different operation (Strip Map (SM), Interferometric Wide (IW), Extra Wide (EW), Wave (WV)) and different polarization modes (HH+HV, VH+VV, HH, VV). Synthetic Aperture Radar (SAR) technology,

specifically Interferometric SAR (InSAR), has emerged as a valuable remote sensing tool for damage detection due to its ability to operate under all weather conditions and capture information in both day and night. Interferometric SAR (InSAR) coherence, a measure of the stability or similarity between SAR images acquired over the same area at different times, has proven to be a reliable indicator of surface changes [Hu et al., 2014]. High coherence values typically indicate stable and unchanged surfaces, while low coherence values often suggest disruptions or alterations in the structure, such as vegetation growth, ground displacement, or destruction of buildings.

We download Sentinel-1 SLC images for Beirut from April 6, 2020 to December 26, 2020, for Mariupol from October 7, 2021 to July 22, 2022, and for Gaza from May 9, 2023 to April 9, 2024, using 12-day intervals for full coverage of each analysis period. The original Sentinel-1 SLC images are captured in IW mode with spatial resolutions of 20 m in the azimuth and 5 m in the ground range direction. These images are then multi-looked by a window of 4 pixels in the range direction in order to reduce speckle noise, resulting in MLC (Multi-Looked Complex) images with sampling spacings of 20 meters in both azimuth and ground range directions. All downloaded images are freely available from the Copernicus Open Access Hub of the European Space Agency (ESA) [ESA, 2024].

4.3 Coherence Scores

Interferometric coherence is calculated from two co-registered MLC images taken at the beginning and end of each 12-day time period over the same area. The complex images are multiplied pixel-by-pixel to compute coherence, considering both the amplitude and phase information. Specifically, coherence is calculated by averaging the phase differences between the two images over a local window and normalizing it by the product of their amplitudes [Bamler and Hartl, 1998],

$$\gamma = \frac{\left|\sum_{i=1}^{N} S_1(i) \cdot S_2^*(i)\right|}{\sqrt{\sum_{i=1}^{N} |S_1(i)|^2 \cdot \sum_{i=1}^{N} |S_2(i)|^2}} \tag{1}$$

where $S_1(i)$ and $S_2(i)$ are the complex pixel values from the first and second MLC image, $S_2^*(i)$ is the complex conjugate of $S_2(i)$, and N represents the number of pixels in the averaging window. The coherence value, which ranges between 0 and 1, reflects the similarity between the two images: a coherence close to 1 indicates high similarity (minimal change), while a value close to 0 indicates substantial differences, potentially caused by changes in the surface or disturbances.

Decorrelation sources in an interferometric coherence can be represented by a product of different decorrelation components [Bamler and Hartl, 1998].

$$\gamma = \gamma_{\text{SNR}} \cdot \gamma_{\text{temporal}} \cdot \gamma_{\text{spatial}} \tag{2}$$

where, γ_{SNR} is the noise decorrelation due to the signal-to-noise ratio (SNR). γ_{temporal} accounts for temporal decorrelation resulting from changes in the ground surface between the acquisition times of the two MLC images. γ_{spatial} represents spatial or baseline decorrelation depending on the geometry of the satellite passes, particularly the spatial baseline between the two acquisitions. Each decorrelation component takes a value between 0 and 1, with 1 indicating no decorrelation and 0 indicating complete decorrelation. The overall coherence γ is thus reduced when any of these decorrelation sources are present.

When buildings or infrastructure are damaged or destroyed between two image acquisitions, the change in surface structure results in a loss of coherence due to temporal decorrelation. In order to detect these temporal changes, SNR and spatial decorrelations are approximated through the use of both Digital Elevation Model (DEM) and satellite trajectories based on Interferometric SAR (InSAR) [Yagüe-Martínez et al., 2016, Bamler and Hartl, 1998, Moreira et al., 2013], leaving only the temporal decorrelation as the dominant component in the interferometric coherence. The final coherence scores, which are based on sampling spaces of 20m, are then resampled and projected into UTM, using the corresponding UTM zone of each AOI, to a pixel size of 10m$\times$10m in order to limit the possible loss of information due to the projection.

4.4 Detection of Destruction

To reliably identify damage and destruction from changes in the coherence, we analyze the first differences of the coherence values for each pixel over time. For each pixel, we fit a non-parametric median regression with a flexible trend to these differences [Koenker, 2005]. Median regression is employed to reduce the influence of outliers on the fitted trend, including those resulting from actual destruction. The flexible trend allows us to account for gradual deviations in coherence, such as those caused by atmospheric changes or other non-structural variations.

We then calculate the differences between observed values and fitted trend, i.e., the residuals. Generally, large residuals can be classified as outliers and are likely due to changes in the building's structure, such as those caused by damage or

Table 3: Dataset Summary

Case Study	Pixels	Buildings	# of Time Periods
Beirut	57,581	6,915	22
Mariupol	67,737	10,964	24
Gaza	859,502	172,916	28

Note: Number of pixels, buildings and time periods considered for each case study. The corresponding number of pixels and buildings refers to a single 12-day time period. Each pixel and building is observed over the entire analysis period. Due to the use of first differences in the coherence scores, the first time period drops out.

destruction. However, since noise levels vary between pixels, this classification is challenging. We solve this by using a robust estimator of the standard deviation, Q_n, first discussed in Rousseeuw and Croux [1993].

We find that residual distributions for each pixel to be approximately normally distributed, though with substantially larger tails due to outliers that clearly do not follow the same distribution (see Supp.15). Hence, to estimate the standard deviation of the residuals of each pixel, while being minimally affected by these outliers, we use Q_n, defined as

$$Q_n = d\{|x_i - x_j|; i < j\}_{(k)}, \tag{3}$$

where d is a constant and $k = \binom{h}{2}$, with $h = [\frac{n}{2}] + 1$. This means, we take the k^{th} order statistic of all $\binom{n}{2}$ pairwise distances. The multiplication with d is chosen accordingly to achieve consistency. As demonstrated in Rousseeuw and Croux [1993], this estimator has a low error-sensitivity while simultaneously achieving a high efficiency.

Under the hypothesis of a normal distribution with the estimated standard deviation $\hat{\sigma}_i$, we can now derive the probability to observe each residual $r_{i,t}$ for a pixel i in time period t. Specifically, we derive the probability to observe a given residual or a more extreme one in the negative direction, i.e. a one-sided p-value with

$$p(r_{i,t}) = Pr(R_{i,t} \leq r_{i,t}) \tag{4}$$

where $r_{i,t}$ is the observed residual and $R_{i,t} \sim \text{Normal}(0, \hat{\sigma}_i^2)$ a random variable. The negative direction is required, as we are only interested in drops in coherence over time.

This p-value provides evidence and thus certainty for how likely it is that a given coherence score is not just observed by chance and instead due to structural changes of the building. Higher levels of damage or destruction should lead to lower p-values. These p-values can either be used directly to visualize likely buildings and areas of destruction, or specific cut-offs can be chosen at which pixels are classified as destroyed.

Using the available ground truth data for Beirut, we experimented with different levels of flexibility in the regression (see Supp.14), different robust scale estimators (see Supp.16), and directly using the coherence scores in the regression instead of their first differences (see Supp.17), before we ultimately decided on the described setup. Note, fitting these median regressions can be carried out at scale, as the computation is straightforward to parallelize. This means even for the Gaza case study, computation only took roughly two hours on a single machine.

To classify entire buildings, the p-values of each building need to be combined. We do this by constructing the weighted harmonic mean p-value (HMP) [Wilson, 2019] defined as

$$\overline{p}_{B,t} = \frac{\sum_{j=1}^{N} w_j}{\sum_{j=1}^{N} \frac{w_j}{p_{j,t}}}, \tag{5}$$

where B refers to the building, t to the time period, N to the number of pixels that make up the building, $p_{j,t}$ the p-value of a pixel j that is part of building B, and w_j its corresponding weight based on building coverage. The HMP is often for meta-analyses [Errington et al., 2021], and robust to the expected positive dependencies between the p-values of each building.

4.5 Building Footprints

To identify individual buildings, we use building footprints from OpenStreetMap (OSM), widely used for applications in urban planning [Milojevic-Dupont et al., 2020], public health [Sturrock et al., 2018], and disaster management [Poiani et al., 2016]. We obtain these footprints by querying OSM through the *building* key on the first day of the analysis period of each use case, ensuring that each building remains consistent throughout the entire analysis period. This approach prevents any changes, e.g., due to destruction, from being reflected in the building footprints. Using this building information allows us to only analyze those pixels that constitute buildings. We provide a dataset summary in Table 3.

4.6 Destruction Labels

For the Beirut harbor explosion we utilize georeferenced information on destruction from Kondmann et al. [2021]. The authors draw on ground truth data from the Center for Satellite Based Crisis Information (ZKI) at the German Aerospace Center, who manually annotated buildings based on high-resolution satellite images and field reports. Each labeled building represents a structure at the harbor fully destroyed by the explosion. Note, multiple studies [Pilger et al., 2021, Al-Hajj et al., 2021] have documented destruction and substantial damage to buildings located farther from the explosion site.

For Mariupol we use building damage labels in the city center, Zhovtnevyi district, from UNOSAT, part of the United Nations Institute for Training and Research (UNITAR) [UNITAR, 2024]. Selected buildings are annotated manually based on high-resolution (30cm) optical satellite images from Maxar WorldView-3 on May 12, 2022, with varying levels of visible damage. Note, contrary to Beirut, these annotations are not field validated. Hence, we only retain those annotations marked as 'high confidence' and labeled as destruction. As each annotation is only recorded through a single set of geographic coordinates, we match these to our building footprints to annotate entire buildings.

For the composite damage statistics in Gaza we similarly utilize information provided by UNOSAT, based on high-resolution (30-50cm) satellite images from both Maxar and the French national space agency CNES. Theoretically, manual annotations with exact coordinates are available, however, in practice, we find these coordinates to be imprecise. In a vast amount of cases we cannot successfully match these to our building footprints and hence decided for a composite analysis only. We provide more details in Supp.18.

5 Data Availability

All datasets, except SAR imagery, are available on the OSF [Racek et al., 2025a]. Raw satellite images are not provided due to their large file size. However, they are freely available from the Copernicus Open Access Hub of the European Space Agency (ESA) [ESA, 2024]. Additionally, full reproducibility for the remaining parts of the analysis is possible, as all intermediate results and datasets are also provided.

6 Code Availability

Interferometric coherence scores were calculated using Python 3.7 [Van Rossum and Drake Jr, 1995] and Gamma [Werner et al., 2000]. The coherence scores were further processed using R 4.3.3 [R Core Team, 2013]. A complete list of all used libraries and their corresponding versions is available in the project's GitHub repository [Racek et al., 2025b]. There, we also provide all code needed to reproduce the results.

Acknowledgments

This work is jointly supported by the Helmholtz Association under the joint research school 'Munich School for Data Science - MUDS' and the TUM Innovation Network EarthCare.

References

Zhengyang Hou, Ying Qu, Liqiang Zhang, Jun Liu, Faqiang Wang, Qiwei Yu, An Zeng, Ziyue Chen, Yuanyuan Zhao, Hong Tang, et al. War city profiles drawn from satellite images. *Nature Cities*, 1(5):359–369, 2024.

Matthias Kahl and Zhaiyu Chen. Towards automated building damage detection in the gaza strip: Contextual analysis since october 2023. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 2668–2671. IEEE, 2024.

Olivier Dietrich, Torben Peters, Vivien Sainte Fare Garnot, Valerie Sticher, Thao Ton-That Whelan, Konrad Schindler, and Jan Dirk Wegner. An open-source tool for mapping war destruction at scale in ukraine using sentinel-1 time series. *Communications Earth & Environment*, 6(1):1–10, 2025.

Guangliang Cheng, Yunmeng Huang, Xiangtai Li, Shuchang Lyu, Zhaoyang Xu, Hongbo Zhao, Qi Zhao, and Shiming Xiang. Change detection methods for remote sensing in the last decade: A comprehensive review. *Remote Sensing*, 16(13):2355, 2024.

Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and remote sensing magazine*, 1(1):6–43, 2013.

Maxar Technologies. Beirut explosion, 2024. URL https://www.maxar.com/open-data/beirut-explosion. https://www.maxar.com/open-data/beirut-explosion, Retrieved 2024-12-18.

Richard Bamler and Philipp Hartl. Synthetic aperture radar interferometry. *Inverse problems*, 14(4):R1, 1998.

Néstor Yagüe-Martínez, Pau Prats-Iraola, Fernando Rodriguez Gonzalez, Ramon Brcic, Robert Shau, Dirk Geudtner, Michael Eineder, and Richard Bamler. Interferometric processing of sentinel-1 tops data. *IEEE transactions on geoscience and remote sensing*, 54(4):2220–2234, 2016.

Ayesha Shafique, Guo Cao, Zia Khan, Muhammad Asad, and Muhammad Aslam. Deep learning-based change detection in remote sensing images: A review. *Remote Sensing*, 14(4):871, 2022.

Simon Plank. Rapid damage assessment by means of multi-temporal sar—a comprehensive review and outlook to sentinel-1. *Remote Sensing*, 6(6):4870–4906, 2014.

Chongyuan Zhang, Afef Marzougui, and Sindhuja Sankaran. High-resolution satellite imagery applications in crop phenotyping: An overview. *Computers and Electronics in Agriculture*, 175:105584, 2020.

Salman H Khan, Xuming He, Fatih Porikli, and Mohammed Bennamoun. Forest change detection in incomplete satellite images with deep neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):5407–5423, 2017.

Batuhan Osmanoğlu, Filiz Sunar, Shimon Wdowinski, and Enrique Cabral-Cano. Time series analysis of insar data: Methods and trends. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:90–102, 2016.

Wenzhong Shi, Min Zhang, Rui Zhang, Shanxiong Chen, and Zhao Zhan. Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote Sensing*, 12(10):1688, 2020.

Hannes Mueller, Andre Groeger, Jonathan Hersh, Andrea Matranga, and Joan Serrat. Monitoring war destruction from space using machine learning. *Proceedings of the national academy of sciences*, 118(23):e2025400118, 2021.

Valerie Sticher, Jan D Wegner, and Birke Pfeifle. Toward the remote monitoring of armed conflicts. *PNAS nexus*, 2(6): pgad181, 2023.

Christoph Pilger, Peter Gaebler, Patrick Hupe, Andre C Kalia, Felix M Schneider, Andreas Steinberg, Henriette Sudhaus, and Lars Ceranna. Yield estimation of the 2020 beirut explosion using open access waveform and remote sensing data. *Scientific reports*, 11(1):14144, 2021.

Samar Al-Hajj, Hassan R Dhaini, Stefania Mondello, Haytham Kaafarani, Firas Kobeissy, and Ralph G DePalma. Beirut ammonium nitrate blast: analysis, review, and recommendations. *Frontiers in public health*, 9:657996, 2021.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

Daniel J Wilson. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200, 2019.

Timothy M Errington, Maya Mathur, Courtney K Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. Investigating the replicability of preclinical cancer biology. *Elife*, 10:e71601, 2021.

Holly Ellyatt. Mariupol hasn't surrendered to russia, pm says; at least 5 dead, 20 injured in kharkiv attack, 2022. URL https://www.cnbc.com/2022/04/17/russia-ukraine-live-updates.html. https://www.cnbc.com/2022/04/17/russia-ukraine-live-updates.html, Retrieved 2024-11-21.

Joel Gunter. Ukrainian city of mariupol 'near to humanitarian catastrophe' after bombardment, 2022. URL https://www.bbc.com/news/world-europe-60585603. https://www.bbc.com/news/world-europe-60585603, Retrieved 2024-11-21.

Valerie Hopkins and Marc Santora. Ukraine signals end of bitter battle at azovstal steel plant, 2022. ISSN 1553-8095. URL https://www.nytimes.com/2022/05/16/world/europe/azovstal-mariupol.html. https://www.nytimes.com/2022/05/16/world/europe/azovstal-mariupol.html, Retrieved 2024-11-21.

Planet Labs. Planet pricing, 2025. URL https://www.planet.com/pricing/. https://www.planet.com/pricing/, Retrieved 2025-02-04.

LAND INFO. Buying satellite imagery: Pricing information for high resolution satellite imagery, 2025. URL https://landinfo.com/satellite-imagery-pricing/. https://landinfo.com/satellite-imagery-pricing, Retrieved 2025-02-04.

ESA. Sentinel-1 mission, 2025. URL https://sentiwiki.copernicus.eu/web/s1-mission. https://sentiwiki.copernicus.eu/web/s1-mission, Retrieved 2024-01-28.

Jun Hu, ZW Li, XL Ding, JJ Zhu, Lei Zhang, and Qian Sun. Resolving three-dimensional surface displacements from insar measurements: A review. *Earth-Science Reviews*, 133:1–17, 2014.

ESA. Copernicus open access hub, 2024. URL https://scihub.copernicus.eu/. https://scihub.copernicus.eu/, Retrieved 2024-11-22.

Roger Koenker. Quantile regression. *Cambridge University Press*, 2005.

Peter J Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.

Nikola Milojevic-Dupont, Nicolai Hans, Lynn H Kaack, Marius Zumwald, François Andrieux, Daniel de Barros Soares, Steffen Lohrey, Peter-Paul Pichler, and Felix Creutzig. Learning from urban form to predict building heights. *PLOS one*, 15(12):e0242010, 2020.

Hugh JW Sturrock, Katelyn Woolheater, Adam F Bennett, Ricardo Andrade-Pacheco, and Alemayehu Midekisa. Predicting residential structures from open source remotely enumerated data using machine learning. *PloS one*, 13 (9):e0204399, 2018.

Thiago Henrique Poiani, Roberto Dos Santos Rocha, Lívia Castro Degrossi, and Joao Porto De Albuquerque. Potential of collaborative mapping for disaster relief: A case study of openstreetmap in the nepal earthquake 2015. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 188–197. IEEE, 2016.

Lukas Kondmann, Aysim Toker, Sudipan Saha, Bernhard Schölkopf, Laura Leal-Taixé, and Xiao Xiang Zhu. Spatial context awareness for unsupervised change detection in optical satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2021.

UNITAR. Maps and data, 2024. URL https://www.unitar.org/maps. https://www.unitar.org/maps, Retrieved 2024-10-29.

Daniel Racek, Qi Zhang, Paul Wilhelm Thurner, Xiao Xiang Zhu, and Göran Kauermann. Data repository: Detection of building destruction in armed conflict from publicly available satellite imagery, 2025a. URL https://osf.io/kw5g9/. https://osf.io/kw5g9/.

Guido Van Rossum and Fred L Drake Jr. *Python tutorial*, volume 620. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

Charles Werner, Urs Wegmüller, Tazio Strozzi, and Andreas Wiesmann. Gamma sar and interferometric processing software. In *Proceedings of the ers-envisat symposium, Gothenburg, Sweden*, volume 1620, page 1620. Citeseer, 2000.

R Core Team. R: A language and environment for statistical computing. *Foundation for Statistical Computing, Vienna, Austria*, 2013.

Daniel Racek, Qi Zhang, Paul Wilhelm Thurner, Xiao Xiang Zhu, and Göran Kauermann. Repository: Detection of building destruction in armed conflict from publicly available satellite imagery, 2025b. URL https://github.com/Daniel-Rac/Detecting_destruction_from_public_satellite_images. https://github.com/Daniel-Rac/Detecting_destruction_from_public_satellite_images.

8. The Russian war in Ukraine increased Ukrainian language use on social media

Contributing article

Racek, D., Davidson, B. I., Thurner, P. W., Zhu, X. X., and Kauermann, G. (2024). The Russian war in Ukraine increased Ukrainian language use on social media. *Communications Psychology, 2(1), 1.* https://doi.org/10.1038/s44271-023-00045-6.

Data and code

Available at https://osf.io/48sbc.

Copyright information

Supplementary material

Supplementary material is available online.

Author contributions

The original idea to analyse the language change of tweets in Ukraine stems from Daniel Racek. Xiao Xiang Zhu provided the initial dataset. Daniel Racek collected additional tweets, processed them and designed the bot detection. Daniel Racek and Göran Kauermann designed the generalized additive models to analyse the language shift of users. Daniel Racek conducted the study, analysed the results and created the visualizations. Daniel Racek, Brittany Davidson, Paul Thurner and Göran Kauermann wrote the first draft of the article. All authors were involved in improving and editing the article.

communications psychology

ARTICLE

https://doi.org/10.1038/s44271-023-00045-6 OPEN

The Russian war in Ukraine increased Ukrainian language use on social media

Daniel Racek [1]✉, Brittany I. Davidson [2], Paul W. Thurner [3], Xiao Xiang Zhu [4] & Göran Kauermann[1]

The use of language is innately political, often a vehicle of cultural identity and the basis for nation building. Here, we examine language choice and tweeting activity of Ukrainian citizens based on 4,453,341 geo-tagged tweets from 62,712 users before and during the Russian war in Ukraine, from January 2020 to October 2022. Using statistical models, we disentangle sample effects, arising from the in- and outflux of users on Twitter (now X), from behavioural effects, arising from behavioural changes of the users. We observe a steady shift from the Russian language towards Ukrainian already before the war, which drastically speeds up with its outbreak. We attribute these shifts in large part to users' behavioural changes. Notably, our analysis shows that more than half of the Russian-tweeting users switch towards Ukrainian with the Russian invasion. We interpret these findings as users' conscious choice towards a more Ukrainian (online) identity and self-definition of being Ukrainian.

[1] Institute of Statistics, Ludwig-Maximilians-University Munich, Munich, Germany. [2] School of Management, University of Bath, Bath, UK. [3] Institute of Political Science, Ludwig-Maximilians-University Munich, Munich, Germany. [4] School of Engineering and Design, Technical University of Munich, Munich, Germany. ✉email: daniel.racek@stat.uni-muenchen.de

Social media is critically important in today's society[1–3]. In recent years, it has played a key role in a number of political shifts and crises[4,5]. While social media has been found to amplify all manners of misinformation, propaganda, populism, and xenophobia[6–8], it can also serve as a mechanism to call for aid and as a source for live updates of major events unfolding[9–12].

In this article, we analyse language use of Ukrainian citizens on social media before and during the Russian invasion of Ukraine (subsequently referred to as war), where after years of tensions and open aggression between Russia and Ukraine[13], on 24th February 2022, Russian forces began to invade and occupy parts of Ukraine[14]. At the time of writing, it has been estimated that the war has led to over 23,000 civilian casualties[15] and hundreds of billions of dollars worth of damage[16,17]. This has caused worldwide unrest, alongside 8.2 million Ukrainian refugees recorded across Europe and 5 million registered for temporary protection[18,19].

