AUGUST 2019



COMPARING CONFLICT DATA

SIMILARITIES AND DIFFERENCES ACROSS CONFLICT DATASETS

COMPARING CONFLICT DATA:
SIMILARITIES AND DIFFERENCES ACROSS CONFLICT DATASETS

By Prof. Clionadh Raleigh and Dr. Roudabeh Kishi

TABLE OF CONTENTS

I. Introduction 3

A. Data Collection 4

1. Inclusion criteria for researcher-led collections 4
2. Inclusion criteria for machine-led collections 8

B. Event Types, Frequency, and Geography 15

C. Sourcing 22

D. Useability 27

II. Conclusions 29

Annex: Comparison Table 32

I. INTRODUCTION

Conflict data users have a broad range of public options that differ with respect to coverage, depth, usability, and content. This review considers a selection of publicly available datasets which purport to cover similar forms of conflict activity. We specifically compare the inclusion criteria, methodology, and sourcing of select sets. We include here: the Global Terrorism Database (GTD); the Integrated Crisis Early Warning System (ICEWS) dataset; the Phoenix event dataset; the Global Database of Events, Language, and Tone (GDELT); and the Uppsala Conflict Data Programme Georeferenced Event Dataset (UCDP GED). The intent of this comparison is to demonstrate how the collection mandates, coding rules, and sourcing methods can result in drastically different information on political violence and interpretations of conflict. The review contrasts these variations with the relative advantages of the Armed Conflict Location & Event Data Project (ACLED).

At ACLED, we have continuously sought to update our methodology, sources, usability, and scope for our varied user community. We believe it is especially useful to illustrate the added value of using ACLED, compared to other existing sources. We are often asked to provide a summary of the key ways that using ACLED will affect analysis compared to other sources. We respect that many of our users have to work within probable scenarios, and hence must gauge the influence of data sources, different data aggregations, or coding strategies over others. To that end, we have concentrated on what, where, and how different data providers collect information; how and to what degree information is reviewed, checked, and cleaned; and how often and when data are released for use. Comments from users as to the thoroughness of this review are welcome.

Each of the data projects reviewed here is a collection of event data. To that end, their objective is to break down larger episodes of conflict or violence into constituent events by date, agent, intensity, or type. We have chosen to review sets that are human collected as well as those that are automated; some have expansive interpretations of conflict, including crime; others have a strict definition of the type of political violence they cover (e.g. 'terrorism'). Some maximize usability by allowing users to create curated and specific data sets; others are designed to be read by computer programs with little human interaction or choice. Still others are open about the sourcing of their information, in contrast to those which are opaque about how decisions are made and how information is collected. In this comparison, we look solely at armed, organized conflict events, and remove references to events that occur outside of that criterion.

Regardless of how data are collected, all sets purport to cover political violence and should therefore have a high degree of overlap. Each conflict dataset's mandate suggests that the *mode* of collection differs, rather than the catchment of events.[1] The types of events that define the universe (or 'catchment') of events from each set differs slightly for armed, organized conflict, as do the sources for each set, so these should account for a relatively small number of disparities. However, this review suggests that content overlap expectations are exaggerated. The definitions and the sourcing lead to extreme differences, and this is especially obvious when observing the variations in the machine-based sets.

[1] Barring GTD, which has a specific and limited mandate.

A. Data Collection

How a dataset is collected is perhaps the key difference across conflict datasets. This initial decision determines which information is privileged, how that information is reviewed, and what is accounted for. Some datasets, like ACLED, use manual coding by human researchers, while others rely on automated machine coding entirely. ACLED relies on a team of human researchers based around the world, each with specific local context and