The war in Ukraine is also taking place in the digital era, with social media coverage documenting the horrific events in up to real-time. This provides a unique digital trace of many first-hand accounts of the war, as citizens are communicating among each other and to the public. This is generally known as crisis informatics, whereby social media data are utilized before, during, or after emergency events for use cases such as disaster monitoring, management, and prevention[9,12,20–22]. Recent studies have demonstrated that tweets can capture events of political violence[23] and can help in monitoring and understanding intra-country conflicts[24].

In our work, the user's choice of language in a tweet is of particular interest. Many people across the world (including most Ukrainian citizens with the Russian and Ukrainian language[25]) are multilingual. This multilingualism comes with a number of links to an individual's identity, as someone may speak one language at work, but another one at home with their family. Thus, different languages are spanning across multiple facets of one's identity[26]. These context-based adaptations of our self-presentation and behaviour are expected by those around us[27–30]. Hence, it is important to note that a user's choice of language online can be argued as an active choice to communicate and a way they seek to present themselves to their audience[26]. For example, many non-natives switch to English in order to ensure a wider intelligibility online[31].

The use of language is also inherently political. Languages can be the cause of conflict and they are often incorporated in cultural and ethnic identity definition and are the basis for nation building and political change[32,33]. After the dissolution of the USSR, most post-soviet countries introduced new language laws in order to assert their original native language and build a new nation[32,34]. In Ukraine, after their independence, many people were considering themselves Russians by nationality or Ukrainian with Russian as their main native language[35,36]. While the government aimed to reverse those effects, they were only moderately successful in achieving this goal, as census results show[35–37]. Only more recently, with the Euromaidan protests and the Russian military intervention in Crimea and the Donbas, surveys between 2012 and 2017 show a consistent and substantial shift away from Russian ethnic and linguistic identification towards Ukrainian practice[37].

We investigate language choice and tweeting activity on Ukrainian Twitter (now called X) from January 2020 to November 2022 using over 4 million geo-tagged tweets from more than 62,000 different users. In doing this, we study how Ukrainian citizens (and non-citizens living there) respond to their country being aggressively attacked and invaded by its direct neighbour they share a long history and language with, and how the use of language evolved before and during this war. Our study allows us to follow the same set of users and observe their (change in) behaviour over both the short- and longer-term as the war breaks out and continues to unfold on an individual level. Hence, we are able to comment on recent news articles outlining shifts in language use from Russian to Ukrainian as a direct result of the war[38,39]. Moreover, we are able to monitor long-term language trends even before the war without the necessity of relying on small-scale surveys nor the infrequent censuses, the last one of which was conducted in 2001.

More specifically, we study overall trends in the number of tweets in the three main languages (Ukrainian, Russian, English) over time. Second, we investigate how these trends translate to users' individual tweeting activity and if changes result from the in- and outflux of users, common in online communities[40–42], or if they result from users changing their behaviour over time[43–45]. We quantify the magnitude of both effects respectively. Third, we study if changes in users' tweeting activity originate from shifts between languages and quantify the magnitude of these shifts. Fourth and finally, we take a closer look at those users that switch from predominately tweeting in Russian to predominately tweeting in Ukrainian with the outbreak of the war.

Methods

This study was ethically approved by the ethics commission of the faculty of mathematics, computer science and statistics at Ludwig-Maximilians-Universität (LMU) München, Germany. The reference identifier is EK-MIS-2022-127. We did not pre-register this study. No information on user demographics such as age, sex, gender or race were collected or determined and - in accordance with the ethics commission - no informed consent by the Twitter users was obtained.

Data

Data collection & final dataset. We collected tweets from 9th January 2020 to 12th October 2022 using the 1% real-time stream of the Twitter API. During collection, we filtered the data such that we only gathered tweets containing geo-information from the API. We then manually filtered the dataset to only retain tweets from Ukraine (denoted by the "UA" country tag), as common in the literature[46], and excluded any retweets, which left us with primary tweets, quotes and replies, all of which contain original tweet texts.

This dataset obtained from the 1% stream consisted of 4,102,982 tweets. As we began cleaning, we noticed gaps with missing tweets, most likely due to server and internet outages during the real-time data collection process. Hence, we retrospectively identified and filled all gaps. To do this, we first identified all time windows >10 min without any tweet and added them to our download queue. Days with more than two of such time windows were added to the queue as a whole. We then queried the Twitter Research API 2.0 using the *tweets/search/all* endpoint to obtain tweets with Ukrainian geoinformation for all time windows in this queue and added the newly obtained tweets to our original dataset. Finally, we repeated this process for the 15 days with the least amount of tweets in our dataset. After removing all duplicates, this meant we added a total of 350,359 additional tweets to our dataset this way. Our subsequently conducted sensitivity analysis shows that through the two-stage filtering process combined with the recollection efforts, we were able to recover almost all geo-tagged tweets from Ukraine during this time period (see section "Sensitivity Analysis" for more info).

We conducted an extensive spam filtering scheme, in which we (1) removed any duplicate tweets, (2) identified and removed potential spam bots by training a bot detection model following[47], (3) removed users with >100 tweets per day, (4) only kept tweets

COMMUNICATIONS PSYCHOLOGY | https://doi.org/10.1038/s44271-023-00045-6 ARTICLE

coming from official Twitter clients or Instagram, and (5) applied additional filtering rules specific to our dataset. This reduced our dataset from originally 4,453,341 tweets (62,712 users) down to 2,845,670 tweets (41,696 users). For a more extensive description and rationale see section "Data Cleaning".

User characteristics. Unsurprisingly, social media is popular in Ukraine, particularly among the younger generation, with almost all citizens aged 18-39 in 2021 reporting that they use social media. For Twitter, user statistics are as follows: 18–29 (13% usage), 30–39 (8%), 40–49 (7%), 50+ (1%)[48].

We provide an overview and descriptive statistics on all user attributes as available from the API in Supplementary Table 4. The relevant user attributes for our main results and their assigned names are described in the following. Followers are the number of accounts that follow a user. Followings reports the number of accounts a user is following. The account age the number of months a user account has existed from account creation to their latest tweet in our dataset. The tweet frequency the number of tweets per day. The like frequency the number of liked tweets (by the user) per day. # of Tweets in Ukraine reports the total number of tweets in our dataset. All Twitter user attributes are a snapshot from the last time we observe a user's respective tweet in our sample.

As described in Supplementary Notes 1, we conduct multilingual topic modelling using BERTopic[49]. War topic 1 reports the number of tweets assigned to first war topic cluster (topic #1), which covers updates about the war and calls for help. War topic 1 (rel.) the relative share of tweets assigned to this topic. War topic 2 reports the number of tweets assigned to second war topic cluster (topic #3), which covers a more political side of the overall conflict. War topic 2 (rel.) the relative share. A full list of all topic clusters is available in Supplementary Table 1.

Sensitivity analysis. After data collection (before the cleaning), we evaluated the completeness of the dataset, i.e. whether we were able to recover most of the tweets published in Ukraine over the course of the study period, using the following strategy. We draw a random subset of 29 days from our analysis period and draw tweets from the Twitter Research API 2.0 using the *tweets/search/all* endpoint, which returns all historic tweets that have not been deleted since. We find a coverage of 98.24% (SD: 3.09%). More importantly, in the opposite direction we are only able to report a coverage of 77.67% (SD: 9.55%). Hence, employing our strategy using the real-time stream offers substantially more tweets, which have been deleted since (for more information on tweet deletion and its effects see ref. 50). Moreover, this suggests we were able to recover most of the geo-tagged tweets from Ukraine.

Data cleaning. For cleaning our dataset, we first train a Twitter bot detection model using a random forest (RF), as described in ref. 47. We use the exact same model as described in the authors' work (except for removing the attribute *profile_use_background_image*, which is no longer available from the Twitter API), using the training datasets *botometer-feedback*, *celebrity*, *political-bots*, as well as 100 manually labelled Twitter accounts from our dataset. To evaluate performance, we first set up a nested cross validation (CV) routine, with both a fivefold CV in the inner and outer loop. The inner CV is used for hyperparameter tuning, tuning both the number of trees as well as the minimum node size of the RF, whereas the outer loop is used for evaluating model performance. This results in an average area under the receicer operator characteristics curve (AUROC) of 0.9837 and an average area under the precision-recall curve (AUPRC) of 0.7707. For our final model, we replicate this procedure, by setting up a 5-fold CV on the entire dataset to find the best performing hyperparameters. We then train our RF on the entire dataset and use this model to identify bots and spam accounts in our dataset.

As we are only interested in removing the most prevalent spam, we opt for a conservative removal strategy to not falsely remove too many real and non-spam users. Hence, we only remove users with a predicted bot probability >50% and more than 10 tweets since account creation as well as users with a predicted bot probability >30% and more than 10,000 tweets. While thresholds of 50% and 30% respectively might not seem conservative, in the given setting, in which the bot class is heavily underrepresented (3.7% of observations in training dataset), an F1-optimizing threshold on the training dataset would lie far below that. We are somewhat less conservative with users that published over 10000 tweets, as in most cases they are spam accounts (e.g. related to bitcoins or NFTs). We do to not remove users with less than 11 tweets, as even for a human it becomes incredibly difficult to determine if a user is a bot with such limited amount of information to draw from. At the same time, we noticed a large influx of new users after the outbreak of the war who exclusively called for help in a short span of time, a behaviour which can easily be mistaken for a bot. Notably, we do not tune the optimal classification threshold, as the outbreak of the war in Ukraine represents an unprecedented event, with an unusual amount of new users joining (see Section "User Activity"). Hence, we expect the distribution between the target label (bot or human) and our features to be different between the bot training dataset and our Ukrainian dataset. Unfortunately, an extensive manual labelling strategy and more elaborate bot detection is beyond the scope of this work and would warrant its own paper. In summary, with this strategy we remove a total of 2021 users and their tweets from our dataset.

To further identify and remove potential spam accounts, we identify all accounts with more than 100 tweets on a single day (the mean is ~ 4.4 and the median = 2), and remove those 257 users from the dataset. We also noticed an unusual amount of Tweets containing the word "BTS" (45,579; referring to the Korean K-Pop band[51]) with spikes on specific days, which we subsequently filter out. Next, we identify and remove any tweets published by the same user that contain the exact same text as their previous tweet if both tweets were published within a one minute window. Fifth and finally, we filter out any tweets with the *source* attribute not being equal to Instagram or Twitter. That way, we discard any tweets automatically published by social media schedulers such as dlvr, which are often used by news agencies or other companies.

Statistical modelling

Tweet modelling. We define the number of tweets $Y_{t,u,l}$ made in week t by user u in language l. As tweets are count data, we model the $Y_{t,u,l}$ to follow a Poisson distribution with intensity $\lambda_{t,u,l}$, where

$$\lambda_{t,u,l} = \exp(\mu + s_l(t) + W_{u,l}). \quad (1)$$

Here, μ is a general time-constant intercept, which captures the average tweet intensity over all users, languages and weeks. The $W_{u,l}$ are language-specific time-constant random intercepts for each user u, assumed to be normally distributed. They capture by how much the average tweeting behaviour (more or less tweets) of each user in each language differs from the general mean μ. Finally, $s_l(t)$ denotes a smooth global time trend for each language l (Ukrainian, Russian, English) and captures changes in the tweeting behaviour over all users over time. Hence, with the latter, we can measure behavioural changes of the users over time (e.g. are users tweeting more with the outbreak of the war?),

whereas the random intercepts measure changes in the user sample over time (e.g. are users that enter the platform after the war tweeting more on average?). This results in a generalized additive mixel model (GAMM). For more information, we refer the reader to ref. 52 and ref. 53. We fit the model with the R package *mgcv* v1.8.41[54] using the GAM implementation for very large datasets *bam*. To speed up the estimation, we use the discrete option, which discretizes covariates to ease storage and increase efficiency. For fitting $s_l(t)$, we employ thin plate regression splines. Our estimation sample consists of $y = 1{,}045{,}245$ observations, with $t = 143$ weeks, $l = 3$ languages and $u = 13{,}643$ users. For our fitted model, we report an explained deviance of 71.3%.

The effect sizes in the results are calculated as follows. For the behavioural effects we derive the change in $s_l(t)$ between two respective dates t_1 and t_2 and take the exp(.), i.e. $\exp(s_l(t_2) - s_l(t_1))$ for each language l. The result is the change in expected tweeting activity due to behavioural changes, when controlling for the in- and outflux of users. The sample effects are derived by averaging the random effects of the active users at the two respective dates and taking the exp(.), i.e. $\exp(\overline{W}_{t_2,l} - \overline{W}_{t_1,l})$. We define $\overline{W}_{t,l}$ as the average random effect in language l over all users u active at time point t. This captures the averaged change in expected tweeting activity due to a change in average tweeting intensity of the active users, when controlling for behavioural changes.

Language modelling. To model users' pairwise language probability, we refrain from a multinomial modelling strategy, as even with a weekly setup our dataset is particularly large. (To the best of our knowledge, a package with a parallel estimation routine for large datasets that can fit a GAMM for a multinomial distribution does not exist.) Instead, we model each pairwise probability separately through a binomial distribution. Our pairwise evaluation gives us a total of three different language pairs (UA over RU, UA over EN, RU over EN), for which we model the probability π to tweet in language one (subsequently l_1) over language two (subsequently l_2). The order in which we specify these pairs is irrelevant, as the probability to tweet in l_2 over l_1 is simply $1 - \pi$. More specifically, we define $X_{t,u}$ as the number of tweets made in week t by user u in l_1. We assume $X_{t,u} \sim Binomial(n_{t,u}, \pi_{t,u})$, where $n_{t,u}$ denotes the total number of tweets made by user u in week t (sum of tweets in l_1 and l_2) and $\pi_{t,u}$ corresponds to the probability to tweet in l_1 over l_2. We assume that $n_{t,u}$ is known and instead model $\pi_{t,u}$ by setting

$$\pi_{t,u} = f(\mu + s(t) + W_u), \tag{2}$$

where $f(.)$ is defined as the logistic function. Similarly to before, μ is a general time-constant intercept, which captures the average mean probability over all users and weeks to tweet in l_1 over l_2. Again, the W_u are time-constant random intercepts for each user u that capture by how much the average probability differs from the general mean μ, and are assumed to be normally distributed. The smooth global time trend $s(t)$ captures changes in the probability over all users over time. Hence, as before, we can measure behavioural changes of the users over time with the latter (are users actively changing the language they are tweeting in?), whereas the random intercepts measure changes in the sample over time (how does the language probability of users entering/leaving the platform evolve?). We estimate this model specification for all three aforementioned language-pairs with the R package *mgcv* v1.8.41[54] using the GAM implementation for very large datasets *bam*. To speed up the estimation, we use the discrete option, which discretizes covariates to ease storage and increase efficiency. For fitting $s(t)$, we employ thin plate regression splines. Users not tweeting in either of the two languages of the respective language pair, need to be discarded by definition. Hence, for UA over RU our estimation sample consists of of $x = 194{,}178$ observations, with $t = 143$ weeks and $u = 10{,}531$ users. For UA over EN: $x = 146{,}984$, $t = 143$, $u = 9{,}133$. For RU over EN: $x = 170{,}853$, $t = 143$, $u = 10777$. For our fitted models, we report explained deviances of: 85.8% (UA over RU), 90.5% (UA over EN) and 90% (RU over EN).

The coefficients of a logistic regression, as employed here, must be interpreted with respect to changes in the odds (also known as odds ratio). The odds ratio is defined as $odds = p/(1-p)$. Hence, it describes how likely an event is going to happen compared to not happen. In this setting, it describes how likely it is to tweet in language 1 over language 2.

The effect sizes in the results are calculated as follows. For the behavioural effects we derive the change in $s(t)$ between two respective dates t_1 and t_2 and take the exp(.), i.e. $\exp(s(t_2) - s(t_1))$ for each of the three models. The result is the change in odds to tweet in l_1 over l_2 due to behavioural changes, when controlling for the in- and outflux of users. The sample effects are derived by averaging the random effects of the active users at the two respective dates and taking the exp(.), i.e. $\exp(\overline{W}_{t_2} - \overline{W}_{t_1})$ for each of the three models. We define $\overline{W}_t$ as the average random effect over all users u active at time point t. This captures the averaged change in odds due to a change in average tweeting probability of the active users, when controlling for behavioural changes.

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results

Descriptive findings. To determine the language of a tweet, in accordance with the literature[55,56], we utilize the language field provided by the Twitter API. Ukrainian (35.8%) and Russian (35.4%) tweets are most prevalent in our dataset, followed by English (11.5%). A large proportion of tweets (11.1%) is labelled as "undefined", which among others consists of tweets that are too short, contain only hashtags, or only have media links. All other languages have shares of 1.2% or less. For the subsequent analysis we focus on tweets coming from the three main languages (English, Russian, Ukrainian) and discard all remaining tweets. A full breakdown of the language distribution is reported in Supplementary Figure 6.

In our dataset, there are clear trends in the aggregate over time (Fig. 1). In the beginning of 2020, we can see that Russian is the predominant language being used on Twitter in Ukraine, however, over time, this number gradually declines. The number of Ukrainian and English tweets on the other hand remains more or less constant over this initial time period. In the figure, we mark two key dates. On 11th November 2021, the United States officially report a mobilization of Russian troops along the Ukrainian border for the first time[57–59]. We will subsequently call this the first signs of aggression. 24th February 2022 marks the begin of the Russian invasion of Ukraine (subsequently referred to as outbreak of the war). As we approach this outbreak, there is a clear spike in tweets across all three languages, with a larger spike in both English and Ukrainian. Afterwards, English and Russian remain mostly constant, although the former on a much higher level than before. For Ukrainian, there is a clear upward trend in the daily number of tweets after the outbreak of the war.

Given these remarkable shifts in the number of tweets in the three considered languages, we want to investigate the underlying factors contributing to these changes. Note, that from the aggregate trends, we can not distinguish whether the observed

COMMUNICATIONS PSYCHOLOGY | https://doi.org/10.1038/s44271-023-00045-6 ARTICLE

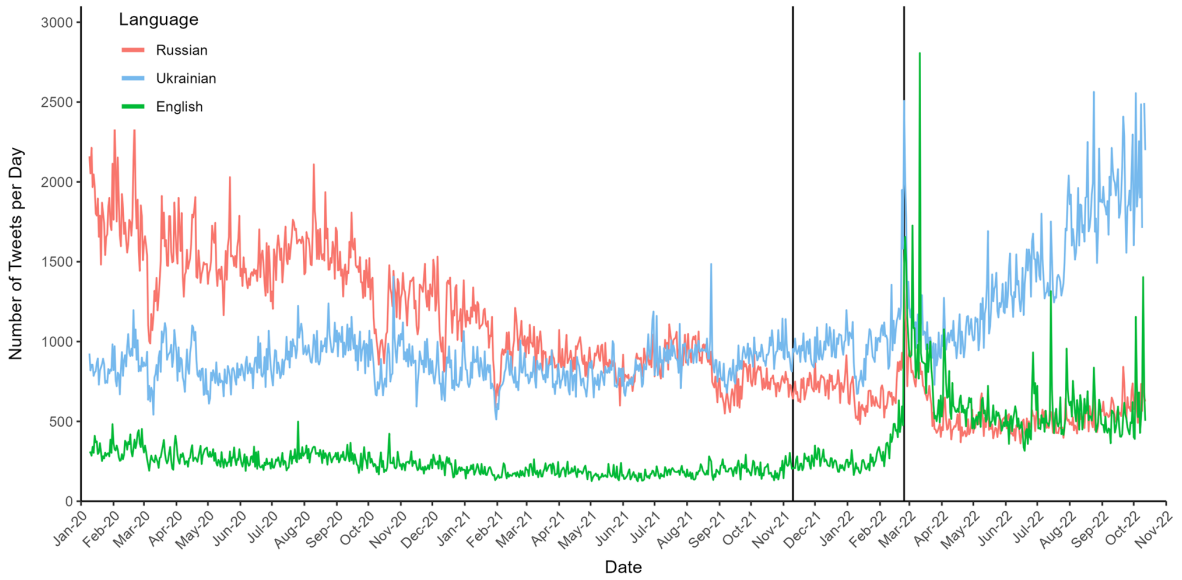


Fig. 1 Daily number of tweets in the three most common languages. Russian in red, Ukrainian in blue, English in green. From 9th January to 12th October (1008 days). The first vertical line denotes the mobilization of the Russian troops along the Ukrainian border (11th November 2021). The second line denotes the outbreak of the war (24th February 2022).

patterns are due to large in- and outfluxes of users, i.e. user turnover, which are common in online communities[40–42], or whether the actively tweeting users change their behaviour over time[43–45]. The disentanglement of this question is the aim of the rest of this article.

User activity. In order to address this question, we restructure our dataset by aggregating the number of tweets made by each user in English (EN), Ukrainian (UA), and Russian (RU) in each week. (Note, that we employ the Ukrainian country code "UA" instead of the official Ukrainian language tag "UK" in order to avoid confusion.) This allows us to study users' individual behaviour over time. To obtain reliable results, we restrict the further analysis to users who have tweeted in total at least ten times in any of the three languages. Furthermore, we choose weeks instead of days, as we are interested in general shifts and overall changes in behaviour over time, which are captured sufficiently well on a weekly basis. Through this weekly definition, we can dramatically reduce the size of our dataset, hence more complex modelling approaches become computationally feasible. We drop the first and last week in our dataset as these are incomplete (less than 7 days) and aggregate the remaining tweets on a weekly basis for each user and language. Finally within this, we are only considering weeks in which users are active (we define this as any week in which a user is tweeting at least once, as well as up to two weeks after), in order to account for the times in which users may be inactive for several weeks at a time or abandon their accounts. Thus, our new sample ranges from 13th January 2020 to 10th October 2022 and consists of 143 analysis weeks, 13,643 users and 1,045,245 observations.

Using this definition of user activity, we can visualize the total amount of active users as well as turnover rates (switch from active to inactive and vice versa) over time (Fig. 2). In the beginning of 2020, we have around 2800 active users per week. This number gradually decreases to roughly 1,800 until we approach the outbreak of the war. Afterwards, the number of active users starts increasing again. Note the drop and subsequent spike in activity shortly before and with the outbreak of the war. Looking at the turnover rates, we find that there is a constant stream of ~250 (potentially different) users per week that switch from active to inactive and vice versa. The aforementioned spikes are also evident in these turnover rates. Finally, we find that there are roughly 50 users per week that join our sample for the first time and about the same amount that leave it altogether. Both of these numbers almost double after the outbreak of the war.

Tweeting activity. To obtain a better understanding on how the average active Ukrainian Twitter user changes over time, we visualize the average number of published tweets by a user in each language in Fig. 3a. From the figure, we can clearly see that there are substantial shifts. Overall, the average number of RU tweets per user decreases constantly over time (from over 6.5 to 2.1), the outbreak of the war being no exception. The average number of EN tweets decreases slightly until the war, where we notice a sudden uptick (from 0.7 to 2.8), followed by a steady decline. Meanwhile, the number of UA tweets slowly but steadily rises (from 2.4 to 2.9), with steeper increases after the first signs of aggression in November 2021 and no appearance of slowing down (5.3 at the end).

By combining these findings with Fig. 2, we can at least partially explain the aggregate trends evident in Fig. 1. While the active user sample is shrinking over time, those users that stay (and join) the sample are tweeting more in UA. Hence, there is no decrease in the overall amount of UA tweets. We find the exact opposite for RU tweets. As the number of active users is declining, the users that stay active are tweeting less in RU, resulting in the visible decrease of aggregate RU tweets over time. Notably, so far, we do not know, if those changes in the average amount of tweets per user are simply driven by shifts in our active user sample (i.e., are those users that initially tweet a lot in RU leaving over time and this is why we see this decrease in the average?), or, if these changes are (at least partially) driven by behavioural changes in those users that remain active on Twitter (i.e., are the same users tweeting less in RU over time?).

We address this through our tweet model described in Section "Tweet Modelling". We fit a generalized additive mixed model

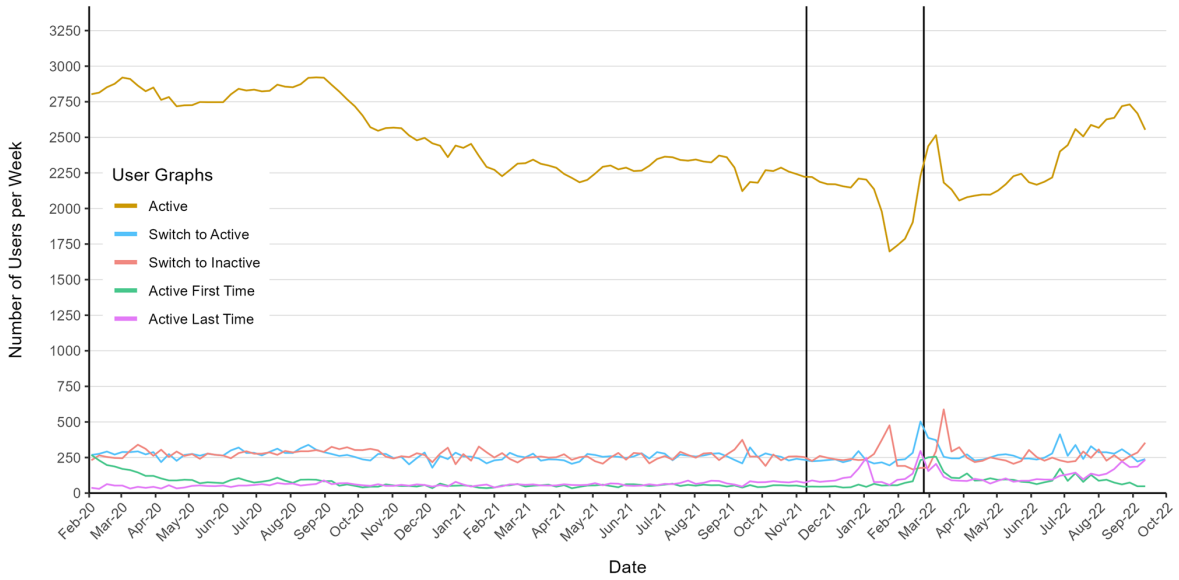


Fig. 2 Weekly user activity graphs. The brown graph reports the number of active users in each week. The blue (red) graph reports the number of users who switch to active (inactive), the green the number of users who switch to active for the first time, the purple the number of users who were active for the last time, i.e. drop out of the sample altogether. All graphs, but particularly the latter two, are skewed upwards respectively downwards towards beginning and end of the study period due to the nature of how the dataset is constructed. Hence, we drop the first and last three weeks for visualization purposes (137 total weeks left). The full plot is available in Supplementary Notes 5.

(GAMM) to predict the number of tweets made by each user in each language in each week, assuming a Poisson distribution. By incorporating both a smooth global time trend for each language, as well as user-specific random effects for each of the languages, we disentangle sample shifts (random effects) from behavioural changes (global trend). Hence, the former capture any changes in the population of active Ukraine-based Twitter users, while the latter strictly measures how these active users change their behaviour.

Figure 3b visualizes the average fitted sample (population) effects, i.e. the graphs depict how the average time-constant tweeting intensity in our active user sample changes over time due to user turnover. The figure shows, that the average RU tweeting intensity is mostly constant over time until November 2021, where aggression starts. From that point onward, in the span of only a few months, we see a decline of 21% in RU tweets from November 2021 to October 2022 (end of study period), solely attributed to changes in the user sample during that period. For EN, we find somewhat of an opposite effect. Similarly, there are only minor fluctuations until November 2021. But afterwards, there is a sharp increase of 107%. Taking a look at UA, we find a long-term increase of about 43% before the aggression starts. This increase comes to a hold shortly before the war, and considerably speeds up in the weeks after (+87%). All (relative) effect sizes calculated between the most relevant dates in our analysis period (start of study period, first signs of aggression, outbreak of war, end of study period) are reported in Table 1. We elaborate on this in Supplementary Notes 6, where we provide an additional figure, which illustrates sample changes over four-weekly intervals (Supplementary Fig. 9). From there, we can observe that the largest shifts clearly take place with and after the outbreak of the war. We also provide an alternative to Table 1, which measures the speed of change between the key dates in Supplementary Table 5. A full breakdown of all model coefficients is available in Supplementary Table 7.

Next, we will investigate behavioural changes using Fig. 3c. The graphs depict how the tweeting behaviour of the active users changes throughout the study period, when controlling for the user turnover (sample effects). Starting with RU, we notice that users are tweeting less and less over time. From January 2020 to November 2021, users tweet 49% less in RU due to behavioural changes. Subsequently, we see a small rise with the outbreak of the war (+5%), followed up by an even steeper decline (−24%). In contrast, UA is reasonably consistent in its use up until the start of aggression. From there, we observe a surge (+36%) until the outbreak of the war, followed by a gentler increase (+15%) after. Finally, looking more closely at EN tweeting behaviour, we can observe a general downward trend (−34%) until November 2021. Once the aggression starts, there is a huge spike (+130%), as users are tweeting a lot more in EN. After the outbreak of the war, this somewhat reverses (−40%), however, without dropping back down to pre-aggression levels. A full breakdown of all changes is reported in Table 1. Again, we elaborate in in Supplementary Notes 6. Supplementary Figure 9 shows that the largest behavioural shifts take place shortly before, with, and after the outbreak of the war. As a robustness check, we also pursue two alternative modelling strategies, one using factor smooths instead of random intercepts, the other implementing a regression discontinuity design, which are discussed extensively in Supplementary Notes 2.1 and 2.2 respectively. Both confirm the behavioural patterns described here.

Overall, we can conclude that there are only minor sample shifts pre-dating aggression that affect tweeting activity, but major shifts thereafter. In terms of behaviour, we can already observe steady changes early on, which considerably intensify with the war. However, as of yet, we cannot exactly pinpoint where those changes come from. Are users that already tweet in UA simply tweeting more with the outbreak of the war, or is it possible that users are actively switching the language they are tweeting in?

COMMUNICATIONS PSYCHOLOGY | https://doi.org/10.1038/s44271-023-00045-6 ARTICLE

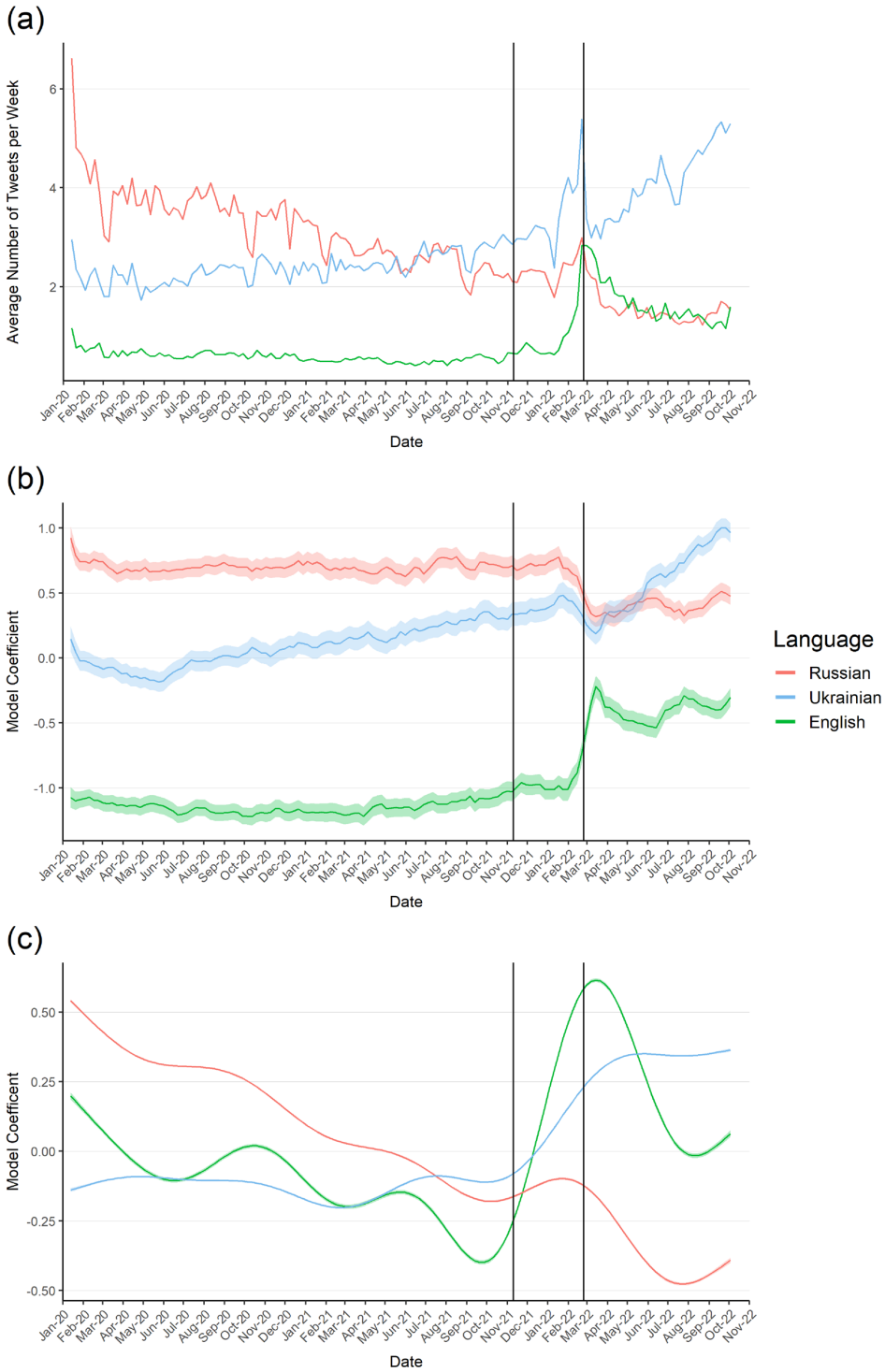


Fig. 3 Changes in the number of tweets per user. Russian in red, Ukrainian in blue, English in green. (**a**) visualizes the descriptive average number of tweets over time, (**b**) the sample effects (average random effects), (**c**) the behavioural effects (global trend). The shaded area depicts the 95% confidence interval. The first vertical line denotes the mobilization of the Russian troops along the Ukrainian border. The second line denotes the outbreak of the war.

Table 1 Tweet activity effect sizes between key dates

Language	Sample effects			
	Start—Aggression	Aggression—War	War—End Study	Aggression—End Study
English	+6.16%	+34.87%	+53.14%	+106.54%
Ukrainian	+43.12%	−0.44%	+87.70%	+86.87%
Russian	−2.91%	−17.41%	−4.12%	−20.82%
	Behavioural Effects			
English	−34.41%	+130.11%	−39.98%	+38.09%
Ukrainian	+4.67%	+35.72%	+15.184%	+56.32%
Russian	−48.90%	+4.68%	−23.86%	−20.30%

Effect sizes for both sample and behavioural changes extracted from the tweet model described in Section "Tweet Modelling" between key dates. All effect sizes are relative increases in the number of tweets between the two respective dates. For the start date calculation, we drop the first two weeks of the study period. Start: start of the study period—27th January 2020. Aggression: first official US report of a mobilization of the Russian troops along the Ukrainian border—11th November 2021. War: outbreak of the war—24th February 2022. End Study: end of the study period—10th October 2022.

Choice of language. We analyze the choice of language more closely in the following. As we are interested in shifts between the individual languages, we look at the pairwise probability to tweet in one language over another over time. Hence, the probability reports how likely it is that a user tweets in language one (e.g. UA) over language two (e.g. EN). With three languages, this pairwise evaluation gives us a total of three different language pairs (UA over RU, UA over EN, RU over EN), where the order in which we specify each pair is irrelevant. Figure 4a visualizes how these pairwise probabilities evolved for an average user over time. For RU over EN the probability is mostly constant (82% to tweet in RU) until aggression starts, from where it continuously drops down to 58%. For UA over EN we see small increases over time (68–72%). With the mobilization of the Russian troops, we see a drop (62%), followed by a rise back to pre-aggression levels months into the war. Finally, for UA over RU we see a completely different pattern. Initially, the probability to tweet in UA is low (32%), from where it continues to rise consistently. In the weeks leading up to the war, there is a considerable speed up in this shift, resulting in a probability of 76% to tweet in UA over RU towards the end of the analysis period in October 2022.

Similarly to before, we can disentangle sample shifts from behavioural changes through statistical modelling. In summary, we fit a GAMM to model users' pairwise language probability to tweet over time, assuming a binomial distribution. As before, we include a smooth global time trend and user-specific random effects into the model. We fit such a model, for all three aforementioned language-pairs. A full description is provided in Section "Language Modelling".

Figure 4b visualizes the fitted average sample effects across all three models, i.e. the graphs depict how the average time-constant tweeting probabilities in the active user sample change over time. As we are working with coefficients of a logistic regression, changes must be interpreted with respect to changes in the odds. The figure shows that for RU over EN, initially, there is only a minor decline (−19%). However, as we approach the outbreak of the war, we can report a large drop in the odds, as users are 62% less likely to tweet in RU over EN than before, with further decreases thereafter (−29%). For UA over EN, we find a small to moderate increase until aggression (+21%) due to sample shifts, followed by a large drop until war outbreak (−52%), which is recovered in the months after (+42%). Finally, for UA over RU, there is a constant increase in the odds over time (+66%), which speeds up once aggression starts (+87% until October 2022). Table 2 details all changes. As before, changes over four-weekly intervals are visualized in Supplementary Fig. 10. The figure shows that the sample effects for the language choice are slightly more erratic, but the major shifts take place with and after war outbreak. The alternative to Table 2, with the speed of changes is available in Supplementary Table 6, the full breakdown of all model coefficients in Supplementary Table 8.

Combining this with the results from the previous section, we can conclude that the user turnover in the first 1.5 years shifts the sample such that users are more likely to tweet in UA (than RU or EN), but not at the expense of either of the two other languages, as the sample effects for tweeting activity are (mostly) steady for both. As we approach the outbreak of the war, this drastically changes. Then, the user sample clearly shifts away from RU, as users are instead tweeting more in EN (initially) and UA (long-term). Upon further investigation (Supplementary Notes 7 and 8), we find that users tweeting in RU start leaving around November 2021 (start of aggression), with EN users joining. The former continue to leave as the war unfolds, with a few of the latter also leaving the sample again over time. This is also reflected in the increase of the UA odds over time (UA over RU consistently, UA over EN as war continues).

Figure 4c reports behavioural language changes across all three language pairs, when controlling for the user turnover. For RU over EN we see a constant decline in the odds over time (−38% to tweet in RU), which further speeds up once aggression starts (−51%). For UA over EN we see the exact opposite, as over time users are more likely to tweet in UA (+64% in odds). This change reverses with the start of aggression and the outbreak of the war (−34%), but subsequently reaches pre-aggression levels as the war unfolds. Finally, we can see a clear shift from UA to RU even early on (+129%). This switch becomes even more striking with the outbreak of the war, as users are actively changing their behaviour such that the average user is 249% more likely to tweet in UA over RU in the span of a single year. Table 2 reports all relevant changes. Supplementary Figure 10 similarly illustrates that the biggest behavioural shifts take place around the outbreak of the war, but also that there already is a constant long-term shift from UA towards RU before. Our alternative modelling strategies in Supplementary Notes 2.1 and 2.2 confirm these findings.

Connecting these language shifts with the results on tweeting activity, we find that the initial decline in EN and RU tweeting activity is not limited to monolingual users. Instead, users are actively shifting towards UA, by reducing their amount of RU and EN tweets (with a stronger shift from RU than EN respectively). Similarly, the temporary increase in EN tweeting behaviour leading up to the war can be linked to both UA and RU users. Finally and most importantly, the decline of RU and the rise of UA tweeting behaviour that manifests itself with the war is strongly driven by a major language shift (2.5 times increase) from RU to UA.

We visualize and demonstrate this substantial behavioural language shift from RU to UA in Figs. 5, 6. Figure 5 plots the language proportion of each user (UA to RU; from 0 to 1) that tweet in either language before (y-axis) and after the war (x-axis). Hence, along the straight black line through the origin we have users that do not switch language (top right UA, bottom left RU), users above the line switch to RU, below the line to UA, with users switching completely from one language to the other being located in either the top left (all tweets in UA to all in RU) or bottom right corner. Statistically significant ($p < 0.05$, $z > 1.96$) language shifts from before to after war outbreak are determined using a two-sided z-test with unequal variances on each user's language proportion, and are marked in the plot (the distributions were assumed to be normal but this was not formally tested). From the figure it becomes evident that there are many users that do not switch language (in both UA and RU), as well as

COMMUNICATIONS PSYCHOLOGY | https://doi.org/10.1038/s44271-023-00045-6 ARTICLE

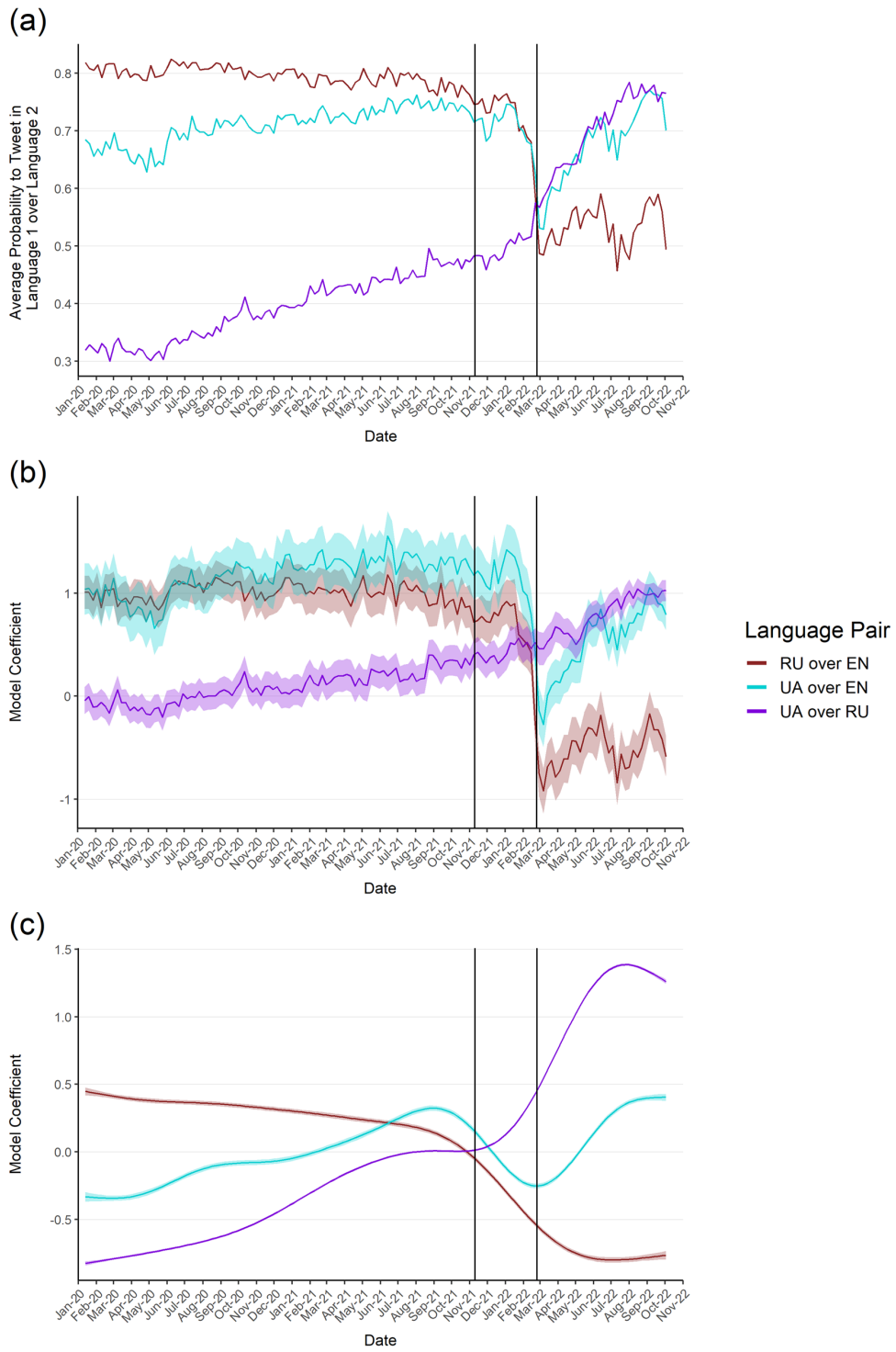


Fig. 4 Changes in the choice of language per user (RU over EN in brown, UA over EN in turquoise, UA over RU in purple). RU over EN in brown, UA over EN in turquoise, UA over RU in purple. (**a**) Visualizes the descriptive average probability to tweet in one language over another, (**b**) the sample effects (average random effects), (**c**) the behavioural effects (global trend). The shaded area depicts the 95% confidence interval. The first vertical line denotes the mobilization of the Russian troops along the Ukrainian border. The second line denotes the outbreak of the war.

many users clearly switching from RU to UA at various levels, whereas there are only very few switching from UA to RU.

In this sample of users who tweet in either RU or UA both before and after the outbreak of the war (3237 users), we have 1363 users who predominately tweet in RU (>80% of tweets) before the war. Of those, 839 (61.6%) tweet more in UA after the war, with 566 (41.5%) reporting a significant behavioural change ($z > 1.96$, $p < 0.05$). Out of those 850 users, 341 (25%) even switch to predominately tweeting in UA (>80% of tweets), i.e. perform a hard-switch, with 296 (21.7%) statistically significant hard-switches ($z > 1.96$, $p < 0.05$). We pick those 296 users and plot their weekly language proportion over time in Fig. 6. Red points denote 100% of the tweets being phrased in RU, blue points denote the same in UA. From the figure, we can clearly see a substantial break and change in behaviour around the time the war breaks out (second black line), as most of the users switch from RU to UA around this mark.

On Ukrainian side, we have 1172 users who predominately tweet in UA (>80% of tweets) before the war. Of those, 471 (40.2%) tweet more in RU after the war, with only 83 (7.1%) reporting a significant behavioural change ($z > 1.96$, $p < 0.05$). More importantly, we only observe 35 (3%) hard-switches, out of

Table 2 Language choice effect sizes between key dates

Language	Sample Eeffects			
	Start—Aggression	**Aggression—War**	**War—End study**	**Aggression—End Study**
UA over RU	+66.13%	+13.00%	+65.72%	+87.25%
UA over EN	+21.43%	−52.08%	+41.96%	−31.98%
RU over EN	−19.01%	−61.74%	−29.33%	−72.96%
	Behavioural effects			
UA over RU	+128.69%	+52.08%	+129.24%	+248.63%
UA over EN	+64.14%	−33.61%	+92.663%	+27.90%
RU over EN	−38.23%	−38.69%	−20.659%	−51.36%

Effect sizes for both sample and behavioural changes extracted from the language model described in Section "Language Modelling" between key dates. All effect sizes are relative increases in the odds between the two respective dates. For the start date calculation, we drop the first two weeks of the study period. Start: start of the study period—27th January 2020. Aggression: first official US report of a mobilization of the Russian troops along the Ukrainian border—11th November 2021. War: outbreak of the war—24th February 2022. End Study: end of the study period—10th October 2022.

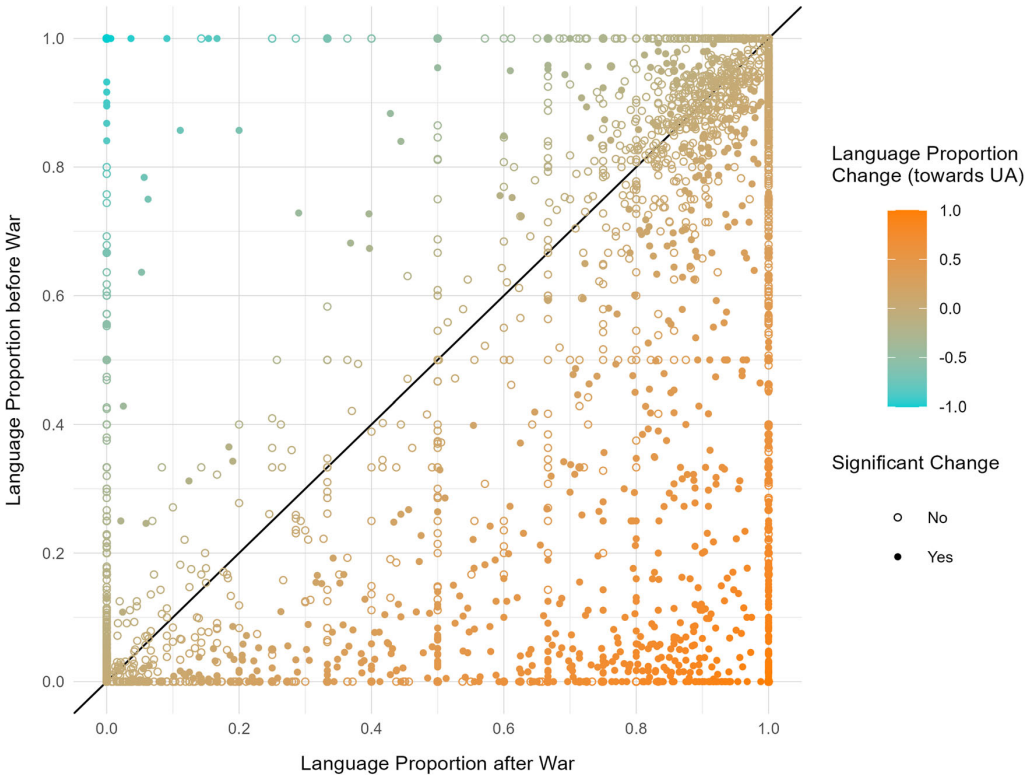


Fig. 5 Scatterplot of users' language proportions before and after the outbreak of the war. We are only considering users who tweet in either RU or UA (or both) before and after ($n = 3237$). The points are coloured with respect to each user's shift in language. 1 (orange) denotes a complete shift to UA, −1 (green) a complete shift to RU, 0 no shift. The straight line through the origin covers all points without a shift. Significant shifts ($z > 1.96$, $p < 0.05$) using a two-sided z-test with unequal variances on each user's language proportion are denoted through full (non-empty) points. $n = 1808$ (821 significant) shifts towards Ukrainian, $n = 818$ (106 significant) shifts towards Russian. Only RU and UA tweets of each user are considered.

COMMUNICATIONS PSYCHOLOGY | https://doi.org/10.1038/s44271-023-00045-6 ARTICLE

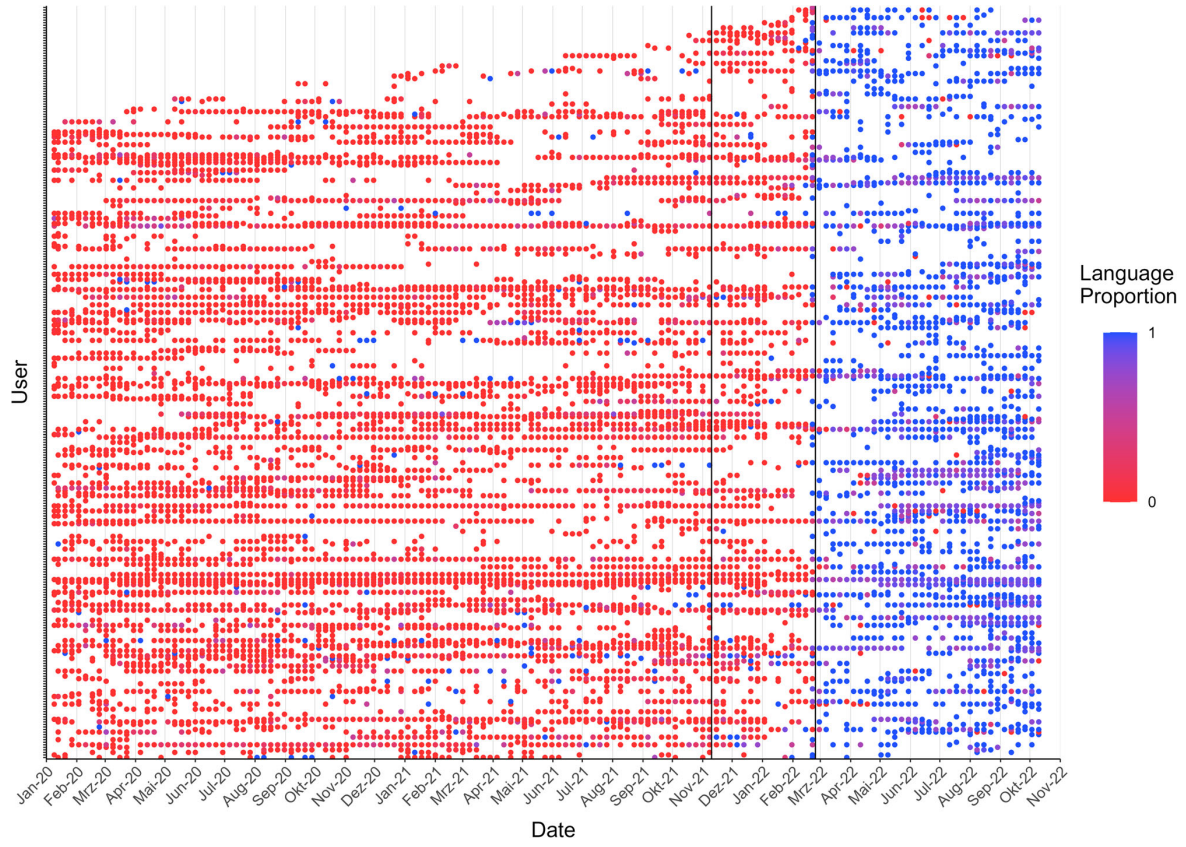


Fig. 6 Scatterplot of users' language proportion in each week over time. Each row (on the y-axis) denotes one of the $n = 295$ users with a statistically significant hard-switch from RU to UA. The points are coloured with respect to each user's language proportion in the respective week (145 total weeks). Blue = 100% Ukrainian, red = 0% Ukrainian (=100% Russian). Missing points indicate that a user was not tweeting in the respective week. Only RU and UA tweets of each user are considered. The first vertical line denotes the mobilization of the Russian troops along the Ukrainian border. The second line denotes the outbreak of the war.

Table 3 Median % differences in user characteristics

Characteristic	No switch	Switch	Difference	*P*-value	$\chi^2(1)$
Followers	77	119	+**54.54%**	0.004	8.223
Followings	116	132	+13.8%	0.155	2.023
Account age (month)	94.15	105.66	+12.22%	0.073	3.196
Tweet frequency	0.79	1.16	+**47.73%**	0.021	5.352
Likes frequency	0.84	1.25	+**48.93%**	0.021	5.352
# of tweets in Ukraine	57	85	+**49.12%**	0.001	10.639
War topic 1	4.0	6.5	+**62.5%**	<0.001	22.061
War topic 1 (rel.)	0.061	0.063	+4.71%	0.801	0.063
War topic 2	1	2	+**100%**	0.007	7.312
War topic 2 (rel.)	0.013	0.015	+17.6%	0.461	0.543

$n = 1067$ users in the no switch group, $n = 296$ users in the switch group. Column 2 reports the median of the respective user characteristic for those Russian users that do not perform a statistically significant hard-switch to Ukrainian with the outbreak of the war, column 3 for the users that do. Significant ($p < 0.05$) differences using a two-sided chi-squared are marked in bold. A description of all user attributes is provided in Section "User Characteristics".

which 20 (1.7%) are significant ($z > 1.96$, $p < 0.05$). Hence, there are only very few UA tweeting users for which we can report a significant switch towards RU after the war.

Finally, we analyze potential differences in those RU users that perform a hard-switch to UA from those that do not (see Table 3). We find that there are significant differences ($p < 0.05$) in the median in various user characteristics between the two groups using a two-sided chi-squared test (no distributional assumptions required). Users switching have more followers (+54.5%, $\chi^2(1) = 8.223$, $p = 0.004$), a higher tweet frequency (+47.7%, $\chi^2(1) = 5.352$, $p = 0.021$) as well as a higher like frequency (+48.9%, $\chi^2(1) = 5.352$, $p = 0.021$) and published more Ukraine geo-tagged tweets during the study period (+49.1%, $\chi^2(1) = 10.639$, $p = 0.001$), whereas there are only small non-significant differences in account age (+12.2%, $\chi^2(1) = 3.196$, $p = 0.07$) and followings (+13.8%, $\chi^2(1) = 2.023$, $p = 0.16$).

We also conduct a multilingual topic modelling on the tweets using BERTopic[49]. The method and its results are thoroughly described in Supplementary Notes 1. We find two topic clusters referencing the war (topic #1 and topic #3). The former is mostly related to updates regarding the situation, asking for help, and supporting the people of Ukraine, the latter covers a more political side of the overall conflict. Both topics are in total more discussed by those RU users that switch language (+62.5%, $\chi^2(1) = 22.061$, $p < 0.001$; +100%, $\chi^2(1) = 7.312$, $p = 0.007$). However, once we control for their total amount of tweets in our dataset, i.e. we compute a relative share of war related tweets for each user, the differences shrink and turn non-significant (+4.71%, $\chi^2(1) = 0.063$, $p = 0.801$; +17.6%, $\chi^2(1) = 0.543$, $p = 0.461$).

Discussion

In our work, we collected geo-tagged tweets from Ukraine and analyzed tweeting activity and language choice before and during the Russian war in Ukraine from 9th January 2020 to 12th October 2022. Due to the nature of our longitudinal dataset and our methodological approach using a generalized additive mixed model (GAMM), we were able to disentangle and quantify shifts in the user sample, arising from user turnover, from behavioural changes of the actively tweeting users. Our GAMMs were able to handle the large sample size and take care of user's varying periods of inactivity within the study period, while at the same time allowing for a flexible non-linear but interpretable model fit.

Our analysis shows a steady long-term shift away from Russian towards Ukrainian already before the war, as the Ukrainian tweet probability rises substantially (vs. Russian; 33% to 48%). This shift is majorly driven by behavioural changes. The actively tweeting users reduce their number of Russian tweets in favour of Ukrainian over time. This is likely a conscious choice and thus shift in how the users communicate and present themselves to their online audience[26–29,31]. This finding is also in line with trends observed over a 20-year period between the 1989 and the last conducted census in 2001[36] and more recently across surveys[37], where the share of people reporting Ukrainian as their native language perpetually rose over time. Notably, with the Euromaidan protests and the subsequent Russian military intervention in 2014, this shift seems to have sped up, as citizens ethnonational identification and everyday language use is substantially shifting towards Ukrainian. This recent shift towards Ukrainian has also been identified in a small qualitative study on Facebook posts[60]. We can confirm these findings quantitatively both at-scale and in an ecologically valid setting.

We find this gradual shift to drastically speed up with the start of Russian aggression in November 2021 and the subsequent outbreak of the war. In the span of a few months, Ukrainian tweet probability rises from 48% to a remarkable 76%. While some of this increase can be explained by Russian tweeting users leaving and Ukrainian users joining (+87% in odds to tweet in Ukrainian), the major factor is a behavioural change (+249% in odds to tweet in Ukrainian), with a rise in Ukrainian (+56%) and a decrease in Russian tweeting activity (−20%). Notably, we show that out of those users predominately tweeting in Russian before the war, roughly half of them tweet more in Ukrainian after. Strikingly, around a quarter of them switch to predominately tweeting in Ukrainian, i.e., they are performing a hard-switch. It is worth noting, that we do not observe more than a handful of switches in the other direction. This shift from Russian to Ukrainian is in line with news reports and small-scale surveys outlining the war as the cause for the recent changes in language use across Ukraine[38,39]. We theorize that this is a highly politicized response. Users want to distance themselves from any support of the war by no longer using Russian, and consciously change their self-expressed (online) identity[26–29,31], as also already to some extent reported after the Russian military intervention in 2014 both on- and offline[37,60] and confirmed in our study through the gradual shift before the war. However, with the Russian invasion, this shift seems to have sped up massively. Moreover, the distancing from supporting the war may also explain why Russian users that perform a hard-switch to Ukrainian seem to be more active on Twitter (including discussions on the war) and have a larger follower base (median of 119 vs. 77). Pressure and general interactions on social media were already reported among the main reasons for the language switch after 2014[60]. Note, that this might also (partially) explain the sample of active users shifting from Russian towards Ukrainian (sample effects).

In addition, we observe a long-term behavioural shift away from English tweeting activity up until November 2021. This could be interpreted as a reduction in talking to a broader international audience during that time[61–63], due to the fact that English is the most widely understood language on the internet by far[31,64]. However, not surprisingly, with the mobilization of the Russian troops along the Ukrainian border and specifically in the weeks leading up to the war, with a spike during outbreak, we observe a substantial shift towards English. We hypothesize users wanted to let the world know what was happening and called for aid[31], which is supported by the fact that we observe a heavy spike in English tweets assigned to the first war topic (more related to help, support and updates). While we record a large influx of English speaking users during that time (+35% in number of tweets), we can also see a substantial behavioural shift (+130%). Already active users tweet substantially more in English, independent of the language they were normally tweeting in. As the war continues to unfold, this somewhat reverses, with some of the newly joined English users leaving and behaviour reverting, although not to pre-aggression levels. With the world being more aware of the situation, and the international community supporting Ukraine in various ways[65,66], users likely have less reasons to continue tweeting in English. Instead, they return back to intra-national discussions and thus their native language(s).

Limitations. We recognize that while our study provides a strong foundation towards a better understanding on how the Ukrainian population reacted to the Russian invasion both on- and offline, possible limitations need to be acknowledged. The sample of users investigated here is not representative of the entire Ukrainian population. Indeed, it is skewed towards the younger and middle-aged part of the population (aged 18-49, see also Section "User Characteristics"). Additionally, we want to emphasize that geo-information is not included on most Twitter clients by default, which might further skew the sample. As on most other social media platforms, users have the option to create new accounts, which we cannot match to their prior ones. Hence, some of the behavioural effects might even be underestimated and instead accounted for as sample effects. Moreover, users might stop tweeting (with Ukrainian geo-information) for various reasons (e.g. because they fled the country). One should keep in mind that the behavioural language shifts taking place with the outbreak of the war are only demonstrated for those users who continue to tweet at and/or after the outbreak, which could potentially lead to a selection bias. Future work may analyze the content and sentiment of the tweets more closely. This could be augmented through the use of media objects attached to the tweets such as images and videos. An investigation of retweet and follower networks may reveal additional differences between those users that are shifting language to those that are not.

Naturally, any analysis can be repeated and extended to other social media platforms.

Conclusion

In summary, our work investigated tweeting activity and language choice on Ukrainian Twitter before and during the Russian war in Ukraine through a large-scale longitudinal study. We demonstrate substantial shifts away from the Russian language to Ukrainian, which we interpret as users' conscious choice towards a more Ukrainian (online) identity. More than half of the predominately Russian-tweeting users shift towards Ukrainian, and a quarter of them even perform a hard-switch to Ukrainian, as the war breaks out. This can be seen as citizens' increasing opposition to Russia and a return to the country's linguistic roots as well as a push towards a conscious self-definition of being Ukrainian.

Data availability

Data are available at the Open Science Framework (OSF) using https://osf.io/48sbc or with the https://doi.org/10.17605/OSF.IO/48SBC. As per Twitter developer agreement 18th April 2023, we are legally not allowed to share tweets beyond their IDs. Hence, we share our data in two ways. First, by sharing all tweet IDs needed to construct our aggregated datasets. Second, by sharing our aggregated datasets that are used for all our analyses. All of these are provided in the OSF repository with a corresponding documentation.

Code availability

All of our code (including the aggregation scripts) is also available in the OSF repository at https://osf.io/48sbc or with the https://doi.org/10.17605/OSF.IO/48SBC. Descriptions for each script are provided there. We conducted our main analyses using R 4.1.3. In the OSF repository, we provide a session info file, which lists the version of every R package employed to conduct our analyses. We conducted the topic modelling using Python 3.10 and BERTopic 0.15.0.

Received: 5 June 2023; Accepted: 28 November 2023;
Published online: 10 January 2024

References

1. Saroj, A. & Pal, S. Use of social media in crisis management: A survey. *Int. J. Disaster Risk Reduction* **48**, 101584 (2020).
2. Dwivedi, Y. K., Ismagilova, E., Rana, N. P. & Raman, R. Social media adoption, usage and impact in business-to-business (b2b) context: A state-of-the-art literature review. *Inf. Syst. Front.* 1–23 (2021).
3. Wong, A., Ho, S., Olusanya, O., Antonini, M. V. & Lyness, D. The use of social media and online communications in times of pandemic covid-19. *J. Intensive Care Soc.* **22**, 255–260 (2021).
4. Mäkinen, M. & Wangu Kuira, M. Social media and postelection crisis in Kenya. *Int. J. Press/politics* **13**, 328–335 (2008).
5. Sadri, A. M., Hasan, S., Ukkusuri, S. V. & Cebrian, M. Crisis communication patterns in social media during hurricane sandy. *Transp. Res. Record* **2672**, 125–137 (2018).
6. Morozov, E.*The net delusion: The dark side of Internet freedom* (PublicAffairs, 2012).
7. Zhuravskaya, E., Petrova, M. & Enikolopov, R. Political effects of the internet and social media. *Ann. Review Econ.* **12**, 415–438 (2020).
8. Flamino, J. et al. Political polarization of news media and influencers on twitter in the 2016 and 2020 US presidential elections. *Nat. Human Behav.* **7**, 904–916 (2023).
9. Sacco, V. & Bossio, D. Using social media in the news reportage of war & conflict: Opportunities and challenges. *J. Media Innov.* **2**, 59–76 (2015).
10. Rogstadius, J. et al. Crisistracker: Crowdsourced social media curation for disaster awareness. *IBM J. Res. Dev.* **57**, 4–1 (2013).
11. Allcott, H. & Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**, 211–236 (2017).
12. Kaufhold, M.-A., Rupp, N., Reuter, C. & Habdank, M. Mitigating information overload in social media during conflicts and crises: design and evaluation of a cross-platform alerting system. *Behav. Inf. Technol.* **39**, 319–342 (2020).
13. Marples, D. R.*The War in Ukraine's Donbas: Origins, Contexts, and the Future* (Central European University Press, 2021).
14. Bigg, M. M. Russia invaded ukraine more than 200 days ago. here is one key development from every month of the war. https://www.nytimes.com/article/ukraine-russia-war-timeline.html (2022). Retrieved 2023-01-14.
15. OHCHR. Ukraine: civilian casualty update 24 April 2023. https://www.ohchr.org/en/news/2023/04/ukraine-civilian-casualty-update-24-april-2023 (2023). Retrieved 2023-04-26.
16. Lamb, W. Rebuilding Ukraine will cost at least $349 billion, a new report estimates.*The New York Times*https://www.nytimes.com/live/2022/09/10/world/ukraine-russia-war#rebuilding-ukraine-349-billion-dollars (2022). Retrieved 2023-04-14.
17. World Bank. Ukraine rapid damage and needs assessment: February 2022 - february 2023 (english). *Washington, D.C. : World Bank Group*. http://documents.worldbank.org/curated/en/099184503212328877/P1801740d1177f03c0ab180057556615497 (2023).
18. UNHCR. Ukraine refugee situation https://data.unhcr.org/en/situations/ukraine (2023). Retrieved 2023-04-14.
19. Ratten, V. The ukraine/russia conflict: Geopolitical and international business strategies. *Thunderbird Int. Bus. Rev.* **65**, 265–271 (2023).
20. Reuter, C., Hughes, A. L. & Kaufhold, M.-A. Social media in crisis management: An evaluation and analysis of crisis informatics research. *Int. J. Human–Comput. Interact.* **34**, 280–294 (2018).
21. Jurgens, M. & Helsloot, I. The effect of social media on the dynamics of (self) resilience during disasters: A literature review. *J. Contingencies Crisis Manag.* **26**, 79–88 (2018).
22. Dwarakanath, L., Kamsin, A., Rasheed, R. A., Anandhan, A. & Shuib, L. Automated machine learning approaches for emergency response and coordination via social media in the aftermath of a disaster: A review. *IEEE Access* **9**, 68917–68931 (2021).
23. Dowd, C., Justino, P., Kishi, R. & Marchais, G. Comparing 'new'and 'old'media for violence monitoring and crisis response: evidence from kenya. *Res. Politics* **7** (2020).
24. Steinert-Threlkeld, Z. C., Chan, A. M. & Joo, J. How state and protester violence affect protest dynamics. *J. Polit.* **84**, 798–813 (2022).
25. Kulyk, V. The age factor in language practices and attitudes: continuity and change in Ukraine's bilingualism. *Nationalities Pap.* **43**, 283–301 (2015).
26. Lee, C. Language choice and self-presentation in social media: the case of university students in hong kong. In *The language of social media: Identity and community on the Internet*, 91–111 (Springer, 2014).
27. Fiske, S. T. *Social beings: Core motives in social psychology* (John Wiley & Sons, 2018).
28. Herrmann, T., Jahnke, I. & Loser, K.-U. The role concept as a basis for designing community systems. In *Coop*, 163–178 (2004).
29. Hogg, M. A., Terry, D. J. & White, K. M. A tale of two theories: A critical comparison of identity theory with social identity theory. *Social psychology quarterly* 255–269 (1995).
30. Davidson, B. I. & Joinson, A. N. Shape shifting across social media. *Soc. Media + Soc.* **7**, 2056305121990632 (2021).
31. Lee, C. Multilingual resources and practices in digital communication. In *The Routledge handbook of language and digital communication*, 118–132 (Routledge, 2015).
32. Smagulova, J. Kazakhstan: Language, identity, and conflict. *Innovation: Eur. J. Soc. Sci. Res.* **19**, 303–320 (2006).
33. Wright, S.*Language policy, the nation and nationalism*, 59-78. Cambridge Handbooks in Language and Linguistics (Cambridge University Press, 2012).
34. Pavlenko, A. Multilingualism in post-soviet countries: Language revival, language removal, and sociolinguistic theory. *Int. J. Bilingual Educ. Bilingualism* **11**, 275–314 (2008).
35. Marshall, C. A. Post-soviet language policy and the language utilization patterns of kyivan youth. *Lang. Policy* **1**, 237–260 (2002).
36. Stebelsky, I. Ethnic self-identification in ukraine, 1989–2001: why more ukrainians and fewer russians? *Can. Slavonic Pap.* **51**, 77–100 (2009).
37. Kulyk, V. Shedding russianness, recasting ukrainianness: The post-euromaidan dynamics of ethnonational identifications in ukraine. *Post-Soviet Affairs* **34**, 119–138 (2018).
38. Harding, L. 'a generational shift': war prompts ukrainians to embrace their language. *The Guardian*https://www.theguardian.com/world/2023/mar/06/russia-ukrainians-embrace-language-war (2023). Retrieved 2023-03-29.
39. Warner, A. War in ukraine spurs decline in russian-language use, survey shows. *Multilingual* https://multilingual.com/war-in-ukraine-spurs-decline-in-russian-language-use-survey-shows/ (2022). Retrieved 2023-03-29.
40. Dabbish, L., Farzan, R., Kraut, R. & Postmes, T. Fresh faces in the crowd: turnover, identity, and commitment in online groups. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 245–248 (2012).
41. Panek, E., Hollenbach, C., Yang, J. & Rhodes, T. The effects of group size and time on the formation of online communities: Evidence from reddit. *Soc. Media+ Soc.* **4**, 2056305118815908 (2018).

42. Ransbotham, S. & Kane, G. C. Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in wikipedia. *Mis Quarterly* 613–627 (2011).
43. Davidson, B. I., Jones, S. L., Joinson, A. N. & Hinds, J. The evolution of online ideological communities. *PloS One* **14**, e0216932 (2019).
44. Eichstaedt, J. C. & Weidman, A. C. Tracking fluctuations in psychological states using social media language: A case study of weekly emotion. *Eur. J. Personality* **34**, 845–858 (2020).
45. Dzogang, F., Lansdall-Welfare, T. & Cristianini, N. Seasonal fluctuations in collective mood revealed by wikipedia searches and twitter posts. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 931–937 (IEEE, 2016).
46. Hu, Y. & Wang, R.-Q. Understanding the removal of precise geotagging in tweets. *Nat. Human Behav.* **4**, 1219–1221 (2020).
47. Yang, K.-C., Varol, O., Hui, P.-M. & Menczer, F. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 1096–1103 (2020).
48. Statista. Most popular social media by age Ukraine 2021 https://www.statista.com/statistics/1256255/most-popular-social-media-by-age-ukraine/ (2022). Retrieved 2023-03-28.
49. Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794* (2022).
50. Pfeffer, J., Mooseder, A., Hammer, L., Stritzel, O. & Garcia, D. This sample seems to be good enough! assessing coverage and temporal reliability of twitter's academic api. *arXiv preprint arXiv:2204.02290* (2022).
51. Lee, J. H. & Nguyen, A. T. How music fans shape commercial music services: A case study of bts and army. In *ISMIR*, 837–845 (2020).
52. Faraway, J. J.*Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models* (CRC press, 2016).
53. Hastie, T. J. Generalized additive models. In *Statistical models in S*, 249–307 (Routledge, 2017).
54. Wood, S. N.*Generalized additive models: an introduction with R* (CRC press, 2017).
55. Mosleh, M., Pennycook, G., Arechar, A. A. & Rand, D. G. Cognitive reflection correlates with behavior on twitter. *Nat. Commun.* **12**, 921 (2021).
56. Barbieri, F., Anke, L. E. & Camacho-Collados, J. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 258–266 (2022).
57. Stewart, P. & Ali, I. Pentagon says it continues to see unusual Russian military activity near Ukraine border. *Reuters* https://www.reuters.com/world/europe/pentagon-says-it-continues-see-unusual-russian-military-activity-near-ukraine-2021-11-15/ (2021).
58. Euronews. US alleges 'unusual' Russian troop movements near Ukrainian border. *Euronews* https://www.euronews.com/2021/11/11/us-alleges-unusual-russian-troop-movements-near-ukrainian-border (2021). Retrieved 2023-03-30.
59. NDTV. Soldiers, Separatists, Sanctions: A Timeline Of The Russia-Ukraine Crisis. *NDTV* https://www.ndtv.com/world-news/soldiers-separatists-sanctions-a-timeline-of-the-russia-ukraine-crisis-2782377 (2022). Retrieved 2023-03-30.
60. Kulyk, V. et al. Between the "self" and the "other": Representations of ukraine's russian-speakers in social media discourse. *East/West: J. Ukrainian Stud. (EWJUS)* **5**, 65–88 (2018).
61. Smith, L. E. English as an international language: No room for linguistic chauvinism. *J. Engl. as a Lingua Franca* **4**, 165–171 (2015).
62. Christiansen, T. W. The rise of english as the global lingua franca. is the world heading towards greater monolingualism or new forms of plurilingualism? *Lingue e Linguaggi* 129–154 (2015).
63. Moreno-Fernández, F. & Mella, H. Á. Reexamining the international importance of languages. *HCIAS Working Papers on Ibero-America* (2022).
64. Statista. Infographic: English Is the Internet's Universal Language https://www.statista.com/chart/26884/languages-on-the-internet (2022). Retrieved 2023-03-27.
65. European Commission. EU-Ukraine: Standing together https://eu-solidarity-ukraine.ec.europa.eu/eu-ukraine-standing-together_en (2023). Retrieved 2023-03-30.
66. White House. FACT SHEET: One Year of Supporting Ukraine https://www.whitehouse.gov/briefing-room/statements-releases/2023/02/21/fact-sheet-one-year-of-supporting-ukraine/ (2023). Retrieved 2023-03-30.

Acknowledgements

We would like to thank Matthias Häberle for his support in collecting the data. This work is supported by the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS". This work is also supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. [ERC-2016-StG-714087], Acronym: *So2Sat*). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

D.R., B.I.D., P.W.T., X.X.Z. and G.K. conceived the research. P.W.T., B.I.D. and G.K. supervised the research. D.R. and G.K. designed the methodology. D.R. collected and processed the data, conducted the study and analyzed the results. D.R. created the visualizations. D.R., B.I.D., P.W.T. and G.K. wrote the first draft. All authors edited and approved the article.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44271-023-00045-6.

Correspondence and requests for materials should be addressed to Daniel Racek.

Peer review information : *Communications Psychology* thanks Han-Wu-Shuang Bao and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Jixing Li and Antonia Eisenkoeck. A peer review file is available.

Part III.

Developing New Models

9. Capturing the Spatio-Temporal Diffusion Effects of Armed Conflict: A Non-parametric Smoothing Approach

Contributing article

Racek, D., Thurner, P., and Kauermann, G. (2025). Capturing the Spatio-Temporal Diffusion Effects of Armed Conflict: A Non-parametric Smoothing Approach. *Center for Open Science, (No. q59dr_v2).* https://doi.org/10.31219/osf.io/q59dr_v2. Accepted at *Journal of the Royal Statistical Society Series A: Statistics in Society.*

Data and code

Available at https://osf.io/ypwuv/.

Supplementary material

Supplementary material is available online.

Author contributions

The idea to capture the diffusion of armed conflict in a statistical model stems from all authors. Daniel Racek and Göran Kauermann designed the smoothing approach to capture the spatio-temporal history of conflict for each grid cell. Daniel Racek processed the datasets, conducted the study, analysed the results and created the visualizations. Göran Kauermann provided valuable feedback for the visualizations. All authors wrote the first draft and were involved in improving and editing the article.

CAPTURING THE SPATIO-TEMPORAL DIFFUSION EFFECTS OF ARMED CONFLICT: A NON-PARAMETRIC SMOOTHING APPROACH

Daniel Racek*
Institute of Statistics, LMU Munich
München, Germany

Paul W. Thurner
Institute of Political Science, LMU Munich
München, Germany

Göran Kauermann
Institute of Statistics, LMU Munich
München, Germany

ABSTRACT

Facilitated by advancements in conflict event databases, studies have moved towards predicting armed conflict and understanding its determinants subnationally. However, existing statistical models do not analyze nor capture the diffusion of armed conflict, and hence do not adequately account for its dependence across both time and space. To address this, we introduce a regression approach that simultaneously captures both spatial and temporal dimension of the diffusion of armed conflict through non-parametric smoothing, while all its effects and parameters remain fully interpretable. Using fine-grained conflict data on Africa, we observe that diffusion exhibits long-lasting and far-reaching dependencies that decay exponentially in both space and time, thus highlighting the importance of controlling for these effects. We illustrate the flexibility of our method for studying conflict diffusion, by investigating the role of population in the transmission of conflict. We find that conflict typically breaks out in densely populated areas, and from there diffuses, specifically to lower population areas.

Keywords Armed Conflict, Diffusion, Smoothing, Spatio-temporal Modelling

1 Introduction

Predicting armed conflict and understanding its determinants has been the key focus of conflict research for decades (Blattman and Miguel, 2010; Hegre et al., 2017). However, only in recent years the field has moved away from country-level designs towards analyzing conflict in more fine-grained subnational areas (Bazzi et al., 2022), as new conflict event databases such as the Uppsala Conflict Data Program (UCDP) Georeferenced Event Dataset (GED) (Sundberg and Melander, 2013) and the Armed Conflict Location & Event Data Project (ACLED) (Raleigh et al., 2010) have become available. Paired with the emergence of new data sources such as social media and remote sensing data obtained from satellite imagery (Racek et al., 2024), numerous new studies have been published that analyze conflict at a local, disaggregate level (Von Uexkull et al., 2016; Abidoye and Calì, 2021; Mueller et al., 2022). As a consequence, more advanced statistical models are finding their way into the field (Fritz et al., 2022).

Recent studies at the subnational level have often utilized monthly observations across a lattice grid of equally-sized cells (Ge et al., 2022; Rød and Weidmann, 2023; McGuirk and Nunn, 2024). However, most of the employed models do not adequately account for the dependence structure of conflict over both time and space. Using Monte Carlo simulations, Schutte and Weidmann (2011) have shown that armed conflict indeed exhibits patterns of spatio-temporal diffusion, meaning, that future conflict is influenced by past conflict within the same grid cell but also by past conflict in its (further-away) neighbours. Although numerous studies have reported that past conflict is the best predictor of

*Corresponding author: daniel.racek@lmu.de

future conflict (Bazzi et al., 2022; Racek et al., 2024), conflict diffusion has not yet received the necessary attention, especially compared to other fields such as criminology (Fitzpatrick et al., 2019; Kounadi et al., 2020; Butt et al., 2020) or epidemiology (Meyer and Held, 2017; Briz-Redón and Serrano-Aroca, 2020). Diffusion follows patterns of social behaviour and can often be attributed to rational decisions of actors (Goyal, 2023, p.590). For example, Mueller et al. (2022) argue that ethnic groups act strategically and rationally choose where and when to attack. Schutte (2017) shows that conflict takes place both close as well as far away from capital cities, likely due to differences in location-dependent power between both rebels and the government. However, including such underlying diffusion patterns into statistical models remains a challenge, both practically, i.e., in terms of design, as well as computationally (Weidmann and Ward, 2010). Hence, empirical insights on the diffusion of armed conflict are still very limited. Instead, most studies treat the dependence that arises from this diffusion "as a nuisance" (Schutte and Weidmann, 2011, p.152) and try to control for it by including simple temporal and (rarely) spatial lags. Indeed, many studies do not consider conflict in neighbouring grid cells at all (see e.g., Bazzi et al., 2022; Fritz et al., 2022; Chadefaux, 2022; Schon et al., 2023), or make the simplifying assumption that spatial dependence is limited to the direct neighbours and only existent for a small number of time lags (Weidmann and Ward, 2010). While the situation is slowly starting to improve for black-box machine learning models (see e.g., Radford, 2022; Brandt et al., 2022), for interpretable regression models and other statistical models it has not. Consequently, (causal) analyses on the determinants of conflict may over- or underestimate the impact of the respective predictors of interest due to omitted-variable bias arising from the uncontrolled dependence (see Cook et al., 2023 for an recent in-depth discussion on this), and thus may derive incorrect policy implications (Schutte and Weidmann, 2011).

To address these limitations, in this work, we propose a regression model that can flexibly incorporate both the spatial and the temporal dimension of the diffusion of armed conflict, while – contrary to black-box machine learning models – all its effects and parameters remain fully interpretable. Our main contributions are the following. First, our findings highlight that organized armed conflict indeed exhibits substantial diffusion across time and space, and that this dependence cannot be captured by traditional models employed in the field. Using our proposed model, which covers diffusion up to 550km in distance and 24 months in the past, we demonstrate how conflict is triggered across cells, over both varying distances and time lags. Second, we exemplify the flexibility of our approach for studying the diffusion of conflict, by exploring the role of population in the transmission process. We find that diffusion is heavily driven by population structures. Conflict generally breaks out in densely populated areas and from there diffuses across the region, disproportionately affecting less populated areas. Third, although our method is designed to study armed conflict, it is highly adaptable and can be applied to analyze or account for any other diffusion process across spatio-temporal units with long-lasting and far-reaching dependencies. Analogue to other regression models, the distribution and thus unit of analysis can be specified freely. Fourth and finally, model estimation can be carried out with existing R estimation routines and packages. We provide a tutorial on how researchers can apply our method for their own work in our Supplementary Material and repository (https://osf.io/ypwuv/).

To study conflict diffusion, we draw on conflict fatalities from UCDP GED and employ the 0.5 x 0.5 decimal degree (roughly 55 x 55 km^2 at the equator) lattice grid structure across Africa, predominantly used in the literature. We design a generalized additive model (GAM) (Wood, 2017) with a flexible non-parametric spatio-temporal smoothing component over past conflict, to predict the monthly conflict fatalities in each grid cell, assuming a Poisson distribution. Our smoothing basis is constructed through a set of exponential decay functions with varying decay rates.

Thus, our proposed model has similarities to a spatio-temporal Hawkes process (Reinhart, 2018; Jun and Cook, 2022), a self-exciting point-process with an underlying Poisson intensity, that typically employs exponentially decaying triggering functions across time and space, increasingly used in epidemiology (Meyer et al., 2012; Schoenberg et al., 2019) and to model the dynamics of crime (Mohler, 2014; Reinhart, 2018; Reinhart and Greenhouse, 2018). Likewise, (spatio-temporal) log-Gaussian Cox processes (LGCPs) (Zammit-Mangion et al., 2012) can model complex spatial point patterns through the use of Gaussian fields. Note, the close connection between Gaussian processes and kriging (Zammit-Mangion and Cressie, 2021; Christianson et al., 2023). The main difference between LGCPs and Hawkes processes is that the latter explicitly differentiate between exogenous and endogenous effects through their self-excitation mechanism (for a comparison, see Miscouridou et al., 2023). Our model is more closely aligned with Hawkes-type processes, as our diffusion component is included directly in the linear predictor of the GAM. This allows us to investigate specific (endogenous) spatio-temporal diffusion mechanisms, separate from other (exogenous) covariates, while being able to use existing GAM software for efficient model estimation. Although point processes (see e.g., Daley and Vere-Jones (2006), and Møller and Waagepetersen (2017), for an introduction and overview), such as Hawkes processes and LGCPs, are, in principle, the true generators of individual conflict events, each event is only coded with an approximate date and location, i.e., the events come with a high spatial and temporal uncertainty (e.g., they are often assigned to the capital city of a district or state). Thus, conflict is generally analyzed in coarser spatio-temporal units, instead of as individual events, covering larger areas (e.g., grid cells or administrative units) and longer time intervals (e.g., months or years). Under these conditions, employing point process models may lead to

biased estimates, incur unnecessarily high computational costs and they may become unstable due to the low number of events in many regions. Hence, in our proposed diffusion model, we make explicit use of the lattice grid, i.e., discretize both time and space, which leads to a stable and comparably low-cost estimation routine. This also means, fitting more complex model specifications over larger sets of areas, such as entire continents, becomes computationally feasible.

Our approach is also related to models from both the spatial and temporal econometrics literature, such as the autoregressive (AR) and autoregressive exogenous input (ARX) model (Box et al., 2015), the spatial-autoregressive (SAR) (LeSage and Pace, 2009) and Durbin model (Mur and Angulo, 2006), as well as the recently proposed spatio-temporal autoregressive distributed lag (STADL) model (Cook et al., 2023). Contrary to these, we do not only include temporal and/or spatial lags, but an exhaustive combination of both. Hence, we allow for (varying) direct effects of the spatio-temporal lags of the dependent variable and/or potentially any other exogenous variable(s). One could conceive this design as including an entire spatio-temporal cylinder of past information into the model for each observation (see Figure 2). This implies that a large number of coefficients needs to be fitted. Hence, smoothing becomes a necessity in order to reduce the impact of noise on the model fit and to correctly capture the diffusion. Predicting subnational conflict, specifically its onset and escalation, has continuously been recognized as an extremely difficult task, as there are very few observations of actual conflict (Hegre et al., 2021, 2022; Bazzi et al., 2022; Racek et al., 2024). Hence overfitting (Cawley and Talbot, 2010; Mullainathan and Spiess, 2017), in which a model does not generalize (as well) to new unseen data points, becomes an issue. Our proposed smoothing approach overcomes these difficulties, as will be demonstrated by our out-of-sample evaluation.

One can also draw parallels to time-space (adaptive) smoothing approaches such as the one recently applied for conflict in Tapsoba (2023), and more generally in other fields such as climatology (Lee and Durbán, 2011) and seismology (Helmstetter and Werner, 2014). Contrary to these, we do not smooth across explicit points in time and space, but instead across the spatio-temporal history of each observation. While the former captures the general intensity of conflict in certain regions at specific points in time, our approach explicitly models the diffusion and thus dependence of conflict. Note, these two approaches can be seen as complementary. In fact, our model additionally includes a spatial effect across the lattice grid, separate from the spatio-temporal diffusion component, in order to account for the varying levels of conflict intensity across the continent.

Finally, our contribution is closest to the small body of literature investigating spatial patterns in armed conflict. Mueller et al. (2022) find exponentially decaying diffusion effects in distance that differ between selected countries based on the composition of ethnic groups. Studying conflict in the North Caucasus, Zhukov (2012) observes spatial diffusion facilitated by road networks. Developing a statistical test using Monte Carlo simulations, Schutte and Weidmann (2011) find statistically significant spatio-temporal diffusion effects for first-order neighbouring cells in four cases of civil war. The main difference of our work is that we do not only account for the spatial dimension of the diffusion, but instead also combine it with the temporal dimension, to consider spatio-temporal lags for both large distances as well as many past points in time. Notably, contrary to e.g. Schutte and Weidmann (2011), we capture this through a regression model. Hence, we can measure effect sizes of the diffusion, combine it with other explanatory variables, i.e., investigate interactions, or use the model to study other mechanisms while controlling for the diffusion.

The remainder of the article is organized as follows. Section 2 describes all utilized data sources and how they are pre-processed. In Section 3 we thoroughly describe our proposed method, introduce reference models employed in the literature, explain our model evaluation criteria and discuss the model extension to investigate the role of population. Then, in Section 4, we present our results. Finally, in Section 5, we discuss the results and conclude our work.

2 Data

We rely on conflict data from the widely known UCDP GED (Sundberg and Melander, 2013), which reports events of organized violence resulting in at least one estimated death from 1989 onwards. These events are systematically collected by an experienced team of researchers, and draw on information from national and international news reports, international organizations as well as NGOs. Each event is assigned an approximate date and location, type of violence and an estimated number of fatalities. We utilize battle-related fatalities, i.e., deaths which result from either state-based or non-state violence between organized parties (similar to other recent studies, see e.g., Vesco et al., 2022; Rød and Weidmann, 2023). We discard a small number of events with unknown/imprecise time (1.6%) and location estimates (14.3%), and only keep those that have a time precision $\leq$ 1 month and for which at least the second order administrative region is known.

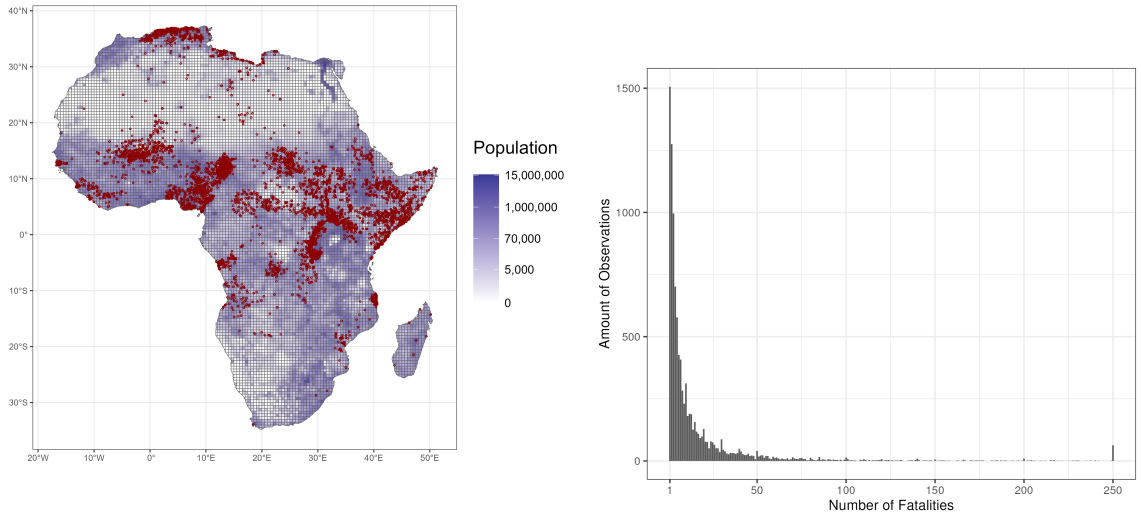
The remaining events are matched on a monthly basis to the commonly employed PRIO grid cells (Tollefsen et al., 2012), used across numerous studies in conflict research (Koren and Bagozzi, 2017; Vesco et al., 2022; McGuirk and Nunn, 2024). They have a size of 0.5 x 0.5 decimal degrees, which is roughly equivalent to 55 x 55 km^2 at the equator,

and cover the whole world. We employ these grid cells instead of administrative units, in order to avoid problems with varying sizes of the latter within and across countries (see also Racek et al., 2024). For reasons of comparability, we study Africa, where armed conflict is prevalent and many research studies have been conducted in (Von Uexkull et al., 2016; Schutte, 2017; Bagozzi et al., 2017; Koren and Bagozzi, 2017; Abidoye and Calì, 2021; Maconga, 2023). A visualization of the grid cells with the considered events is provided in 1a. We analyze conflict on a monthly basis, as a more fine-grained temporal resolution becomes problematic due to the imprecision in reported event dates.

We restrict our analysis period to the years 2000 to 2020, due to the limited availability of some of our data sources, and focus on the monthly amount of battle-related fatalities in each cell. Altogether, this means we have a total of 10,640 grid cells and 252 months, resulting in 2,681,280 observations. Note, armed conflicts are in most parts of the world, including Africa, extremely rare events. Only 0.38% of all monthly observations contain one or more fatalities (mean = 0.068), with a maximum of 1024 fatalities in a single month. The distribution of all non-zero observations is reported in 1b.

We draw on population data from the WorldPop project (Tatem, 2017). Using satellite imagery, surveys, census data and various other geospatial datasets, it estimates yearly population numbers across the world from the year 2000 onwards. We employ the 1km resolution dataset and derive the total amount of population for each cell for each year. In order to avoid a leakage of future information into our models, we lag these numbers in the matching process by a year. As 1a shows, conflict events typically take place in areas with higher population numbers.

From 1a we can observe that grid cells differ in size, as cells at the coast are cropped by definition. Hence, we compute the total area covered by each cell in km^2. Finally, we also draw on yearly country-level PPP-adjusted GDP data from the World Bank (2024), and the revised Polity Score (Polity2) as an indicator for the yearly level of democracy from the Center for Systemtic Peace (Marshall et al., 2017) (both lagged by a year). As cells might cover territory of more than a single country, we determine each cell's country through the majority coverage of its area, using border information provided by CShapes (Schvitz et al., 2022). We provide a Table with descriptive statistics of all variables in Supplementary Material S.1.



(a) Visualization of all grid cells (black-bordered squares) and conflict events (red dots). Each cell is colored with respect to its logged population (from white to blue) in 2011. Events with > 250 fatalities are censored and set to 250 instead for visualization purposes.
in this study are plotted based on the exact latitude and longitude reported in UCDP.

(b) Distribution of number of fatalities for all conflict events. All conflict events consideredfor

Figure 1: Descriptive plots.

3 Methods

3.1 Standard Spatio-Temporal Diffusion Model

Let $Y_{t,s}$ denote the number of fatalities occurring in month t in cell location $\boldsymbol{s}$. We define $\boldsymbol{s} = (r, c)^\top$ as a bivariate location vector, where r refers to the row, and c to the column of the respective location in the grid. As the number of fatalities are count data we assume that

$$Y_{t,s} \sim \text{Poisson}(\lambda_{t,s})$$

with intensity $\lambda_{t,s}$. Note, in theory, the model described subsequently could be fitted with any distribution from the exponential family, with the corresponding change in interpretation as the conditional mean of the chosen distribution. In our spatio-temporal diffusion model (we will refer to this as model M1) , we define the intensity as

$$\lambda_{t,s} = exp(\boldsymbol{x}_{t,s}^\top \boldsymbol{\beta}_x + g(\boldsymbol{s}) + \gamma(H_{t,s})), \tag{1}$$

where $\boldsymbol{x}_{t,s}$ is a feature vector including the intercept, the time-constant cell size, the lagged (logged) population of each cell, the lagged (logged) country-level GDP per capita and the lagged country-level Polity Score. We include the latter two controls, as countries with a lower GDP per capita (Pinstrup-Andersen and Shimokawa, 2008) and a lower level of democracy (Hegre, 2014) typically experience more conflict (outbreaks). We include population, as it has been shown to be among the best and few consistent predictors of armed conflict (Racek et al., 2024), with higher population numbers being associated with an increased risk of conflict (Raleigh and Hegre, 2009). Theoretically, $\boldsymbol{x}_{t,s}$ could be extended to include any additional covariates. The component $g(\boldsymbol{s})$ represents a smooth location effect that captures time-constant levels of conflict intensity across Africa. For this, we employ thin plate regression splines (Wood, 2003), which are low rank approximations of thin plate splines. They avoid the problem of knot placement and are isotropic, i.e., independent to rotations of our grid. We denote $\gamma(H_{t,s})$ as our diffusion effect of conflict, using the notation $H_{t,s}$ to express that we are utilizing the history and neighbouring history H of a cell location $\boldsymbol{s}$ at time point t in our flexible spline representation. More specifically, we make use of a basis function representation leading to the linear structure

$$\gamma(H_{t,s}) = \boldsymbol{b}(H_{t,s})^\top \boldsymbol{u},$$

where $\boldsymbol{b}(H_{t,s})$ is a high-dimensional basis and $\boldsymbol{u}$ are the associated basis coefficients to be estimated. We model the basis as follows. Let $\tau > 0$ be the maximum time lag considered and $\delta \geq 0$ be the maximum distance considered. We then define our basis as

$$\boldsymbol{b}(H_{t,s}) = \sum_{\tilde{t}=t-\tau}^{t-1} \sum_{\tilde{s} \in N_\delta(s)} \boldsymbol{a}(\tilde{t}, t) \otimes \boldsymbol{o}(\tilde{\boldsymbol{s}}, \boldsymbol{s})\ log(Y_{\tilde{t},\tilde{s}} + 1)$$

where $\otimes$ is the Kronecker product of all our basis vectors in time ($\boldsymbol{a}$) and space ($\boldsymbol{o}$), where

$$N_\delta(\boldsymbol{s}) = \{\tilde{\boldsymbol{s}} : ||\tilde{\boldsymbol{s}} - \boldsymbol{s}|| \leq \delta\}$$

defines the neighborhood of location $\boldsymbol{s}$. The distance between $\tilde{\boldsymbol{s}}$ and $\boldsymbol{s}$ is measured through Euclidean distance on grid cell indices. In practice, this means we sum up our time-space basis vector (that results from the Kronecker product) over all past time points for which $t - \tilde{t} \leq \tau$ and all neighbouring cells for which $||\tilde{\boldsymbol{s}} - \boldsymbol{s}|| \leq \delta$.

The individual basis functions in both time and space are defined as exponential decay functions, which are scaled such that their maximum value is 1 and their minimum value is 0 for all w. To be specific, we set

$$a_k(\tilde{t}, t) = \frac{exp(-w_k\ (\tilde{t} - t)) - exp(-w_k\ \tau)}{exp(-w_k) - exp(-w_k\ \tau)},$$

where $w_k > 0$ and $w_k = \{w_1, ..., w_K\}$ are pre-defined decay rates. Accordingly we define the basis in space as

$$o_g(\tilde{\boldsymbol{s}}, \boldsymbol{s}) = \frac{exp(-v_g\ ||\tilde{\boldsymbol{s}} - \boldsymbol{s}||) - exp(-v_g\ \delta)}{1 - exp(-v_g\ \delta)}$$

with decay rates $v_g = \{v_1, ..., v_G\}$. Hence, $\boldsymbol{b}(H_{t,s}) \in \mathbb{R}^{(KG)\times 1}$, where K and G are chosen to be large for model flexibility, necessitating smoothing to prevent overfitting.

Our method can be conceived as including a spatio-temporal cylinder of past information into the model for each observation (see Figure 2), over which we construct our basis. This implies, we allow for direct spatio-temporal dependence for any of the included lags. Similar to other temporal and spatial dependence models, these direct effects are amplified indirectly through spatial and temporal multipliers. For example, conflict in a neighbouring cell in the past increases present conflict in a given cell directly, this in turn (indirectly) increases future conflict in the cell through temporal dependence and in its neighbours through spatio-temporal dependence. This has close similarities to

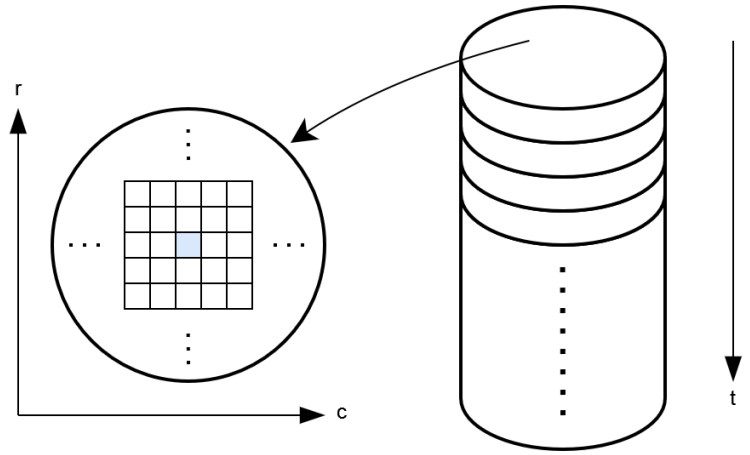


Figure 2: Visualization of all information included in the basis representation of the diffusion model for each observation. The cylinder results from using the Euclidean distance function in the neighborhood definition combined with a maximum time lag. For each time lag t, we have a slice of conflict information comprising of the neighborhood of the blue-colored grid cell. r and c refer to the row respectively column in the grid definition. The dots denote additional grid cells (left part of Figure) and temporal slices (right part of Figure). This spatio-temporal cylinder (times the number of observations) is the input to each basis function over which the smooth effects are ultimately fitted.

the self-exciting behaviour of a Hawkes process (Hawkes, 1971), i.e., the occurrence of conflict in a cell makes future conflict fatalities in the cell itself and its surroundings much more likely in the short- to medium-term, with further conflict fatalities continuing and potentially amplifying this process. This ultimately allows us to model the dynamic spread of conflict across regions.

We utilize a set of ten basis functions each. We set $\tau = 24$ to consider the past 24 lags in time, and set $\delta = 10$ to consider all neighbouring cells up to a distance of (roughly) 550 km from the source cell (horizontally and vertically this would include cells up to the 10th-order neighbour). This gives us in total a combination of $10 \times 10 = 100$ basis functions that differ across time and/or space. Both τ and δ are chosen based on upper bounds of effects identified in the literature (Zhukov, 2012; Mueller et al., 2022; Fritz et al., 2022). Hence, we ensure to include a sufficient amount of temporal lags and neighbouring cells that might still exhibit direct dependencies. Theoretically, due to the smoothing, these upper bounds could be increased arbitrarily. However, to reduce computational complexity, reasonable cut-offs are necessary.

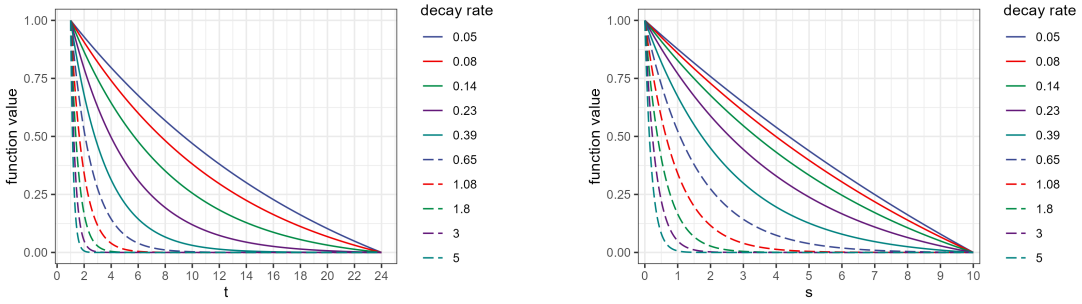
As highest decay rate we choose $w = 5$, resulting in a basis function that only captures the first temporal lag respectively the cell itself (i.e., no spatial lag). As lowest decay rate we choose $w = 0.05$, resulting in a basis function which is (almost) linear. The remaining w are chosen such that there is a constant multiplicative increase in their rate. To be specific, we choose $w_k = v_g = \{0.05, 0.0834, 0.1391, ..., 5\}$, with $w_k = 1.6681\ w_{k-1}$ for $k > 1$. Both individual sets of basis functions are visualized in Figure 3.

Due to the large amount of coefficients to be estimated (a total of $24 \times 44 = 1056$) , smoothing becomes a necessity in order to reduce the impact of noise on the model fit and to correctly capture the diffusion effects. Here, this means we assume that neighbouring points, i.e., lags in both time and space, have similar effect sizes. To guarantee smoothness of $\gamma(H_{t,s})$, we employ a ridge penalty, i.e., we utilize the penalized log-likelihood

$$\ell_{pen}(\boldsymbol{\theta}, \boldsymbol{u}, \rho) = \ell(\boldsymbol{\theta}, \boldsymbol{u}) - \frac{1}{2}\rho\, \boldsymbol{u}^\top \boldsymbol{u},$$

where ℓ is the log-likelihood, ρ the penalty and $\boldsymbol{\theta}$ the remaining parameter vector. This results in a generalized additive model (GAM) that we can estimate with the R package *mgcv* (Wood, 2017). For a discussion on the estimation of the penalty ρ we refer to Supplementary Material S.3.

An additional advantage of our method is, contrary to traditional temporal and spatial dependence models (e.g., AR, SAR), one does not have to choose the presumed amount of lags of the true process in advance. Instead, a researcher may simply choose sufficiently large upper bounds (as done here). Lags beyond the dependency bounds, i.e., without an effect, will be (almost) fully penalized out of the model and their effects will be close to zero.



(a) Temporal basis functions. $t = t_2 - t_1 \leq \tau$ refers to the number of time lags between two time points t_2 and t_1.

(b) Spatial basis functions. $\boldsymbol{s} = ||\boldsymbol{s_1} - \boldsymbol{s_2}|| \leq \delta$ refers to the Euclidean distance between two cell locations $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$.

Figure 3: Visualization of the basis functions employed in the diffusion model.

We make the conscious decision to utilize exponential decay functions in our basis, as exponentially decaying effects have been reported in the conflict literature before (Hegre et al., 2019; Fritz et al., 2022; Mueller et al., 2022) and diffusion across numerous fields typically follows and is modelled through such decays (Reinhart, 2018; Meyer et al., 2012). Thus, the model has an implicit tendency to fit an exponentially decaying effect over both time and space, while it can still fit other patterns if more applicable. We also experimented with other approaches to construct our basis, such as capturing the history and neighbouring history of a cell through separate basis blocks of the spatio-temporal cylinder. While they all lead to closely similar results, as expected, issues arise when trying to achieve a smooth fit due to forgoing this implicit tendency. We refer to Supplementary Material S.6 for details.

3.2 Extending the Model: Interplay between Population and Diffusion

The above modelling strategy can now easily be extended to improve our understanding of specific diffusion patterns and to analyze if and how these depend on additional sets of variables. Here, we explore population as a driver for the underlying diffusion process. In the literature, country-level (Brückner, 2010) as well as disaggregated analyses (Raleigh and Hegre, 2009) have found that higher population numbers are generally associated with an increased risk of conflict. Additionally, both diffusion as well as violent events are rooted in social behaviour and interactions (Goyal, 2023; Ives and Lewis, 2020; Pinckney, 2018). Hence, we suspect that conflict transmission paths can be traced along population numbers.

We incorporate population into the diffusion, by adding a multiplicative interaction between the space-time effect of conflict and the population of each cell. To be specific, we define

$$\gamma_{pop}(H_{t,s}) = \widetilde{lpop}(t-12, \boldsymbol{s})\boldsymbol{b}(H_{t,s})^\top \boldsymbol{u}_{pop}.$$

where

$$\widetilde{lpop}(t, \boldsymbol{s}) = \frac{lpop(t, \boldsymbol{s})}{lpop_{max}},$$

is the normalized logged population (from 0 to 1), of cell location $\boldsymbol{s}$ at time point t. The latter is obtained by dividing the respective logged population by the maximum $lpop_{max}$ over all $\boldsymbol{s}$ and t. We employ the normalized logged population, as it makes the interpretation of the effects substantially easier, and its lagged version in $\gamma_{pop}(.,)$ in order to avoid potential leakage of future information into our model.

Solely adding the interaction $\gamma_{pop}(H_{t,\mathbf{s}})$ into the model equation, would assume by definition that cells with a higher population are more strongly affected by the diffusion of conflict (given that the base diffusion effect is already positive). Hence, we estimate

$$\lambda_{t,s} = exp(\boldsymbol{x}_{t,s}^\top \boldsymbol{\beta}_x + g(\boldsymbol{s}) + \gamma(H_{t,s}) + \gamma_{pop}(H_{t,s})), \tag{2}$$

which allows for a more flexible effect of population on the diffusion. We will refer to this as model M2.

3.3 Reference Models in the Literature

To evaluate our proposed diffusion model, we formulate a range of reference models, allowing us to pursue comparisons with models often employed in the conflict literature. Our focus lies on fully interpretable regression models

used to study and improve our understanding of armed conflict. We explicitly do not include any machine learning model in the comparison, as their fitted effects are generally not interpretable and forecasting is not our main objective. Instead, we use the comparison to showcase that diffusion is not fully captured by the set of interpretable models employed in the field.

For our first reference model (M0-1), we define the monthly cell intensity in the most basic form as

$$\lambda_{t,s} = exp(\boldsymbol{x}_{t,s}^{\top}\boldsymbol{\beta}_x + g(\boldsymbol{s})), \tag{3}$$

i.e., we exclude conflict diffusion and fit the model without any information on past conflict. As before, our feature vector $\boldsymbol{x}_{t,s}$ includes the intercept, the time-constant cell size, the lagged (logged) population of each cell, the lagged (logged) country-level GDP per capita and the lagged country-level Polity Score.

In the conflict literature, models typically only include information on past conflict within a cell, often ranging from 1-12 months (Bazzi et al., 2022; Schon et al., 2023; Chadefaux, 2022; Mueller and Rauh, 2022; Fritz et al., 2022). Hence we set our second reference model (M0-2) to

$$\lambda_{t,s} = exp(\boldsymbol{x}_{t,s}^{\top}\boldsymbol{\beta}_x + g(\boldsymbol{s}) + \boldsymbol{y}_{t,s}^{\top}\boldsymbol{\beta}_y), \tag{4}$$

where we define $\boldsymbol{y}_{t,s}$ to be a T dimensional vector capturing past conflict at time point t for cell location $\boldsymbol{s}$. The inclusion of past conflict can be done in various ways. We opt for preserving as much information as possible and include each lag individually, again using the logged version, which yields $\boldsymbol{y}_{t,s} = (\log(y_{t-1,s}+1), \log(y_{t-2,s}+1), ..., \log(y_{t-T,s}+1))^{\top}$ and set $T = 1, 12, 24$, i.e., we fit the model using 1, 12 and 24 lags of past (logged) conflict fatalities respectively. Note, model M0-2 only accounts for conflict diffusion within a cell, i.e., it only captures the temporal dimension of the diffusion.

A simple strategy to also account for the spatial dimension, without relying on more complex modelling approaches such as ours, is to include first-order neighbouring lags (Weidmann and Ward, 2010). To define this formally in our given setting, we can return to our neighborhood definition from Section 3.1. All first-order neighbouring cells are captured by the neighbourhood

$$N_{\sqrt{2}}(\boldsymbol{s}) = \{\tilde{\boldsymbol{s}} : ||\tilde{\boldsymbol{s}} - \boldsymbol{s}|| \leq \sqrt{2}\}.$$

as first-order neighbours in the diagonal have an Euclidean distance of exactly $\sqrt{2}$ in our given grid definition. In order to avoid an overparameterization of our model, we do not include the lags of each neighbour individually, but instead include the sum of logged fatalities of all neighbours (summing up neighbouring cells is a common strategy in the conflict literature, see e.g., Rød and Weidmann, 2023; Lindholm et al., 2022; D'Orazio and Lin, 2022). We define individual

$$z_{t,s} = \sum_{\tilde{s} \in N_{\sqrt{2}}(s)} log(y_{t,\tilde{s}} + 1)$$

and denote the vector of the first T (first-order) neighbouring lags as $\boldsymbol{z}_{t,s} = (z_{t-1,s}, z_{t-2,s}, ..., z_{t-T,s})^{\top}$. We can now extend M0-2 to

$$\lambda_{t,s} = exp(\boldsymbol{x}_{t,s}^{\top}\boldsymbol{\beta}_x + g(\boldsymbol{s}) + \boldsymbol{y}_{t,s}^{\top}\boldsymbol{\beta}_y + \boldsymbol{z}_{t,s}^{\top}\boldsymbol{\beta}_z). \tag{5}$$

We will refer to this model as M0-3. We similarly fit M0-3 with $T = 1, 12, 24$ lags respectively. For 24 lags this yields $\boldsymbol{z}_{t,s} = (z_{t-1,s}, z_{t-2,s}, ..., z_{t-24,s})^{\top}$.

3.4 Estimation & Evaluation Strategy

We set up the estimation and evaluation of our models the following. We use the years 2000 to 2018 to fit our models, and evaluate out-of-sample performance on all observations from 2019 to 2020. Out-of-sample evaluation is necessary, as predicting conflict has been shown to be a particularly difficult task (Bazzi et al., 2022; D'Orazio and Lin, 2022; Racek et al., 2024), partially due to the large amount of no-conflict observations, which means models can easily pick up and model noise and thus exhibit poor generalization behaviour. In the forecasting literature, this evaluation approach is also known as time-series cross-validation (Petropoulos et al., 2022, p.736).

To allow for a model comparison both in- as well as out-of-sample, rooted in statistical theory (Dunn et al., 2018), we compute the mean Poisson unit deviance defined as

$$\overline{D} = \frac{2}{n}\sum_{i=1}^{n}(y_i \, log\frac{y_i}{\hat{y}_i} - y_i + \hat{y}_i),$$

where y_i are the observed and $\hat{y}_i$ the predicted fatalities for an observation i, and n denotes the total number of observations either in- or out-of-sample. By removing the division (i.e., using the sum instead), this is equivalent to

$-2\ell(\hat{\boldsymbol{\theta}})$ in the Poisson setting (up to a small constant) , where $\ell(\hat{\boldsymbol{\theta}})$ is the loglikelihood of the fitted model. This would be analogous to the mean squared error when assuming a normal distribution.

In order to understand how well the models are performing relatively compared to a null (intercept-only) model, we add a measure of explained deviance into our model analysis. We define this as

$$\text{Explained Deviance} = D_{\text{expl.}} = 1 - \frac{D}{D_0} = 1 - \frac{-2\ell(\hat{\boldsymbol{\theta}})}{-2\ell(\hat{\boldsymbol{\theta}}_0)},$$

where D denotes the deviance of the respective model and the 0-subscript the null model. This would be equivalent to R^2 for a normal distribution. We fit a separate null-model on our out-of-sample observations to better understand generalization behaviour. Finally, we also look at in-sample performance by comparing our models using the AIC.

4 Results

We will start this section by analyzing the diffusion effects of conflict, using model M1. Then, we will compare its performance to reference models employed in the literature. Next, we will analyze the interplay between population and conflict diffusion. Finally, we validate our findings through various robustness checks.

4.1 Diffusion of Conflict

4.1.1 Diffusion Effects

In the first row of Table 1 we report the performance of model M1 for our preferred penalty ρ. Recall that we define $\overline{D}$ as the mean Poisson unit deviance and $D_{\text{expl.}}$ as the explained deviance compared to a null-model. Our diffusion model can explain roughly 37.3% of the deviance in-sample and 33.5% out-of-sample, with a mean Poisson unit deviance of 0.56 and 0.70 respectively. The small difference between in- and out-of-sample performance metrics suggests that our model generalizes well and overfitting is limited.

We visualize the fitted diffusion coefficients in Figure 4. These are obtained by summing over all 100 basis function coefficients and are the predictor of conflict $\gamma(H_{t,s})$ across space (x-axis) and time (y-axis). Hence, these coefficients capture by how much the linear predictor $log(\lambda_{t,s})$ increases when the past logged fatalities increase by one unit. For illustration, the coefficient for the first temporal lag for $d = 0$ is 0.5559, hence one logged fatality in the past month within the same grid cell increases the predicted fatalities by 74.35%, given $\Delta\lambda_{t,s} = exp(0.5559)$. Overall, we observe the desired smooth effect, as well as an exponentially decreasing effect of the diffusion in both time and space. All effects can be interpreted as a direct dependence between present (future) conflict in a cell and past conflict in its surroundings.

Naturally, the impact of conflict diffusion is dependent on the base intensity, which is determined by all control variables and the location of the respective cell in the grid. We provide all fitted coefficients $\boldsymbol{\beta}_x$ in the first column of Table 2 and refer for the location intensity map $g(s)$ to Supplementary Material S.2, as it is not of main interest here. The former, as expected, exhibit a positive (increasing) effect of the logged population and a negative effect for both GDP as well as Polity Score. The negative effect of the cell size is (most likely) an indication for cells at the coast

Table 1: Performance metrics of the diffusion models (top), compared to all reference models (bottom)

Model	AIC	In-sample $\overline{D}$	In-sample $D_{\text{expl.}}$	Out-of-sample $\overline{D}$	Out-of-sample $D_{\text{expl.}}$
M1	1,394,485	0.5622	0.3734	0.6959	0.3347
M2	1,370,278	0.5522	0.3845	0.6610	0.3681
M0-1	1,677,601	0.6789	0.2433	0.9491	0.0926
M0-2, 1 Lag	1,482,589	0.5985	0.3329	0.8005	0.2347
M0-2, 12 Lags	1,451,965	0.5859	0.3469	0.7640	0.2696
M0-2, 24 Lags	1,447,854	0.5842	0.3488	0.7697	0.2641
M0-3, 1 Lag	1,463,068	0.5905	0.3418	0.7767	0.2575
M0-3, 12 Lags	1,416,360	0.5712	0.3633	0.7351	0.2972
M0-3, 24 Lags	1,407,997	0.5677	0.3672	0.7461	0.2867

Notes: Results of all reference models (M0) reported for varying number of time lags. Reference models include lag structures of past conflict often employed in the literature. For diffusion models M1 and M2, we report the performance for our preferred penalty (see Supplementary Material S.3). A lower AIC, a lower $\overline{D}$ and a higher $D_{\text{expl.}}$ indicate a better performance. n = 2,425,920 observation in-sample, n = 255,360 out-of-sample.

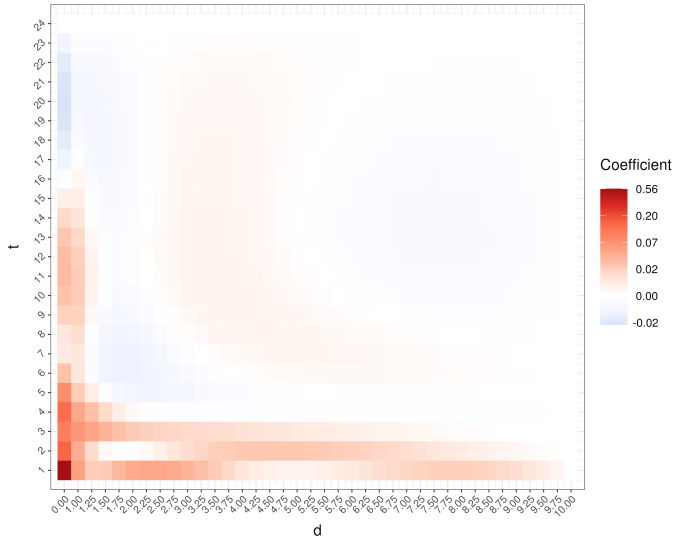


Figure 4: Conflict diffusion coefficients for model M1. The coefficients capture the change in the linear predictor $log(\lambda_{t,s})$ when the past logged fatalities increase by one unit. t refers to the respective monthly time lag, d to the distance of a cell. For $0 < d < 1$ no neighbouring cells exist (the nearest neighbour has a distance of $d = 1$), hence we removed these coefficients from the visualization.

experiencing more conflict. Not taking into account the location effects (centered around zero), a cell with a median value in all controls and no past conflict would have a base intensity of $\lambda = -2.28$, hence experience 0.10 fatalities per month.

In order to better understand the spatial diffusion patterns, we visualize them in a spatial grid map with rows (r) and columns (c) for temporal lags $t = 1, 2, 3, 4$ in Figure 5. In the origin ($c = 0$, $r = 0$), we have our cell of interest. The visualized diffusion effects are symmetrical, i.e., we can understand them as the effect conflict in cell $(0, 0)$ has on all surrounding cells, and, as the effect conflict in all surrounding cells has on cell $(0, 0)$. However, for simplicity, we will restrict interpretation to the latter. The effect is largest at $(0, 0)$ (temporal diffusion only) and exponentially decreases as the spatial distance increases (spatio-temporal diffusion). This pattern holds for all monthly lags (see also Figure 4). Surprisingly, in the Figure, we observe two rings as we move further away from the origin. In the inner ring, according to our model, the diffusion effects are substantially smaller (note the log colour scale), than in the outer ring. As the distance further increases in the outer ring, the effect eventually decreases to (roughly) 0. For $t = 2$, this pattern moves further inwards and then disappears. We will encounter the same pattern later on in Section

Table 2: Coefficients for diffusion models M1 and M2

	M1	M2
Intercept	-5.8667	-7.6843
	(0.0908)	(0.1109)
β_{area}	-0.0497	-0.0467
	(0.0089)	(0.0099)
β_{pop}	0.5399	0.6780
	(0.0020)	(0.0025)
β_{gdp}	-0.1541	-0.1249
	(0.0089)	(0.0057)
β_{polity}	-0.0169	-0.0213
	(0.0020)	(0.0008)

Notes: Standard errors in parentheses. n = 2,425,920 observations in-sample, n = 255,360 out-of-sample. Due to the large sample size, all coefficients are statistically significant ($p < 0.001$). Results reported for our preferred penalties (see Supplementary Material S.3).

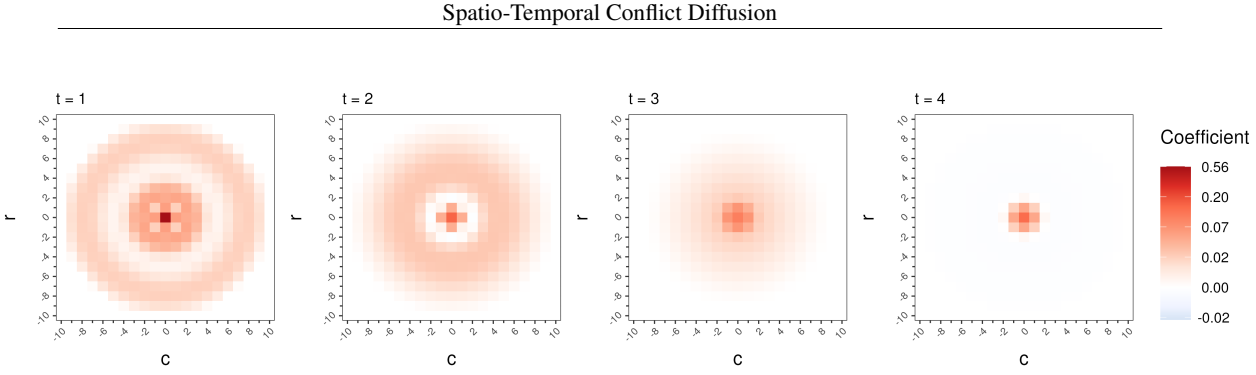


Figure 5: Diffusion coefficients of model (M1) on a spatial grid map. r refers to the row, c to the column in the spatial grid. t refers to the respective monthly time lag. Effects are symmetrical, i.e., they can be understood as the effect conflict in cell $(0, 0)$ has on all surrounding cells, and, as the effect conflict in all surrounding cells has on the cell at $(0, 0)$. These coefficients are analogue to those in Figure 4, but instead visualized on a spatial map.

4.2 and refer for an interpretation and discussion to the following Section. We refer to Supplementary Material S.14 for a larger coefficient map with exact coefficient values for selected distances and temporal lags.

4.1.2 Comparison with Reference Models from the Literature

We report the performance of our reference models (M0), as discussed above, in the bottom rows of Table 1. According to all performance metrics, our proposed diffusion model (M1) outperforms all reference models. In-sample, the differences are minor, with an increase of roughly 0.62 percentage points in explained deviance (+1.68%), compared to the best reference model. However, out-of-sample, our model performs substantially better, with an increase in explained deviance of 3.75 percentage points (+12.61%). Hence, we can conclude that our proposed diffusion model has the superior generalization behaviour, and better captures the underlying patterns of conflict diffusion. Additionally, this means that even the richest reference model, in terms of temporal and spatial lags, does not sufficiently control for the dependence of conflict across time and space. More generally, controlling for the dependence of conflict compared to not (M0-1), improves in-sample performance by 53.47% and out-of-sample performance by 261.45%. Controlling for the spatial dimension of the dependence compared to not (M0-2), increases performance by 7.05% respectively 24.14%.

In the following, we will elaborate on additional insights we can draw from the performance of the reference models. The base model (M0-1), which does not include any variables on past conflict, can explain roughly 24.33% of the null deviance, which decreases to 9.26% out-of-sample. The predictive power can be explained by the fact that some locations never or almost never experience any fatalities (see e.g., Bazzi et al., 2022; Racek et al., 2024) and populous areas are more likely to experience conflict (Raleigh and Hegre, 2009). Both are captured by the constrained set of variables in the base model. However, as evident from the results, the generalization error is substantial. Once we include lagged fatality information into the model, performance improves considerably, both in- as well as out-of-sample with 33.29% and 23.47% explained deviance respectively, given a single cell-only temporal lag (model M0-2). Notably, the out-of-sample increases are substantially larger than the in-sample increases. Hence, we can infer, as expected, that information on lagged fatalities is particularly important to reduce the generalization error. The results also highlight that adding additional information on past conflict, by including more temporal lags into the model, further improves model performance (34.88% and 26.96% explained deviance respectively). Similarly, performance considerably increases when adding information on past conflict of a cell's neighbours into the model (36.72% and 29.72%; model M0-3).

Finally, we can note that the difference between including 12 compared to 24 temporal lags is minor, both cell-only (M0-2) as well as direct neighbours (M0-3), and both in- as well out-of-sample. Our results even show, given our model specification, that including more than 12 lags is detrimental to the out-of-sample performance. This implies that mechanically adding additional information into the model in the form of spatial and temporal lags is not sufficient to improve model fit and performance. Instead, more complex modelling approaches for the diffusion, such as the one proposed in this work, are needed to make use of this information.

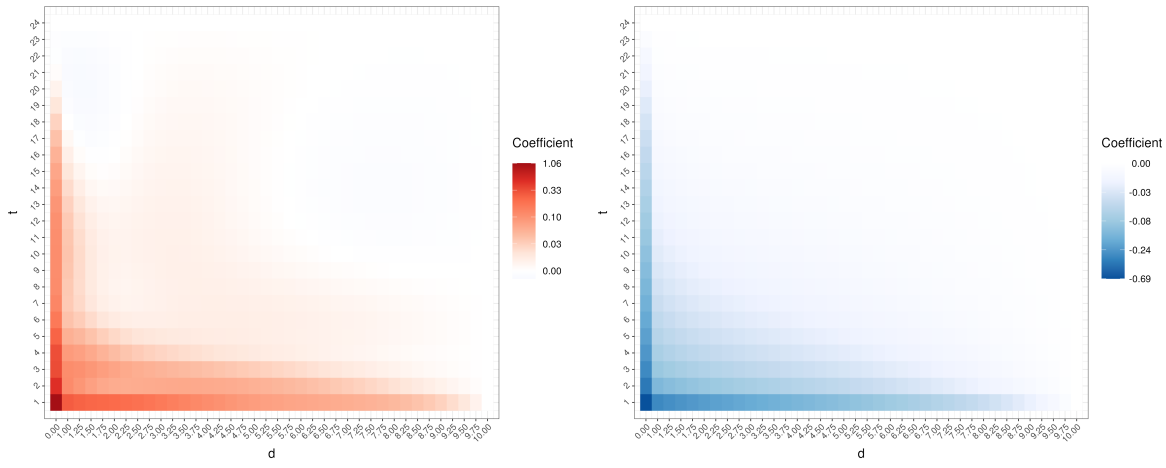
4.2 Interplay between Population and Conflict Diffusion

In the following, we will analyze the interplay between conflict diffusion and population. In the second row of Table 1 we report the performance of our population interaction model M2. As evident from the comparison with M1, the

inclusion of the interaction further improves the model fit substantially, as the explained deviance increases by 3.34 percentage points (+9.98%) out-of-sample. Also note that the difference in explained deviance in-sample compared to out-of-sample is minor (1.64 percentage points), hence model M2 generalizes particularly well. We visualize the estimated diffusion effects in Figure 6. While the base effect is highly positive (6a), the interaction with the population is negative (6b). This implies, the spatio-temporal diffusion of conflict decreases with an increase in the population of the cell of interest.

To facilitate interpretation, we visualize the spatial maps of the combined effect (for lags $t = 1, 2, 3, 4$) for the 0.05, 0.5 and 0.95 population percentile in Figure 7. This means, we fix the logged population to the respective percentile and plot the sum of base diffusion + interaction for each of them. Notably, with the inclusion of the interaction the effects are no longer symmetrical. Instead, we have to interpret the spatial maps with respect to the base cell, i.e., how past conflict in the surrounding cells affects future conflict in the cell at $(0, 0)$. We can observe that the spatial diffusion decreases substantially with an increase in population, i.e., cells with a higher population are much less affected by conflict in their surrounding cells (coefficients first temporal lag, distance $d = 1$: 0.2048, 0.0981, 0.0402). The temporal diffusion also seems to be lower, given that the population count is higher (coefficients first temporal lag, $d = 0$: 0.9025, 0.6590, 0.5269). However, note, the logged population itself is included in the model and exhibits a highly positive effect ($\beta_{pop} = 0.6780$; see Table 2). Hence, the base intensity for conflict fatalities is substantially larger in cells with high population numbers.

In combination with the results on the diffusion, the positive coefficient ($\beta_{pop} > 0$) implies that conflict tends to originate in high population cells, and from there diffuses to other cells and areas, with lower population cells being relatively much more affected (due to the substantially larger diffusion effects). While the increased risk for conflict in populous areas has been well-known and observed across numerous studies (Raleigh and Hegre, 2009; Brückner, 2010), the interplay with diffusion and thus population's dynamic role in conflict outbreaks is, to the best of our knowledge, a novel empirical finding. This is in line with Toft (2010), who finds that dispersed minorities are weakest (compared to urban groups and concentrated minorities) to create the conditions required for separatist conflict, as they are less able to mobilize and control the resources required. Herbst (2014) argues that capital areas are most important for conflict, hence political battles are much more likely to take place there. More generally, densely populated areas have either important strategic locations, or control valuable resources, thus are being important for both rebels as well as the government (Raleigh and Hegre, 2009). In all cases, arguably, conflict would first break out in these highly populous areas and from there spread across the region. Our model is able to demonstrate the corresponding effects empirically.



(a) Base Diffusion.

(b) Population Interaction.

Figure 6: Conflict diffusion coefficients for model M2. The coefficients capture the change in the linear predictor $log(\lambda_{t,s})$ when the past logged fatalities increase by one unit. t refers to the respective monthly time lag, d to the distance of a cell. For $0 < d < 1$ no neighbouring cells exist (the nearest neighbour has a distance of $d = 1$), hence we removed these coefficients from the visualization.

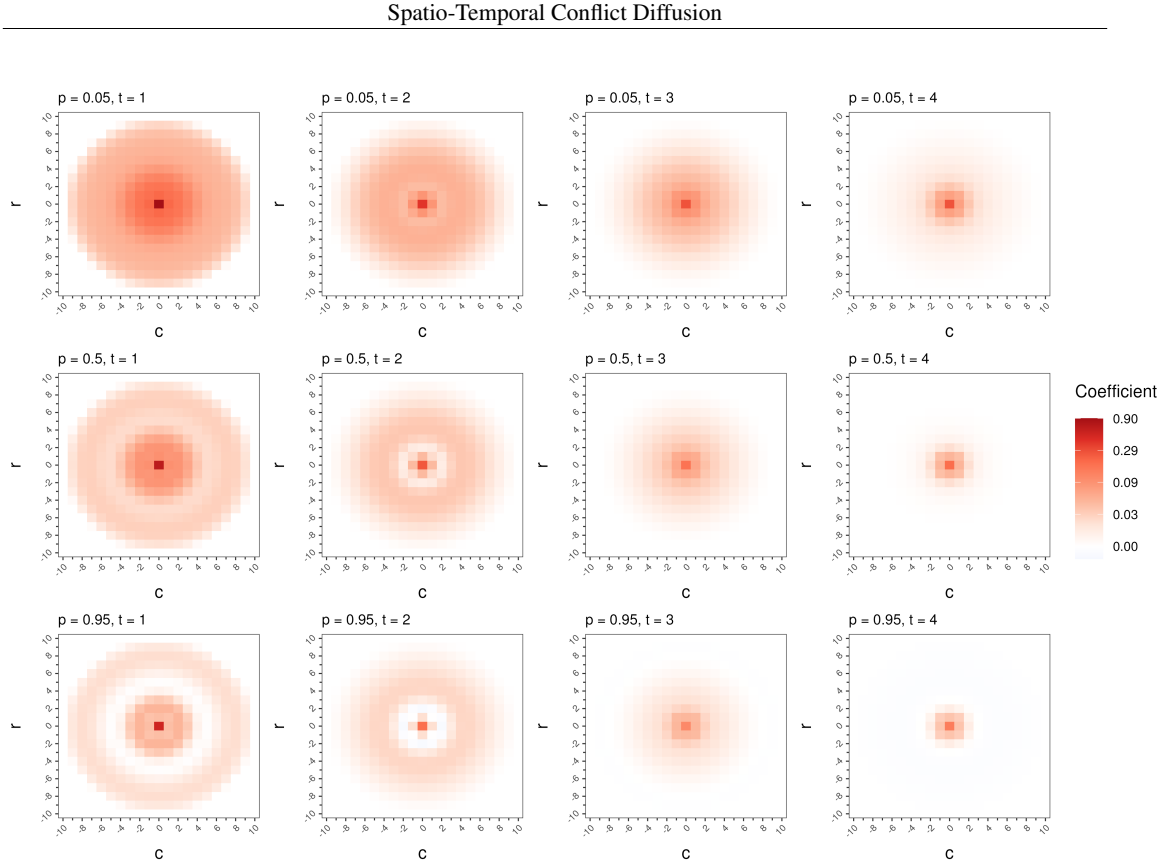


Figure 7: Conflict diffusion coefficients of combined effect (base effect + interaction) for model M2 for fixed population percentiles on a spatial grid map for the first $t = 1, 2, 3, 4$ temporal lags. r refers to the row, c to the column in the spatial grid. p refers to the respective population percentile (upper row: 5%, middle row: 50%, bottom row: 95%).

We provide an illustrative example observed in our dataset in Figure 8. It visualizes a snippet of 11x11 grid cells over the course of eight months, with the central cell located at $31.75°$ latitude and $12.25°$ longitude near the coast in the north-west of Libya. We colour each cell based on its population number percentile with respect to the overall population distribution over all cells and denote the number of logged fatalities in a cell through yellow- to red-coloured circles. We can observe that before February 2011, there were no reported conflict fatalities in this region. In February 2011 we observe the first fatalities, coinciding with the outbreak of the First Libyan Civil War (Britannica, 2024), in a cell that is among the most populous cells in this region (city of Zawiya; 96th population percentile across dataset). In the months that follow, conflict intensifies and spreads across Libya, with many of the cells affected exhibiting lower population numbers.

Returning to the diffusion effects in Figure 7, we note that the previously described ring patterns disappear for low population cells, and persist particularly for high population cells. We theorize, that these may arise from ongoing targeted violence between different ethnic groups (Mueller et al., 2022), and/or rebels and the government (Schutte, 2017). Battles would then take place between the main territories respectively headquarters (higher population) of the respective actor, which are typically located further away in distance. Hence, the higher diffusion effects for larger distances, compared to moderate distances, and thus the ring pattern. As these territories would span across multiple cells and are likely heavily contested, spatial diffusion slows down, thus the inwards movement of the ring pattern for the second monthly lag.

In Supplementary Material S.5, we additionally integrate country borders into the diffusion process, as these may arguably stop or at least slow down diffusion of conflict. However, we find that most of the discussed patterns remain unchanged, as conflict evidently also diffuses across country borders. Notably, the aforementioned ring patterns disappear for across-country (interstate) diffusion, but remain for within-country (intrastate) diffusion, which supports our theory that these may arise from continuous battles between different ethnic groups and/or rebels and the government.

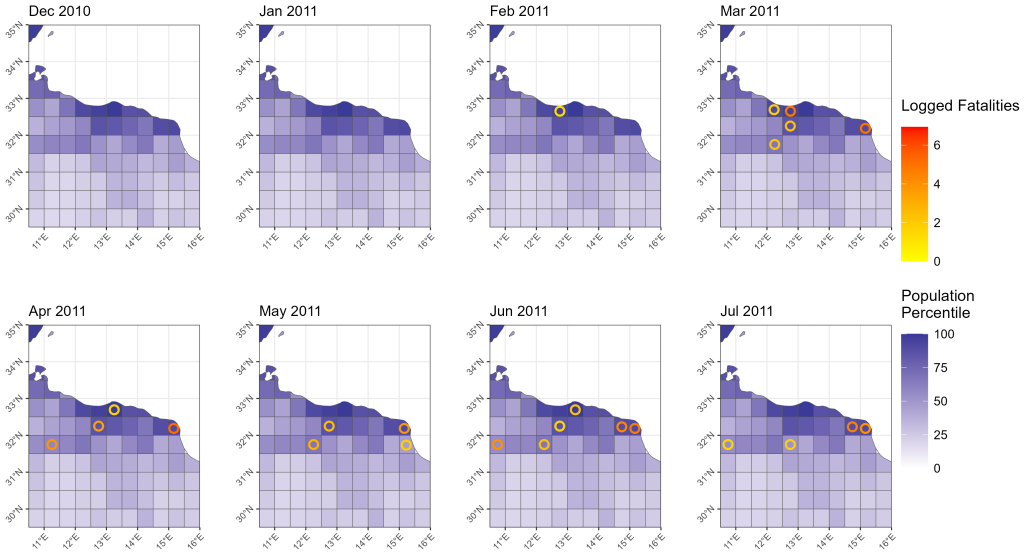


Figure 8: Illustrative data example on how conflict diffuses from high to lower population cells over time. Visualization of 11×11 grid cells in the north-west of Libya ($31.75°$ latitude and $12.25°$ longitude) from December 2010 to July 2011 (8 months). Each cell is coloured with respect to its population percentile over all observations in the dataset from white (low population) to darkblue (high population). The logged fatalities in each cell are denoted by yellow- to red-coloured circles. The white space above the grid is the Mediterranean Sea.

4.3 Robustness & Model Validation

We have conducted a multitude of checks and validation strategies to ensure that our model and its results are fully robust. We provide a detailed discussion on the estimation and the selection of our penalty parameter(s) in Supplementary Material S.3. We present an in-depth evaluation of diffusion model M1 and its residuals in Supplementary Material S.4. As expected, the model has reduced accuracy in predicting some of the rare high-fatality observations. However, we find no evidence of systematic over- or underpredictions across time or space, both in- and out-of-sample. We evaluate an alternative strategy to construct our basis, which does not rely on exponentially decaying basis functions, in Supplementary Material S.6. While we encounter the expected difficulties to achieve a smooth fit, our main findings (exponentially decaying effects; ring patterns) remain unchanged. We demonstrate that the basis in our diffusion model is sufficiently rich and its results are not sensitive to the number of chosen basis functions in Supplementary Material S.7. We reduce the maximum amount of included spatio-temporal lags in our diffusion model (smaller τ and δ) and assess the resulting performance in Supplementary Material S.12. We find this model to perform worse than our standard diffusion model, while still outperforming all reference models. We conclude that a large set of spatio-temporal lags is advantageous.

We examine the use of non-logged fatalities for both diffusion as well as reference models in Supplementary Material S.8. A clear performance gain of our diffusion model over the reference models remains. In Supplementary Material S.9, we find that a negative binomial distribution to account for possible overdispersion is not applicable in the given highly unbalanced data distribution setting. Additionally, we discuss the topic of overdispersion and possible solutions more generally. In Supplementary Material S.10, we conclude that zero-inflation is not required. Finally, we evaluate an alternative set of reference models, for which we gradually increase the included spatio-temporal lags, in Supplementary Material S.11. The results highlight that the performance gap between reference models and diffusion model does not arise from the excess information included in the latter. In fact, a reference model including the exact same amount of spatio-temporal lags as the diffusion model performs almost as poorly as base model M0-1, which does not include any information on past conflict.

5 Conclusion

In this work we proposed a regression approach that is able to capture the dynamic spreading of armed conflict in both time and space through a non-parametric smoothing component. Our proposed model is fully interpretable, generalizes well to new data points and considerably improves predictive performance compared to typical reference models both in- as well as out-of-sample. We demonstrate the flexibility of our approach for studying the transmission of conflict by investigating and identifying a relationship between diffusion and population numbers.

Our results highlight that armed conflict exhibits substantial long-lasting and far-reaching spatio-temporal diffusion. We capture diffusion up to 550km in distance and 24 months in the past, and are able to examine and study these effects more closely. We find that predictions can be improved by 24% (compared to temporal lags only) respectively 13% (compared to also including direct neighbours) using our model, which further increase when integrating population into the diffusion process. This demonstrates that existing models cannot capture the full complexity of these transmission mechanisms.

Extending our model to analyze the interaction between conflict diffusion and population, we find that conflict generally breaks out in densely populated areas, and from there diffuses across the entire region, with less populated areas being disproportionately affected. We are aware that this finding does not provide conclusive evidence for a specific mechanism driving conflict, but it offers novel insights into the dynamic nature and interdependencies of conflict as well as its close connection to population, which we believe to be valuable for future research. Additionally, in conjunction with our analysis on the impact of country borders, it allows us to demonstrate the flexibility of our approach.

As discussed in the introduction, many studies in the field do not incorporate spatial lags, i.e., information from neighboring cells, which is a notable limitation in light of our findings. This lack of control for spatio-temporal dependence can bias predictors in regression analyses and the identification of causal mechanisms (see e.g., Cook et al., 2023, for a discussion on this). This, in turn, may risk drawing inaccurate conclusions and misinform subsequent analyses and applications. Addressing this issue requires modelling approaches that fully incorporate the spatio-temporal conflict history while avoiding overfitting. Our smoothing strategy offers one such approach. We would also like to encourage researchers to evaluate their models on out-of-sample observations. As shown, even slightly more complex models can easily overfit on the noisy and highly-unbalanced conflict data.

Although our approach is designed to study armed conflict, it could similarly be applied to capture any diffusion process with long-lasting and far-reaching spatio-temporal dependencies across any type of spatio-temporal units. Notably, without using a regular lattice grid, the spatial neighbourhood definition becomes more difficult. Either, one could employ a graph-based approach using the shortest path of spatial units, or, one could utilize distances (e.g., border, centroid) to implicitly define the neighbourhood. As long as the diffusion effects are (mostly) decreasing across space and time, our approach should be able to correctly capture them. If the effects are expected to exhibit other patterns, we recommend the alternative approach described in Supplementary Material S.6.

In summary, this paper sheds light on some of the complex diffusion dynamics of conflict, the importance of controlling for its spatio-temporal dependence and offers a solution to integrate conflict diffusion into regression models. We believe that this approach provides a valuable framework for future research, both for exploring the determinants of conflict and for understanding its patterns of diffusion. Our work can be expanded in several directions. First, our model can be extended to account for different diffusion patterns across different types of conflicts, for example through a hidden Markov model. Instead, one could also fit the model separately to specific countries, regions and/or years, or include space-time interaction effects. Second, as already discussed in our analysis on country borders, actor-specific variables, such as information on politically relevant ethnic groups (Mueller et al., 2022), can be integrated in a dyadic fashion into the model to study specific mechanisms. Third, comprehensive analyses on the diffusion effects of other variables could investigate how other factors, such as political protests (Weidmann and Rød, 2019; Rød and Weidmann, 2023), contribute to the spread of conflict. Fourth, existing work studying causal mechanisms such as the effect of droughts (Von Uexkull et al., 2016; Maconga, 2023) or climate change (Selby et al., 2017; Ge et al., 2022) on conflict can be replicated while more thoroughly controlling for the dependence of conflict. Naturally, all armed conflicts are context-specific and unique, thus models will never be able to capture their full complexity. However, incorporating spatio-temporal dependencies represents a first step towards more robust and realistic analyses.

6 Funding

This work is supported by the Helmholtz Association under the joint research school ”Munich School for Data Science - MUDS”.

7 Data Availability Statement

Replication material is available at https://osf.io/ypwuv/.

References

Abidoye, B. and M. Calì (2021). Income shocks and conflict: Evidence from nigeria. *Journal of African Economies 30*(5), 478–507.

Bagozzi, B. E., O. Koren, and B. Mukherjee (2017). Droughts, land appropriation, and rebel violence in the developing world. *The Journal of Politics 79*(3), 1057–1072.

Bazzi, S., R. A. Blair, C. Blattman, O. Dube, M. Gudgeon, and R. Peck (2022). The promise and pitfalls of conflict prediction: evidence from colombia and indonesia. *Review of Economics and Statistics 104*(4), 764–779.

Blattman, C. and E. Miguel (2010). Civil war. *Journal of Economic literature 48*(1), 3–57.

Box, G. E., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Brandt, P. T., V. D'Orazio, L. Khan, Y.-F. Li, J. Osorio, and M. Sianan (2022). Conflict forecasting with event data and spatio-temporal graph convolutional networks. *International Interactions 48*(4), 800–822.

Britannica, T. E. o. E. (2024). Libya Revolt of 2011. Available at `https://www.britannica.com/event/Libya-Revolt-of-2011`, Retrieved 2024-07-02.

Briz-Redón, Á. and Á. Serrano-Aroca (2020). The effect of climate on the spread of the covid-19 pandemic: A review of findings, and statistical and modelling techniques. *Progress in physical geography: Earth and Environment 44*(5), 591–604.

Brückner, M. (2010). Population size and civil conflict risk: Is there a causal link? *The Economic Journal 120*(544), 535–550.

Butt, U. M., S. Letchmunan, F. H. Hassan, M. Ali, A. Baqir, and H. H. R. Sherazi (2020). Spatio-temporal crime hotspot detection and prediction: a systematic literature review. *IEEE access 8*, 166553–166574.

Cawley, G. C. and N. L. Talbot (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research 11*, 2079–2107.

Chadefaux, T. (2022). A shape-based approach to conflict forecasting. *International Interactions 48*(4), 633–648.

Christianson, R. B., R. M. Pollyea, and R. B. Gramacy (2023). Traditional kriging versus modern gaussian processes for large-scale mining data. *Statistical Analysis and Data Mining: The ASA Data Science Journal 16*(5), 488–506.

Cook, S. J., J. C. Hays, and R. J. Franzese (2023). Stadl up! the spatiotemporal autoregressive distributed lag model for tscs data analysis. *American Political Science Review 117*(1), 59–79.

Daley, D. J. and D. Vere-Jones (2006). *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer Science & Business Media.

Dunn, P. K., G. K. Smyth, et al. (2018). *Generalized linear models with examples in R*, Volume 53. Springer.

D'Orazio, V. and Y. Lin (2022). Forecasting conflict in africa with automated machine learning systems. *International Interactions 48*(4), 714–738.

Fitzpatrick, D. J., W. L. Gorr, and D. B. Neill (2019). Keeping score: Predictive analytics in policing. *Annual Review of Criminology 2*, 473–491.

Fritz, C., M. Mehrl, P. W. Thurner, and G. Kauermann (2022). The role of governmental weapons procurements in forecasting monthly fatalities in intrastate conflicts: A semiparametric hierarchical hurdle model. *International Interactions 48*(4), 778–799.

Ge, Q., M. Hao, F. Ding, D. Jiang, J. Scheffran, D. Helman, and T. Ide (2022). Modelling armed conflict risk under climate change with machine learning and time-series data. *Nature communications 13*(1), 2839.

Goyal, S. (2023). *Networks: An economics approach*. MIT Press.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika 58*(1), 83–90.

Hegre, H. (2014). Democracy and armed conflict. *Journal of Peace Research 51*(2), 159–172.

Hegre, H., M. Allansson, M. Basedau, M. Colaresi, M. Croicu, H. Fjelde, F. Hoyles, L. Hultman, S. Högbladh, R. Jansen, et al. (2019). Views: A political violence early-warning system. *Journal of peace research 56*(2), 155–174.

Hegre, H., N. W. Metternich, H. M. Nygård, and J. Wucherpfennig (2017). Introduction: Forecasting in peace research. *Journal of Peace Research 54*(2), 113–124.

Hegre, H., H. M. Nygård, and P. Landsverk (2021). Can we predict armed conflict? how the first 9 years of published forecasts stand up to reality. *International Studies Quarterly 65*(3), 660–668.

Hegre, H., P. Vesco, and M. Colaresi (2022). Lessons from an escalation prediction competition. *International Interactions 48*(4), 521–554.

Helmstetter, A. and M. J. Werner (2014). Adaptive smoothing of seismicity in time, space, and magnitude for time-dependent earthquake forecasts for california. *Bulletin of the Seismological Society of America 104*(2), 809–822.

Herbst, J. (2014). *States and power in Africa: Comparative lessons in authority and control*, Volume 149. Princeton University Press.

Ives, B. and J. S. Lewis (2020). From rallies to riots: Why some protests become violent. *Journal of Conflict Resolution 64*(5), 958–986.

Jun, M. and S. Cook (2022). Flexible multivariate spatio-temporal hawkes process models of terrorism. *arXiv preprint arXiv:2202.12346*.

Koren, O. and B. E. Bagozzi (2017). Living off the land: The connection between cropland, food security, and violence against civilians. *Journal of Peace Research 54*(3), 351–364.

Kounadi, O., A. Ristea, A. Araujo, and M. Leitner (2020). A systematic review on spatial crime forecasting. *Crime science 9*, 1–22.

Lee, D.-J. and M. Durbán (2011). P-spline anova-type interaction models for spatio-temporal smoothing. *Statistical modelling 11*(1), 49–69.

LeSage, J. and R. K. Pace (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

Lindholm, A., J. Hendriks, A. Wills, and T. B. Schön (2022). Predicting political violence using a state-space model. *International Interactions 48*(4), 759–777.

Maconga, C. W. (2023). Arid fields where conflict grows: How drought drives extremist violence in sub-saharan africa. *World Development Perspectives 29*, 100472.

Marshall, M. G., T. R. Gurr, and K. Jaggers (2017). Polity iv project manual. *Polity IV Project*. Retrieved 2024-06-24.

McGuirk, E. F. and N. Nunn (2024). Transhumant pastoralism, climate change, and conflict in africa. *Review of Economic Studies*, rdae027.

Meyer, S., J. Elias, and M. Höhle (2012). A space–time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics 68*(2), 607–616.

Meyer, S. and L. Held (2017). Incorporating social contact data in spatio-temporal models for infectious disease spread. *Biostatistics 18*(2), 338–351.

Miscouridou, X., S. Bhatt, G. Mohler, S. Flaxman, and S. Mishra (2023). Cox-hawkes: doubly stochastic spatiotemporal poisson processes. *Transactions on Machine Learning Research*.

Mohler, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting 30*(3), 491–497.

Møller, J. and R. Waagepetersen (2017). Some recent developments in statistics for spatial point patterns. *Annual Review of Statistics and Its Application 4*(1), 317–342.

Mueller, H. and C. Rauh (2022). Using past violence and current news to predict changes in violence. *International Interactions 48*(4), 579–596.

Mueller, H., D. Rohner, and D. Schönholzer (2022). Ethnic violence across space. *The Economic Journal 132*(642), 709–740.

Mullainathan, S. and J. Spiess (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.

Mur, J. and A. Angulo (2006). The spatial durbin model and the common factor tests. *Spatial Economic Analysis 1*(2), 207–226.

Petropoulos, F., D. Apiletti, V. Assimakopoulos, M. Z. Babai, D. K. Barrow, S. B. Taieb, C. Bergmeir, R. J. Bessa, J. Bijak, J. E. Boylan, et al. (2022). Forecasting: theory and practice. *International Journal of Forecasting 38*(3), 705–871.

Pinckney, J. (2018). *When civil resistance succeeds: building democracy after popular nonviolent uprisings*. International Center on Nonviolent Conflict Press.

Pinstrup-Andersen, P. and S. Shimokawa (2008). Do poverty and poor health and nutrition increase the risk of armed conflict onset? *Food Policy 33*(6), 513–520.

Racek, D., P. W. Thurner, B. I. Davidson, X. X. Zhu, and G. Kauermann (2024). Conflict forecasting using remote sensing data: An application to the syrian civil war. *International Journal of Forecasting 40*(1), 373–391.

Radford, B. J. (2022). High resolution conflict forecasting with spatial convolutions and long short-term memory. *International Interactions 48*(4), 739–758.

Raleigh, C. and H. Hegre (2009). Population size, concentration, and civil war. a geographically disaggregated analysis. *Political geography 28*(4), 224–238.

Raleigh, C., r. Linke, H. Hegre, and J. Karlsen (2010). Introducing acled: An armed conflict location and event dataset. *Journal of peace research 47*(5), 651–660.

Reinhart, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science 33*(3), 299–318.

Reinhart, A. and J. Greenhouse (2018). Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *Journal of the Royal Statistical Society Series C: Applied Statistics 67*(5), 1305–1329.

Rød, E. G. and N. B. Weidmann (2023). From bad to worse? how protest can foster armed conflict in autocracies. *Political Geography 103*, 102891.

Schoenberg, F. P., M. Hoffmann, and R. J. Harrigan (2019). A recursive point process model for infectious diseases. *Annals of the Institute of Statistical Mathematics 71*, 1271–1287.

Schon, J., B. Koehnlein, and O. Koren (2023). The need for willingness and opportunity: analyzing where and when environmental variability influences conflict in the sahel. *Population and Environment 45*(1), 2.

Schutte, S. (2017). Regions at risk: predicting conflict zones in african insurgencies. *Political Science Research and Methods 5*(3), 447–465.

Schutte, S. and N. B. Weidmann (2011). Diffusion patterns of violence in civil wars. *Political Geography 30*(3), 143–152.

Schvitz, G., L. Girardin, S. Rüegger, N. B. Weidmann, L.-E. Cederman, and K. S. Gleditsch (2022). Mapping the international system, 1886-2019: the cshapes 2.0 dataset. *Journal of Conflict Resolution 66*(1), 144–161.

Selby, J., O. S. Dahi, C. Fröhlich, and M. Hulme (2017). Climate change and the syrian civil war revisited. *Political Geography 60*, 232–244.

Sundberg, R. and E. Melander (2013). Introducing the ucdp georeferenced event dataset. *Journal of peace research 50*(4), 523–532.

Tapsoba, A. (2023). The cost of fear: Impact of violence risk on child health during conflict. *Journal of Development Economics 160*, 102975.

Tatem, A. J. (2017). Worldpop, open data for spatial demography. *Scientific data 4*(1), 1–4.

Toft, M. D. (2010). *The geography of ethnic violence: Identity, interests, and the indivisibility of territory*. Princeton University Press.

Tollefsen, A. F., H. Strand, and H. Buhaug (2012). Prio-grid: A unified spatial data structure. *Journal of Peace Research 49*(2), 363–374.

Vesco, P., H. Hegre, M. Colaresi, R. B. Jansen, A. Lo, G. Reisch, and N. B. Weidmann (2022). United they stand: Findings from an escalation prediction competition. *International Interactions 48*(4), 860–896.

Von Uexkull, N., M. Croicu, H. Fjelde, and H. Buhaug (2016). Civil conflict sensitivity to growing-season drought. *Proceedings of the National Academy of Sciences 113*(44), 12391–12396.

Weidmann, N. B. and E. G. Rød (2019, 10). Internet Coverage and the Spatial Diffusion of Protest. In *The Internet and Political Protest in Autocracies*. Oxford University Press.

Weidmann, N. B. and M. D. Ward (2010). Predicting conflict in space and time. *Journal of Conflict Resolution 54*(6), 883–901.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology 65*(1), 95–114.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

World Bank (2024). World development indicators 2024. Retrieved 2024-06-24.

Zammit-Mangion, A. and N. Cressie (2021). Frk: An r package for spatial and spatio-temporal prediction with large datasets. *Journal of Statistical Software 98*, 1–48.

Zammit-Mangion, A., M. Dewar, V. Kadirkamanathan, and G. Sanguinetti (2012). Point process modelling of the afghan war diary. *Proceedings of the National Academy of Sciences 109*(31), 12414–12419.

Zhukov, Y. M. (2012). Roads and the diffusion of insurgent violence: The logistics of conflict in russia's north caucasus. *Political Geography 31*(3), 144–156.

Contributing Publications

Racek, D., Thurner, P. W., Davidson, B. I., Zhu, X. X., and Kauermann, G. (2024). Conflict forecasting using remote sensing data: An application to the Syrian civil war. *International Journal of Forecasting, 40(1):373–391.* https://doi.org/10.1016/j.ijforecast.2023.04.001.

Racek, D., Zhang, Q., Thurner, P., Zhu, X. X., and Kauermann, G. (2025). Unsupervised Detection of Building Destruction during War from Publicly Available Radar Satellite Imagery. *Center for Open Science, (No. 86t3g_v2).* https://doi.org/10.31219/osf.io/86t3g_v2.

Racek, D., Davidson, B. I., Thurner, P. W., Zhu, X. X., and Kauermann, G. (2024). The Russian war in Ukraine increased Ukrainian language use on social media. *Communications Psychology, 2(1), 1.* https://doi.org/10.1038/s44271-023-00045-6.

Racek, D., Thurner, P., and Kauermann, G. (2025). Capturing the Spatio-Temporal Diffusion Effects of Armed Conflict: A Non-parametric Smoothing Approach. *Center for Open Science, (No. q59dr_v2).* https://doi.org/10.31219/osf.io/q59dr_v2. Accepted at *Journal of the Royal Statistical Society Series A: Statistics in Society.*

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 26.05.2025 Daniel Racek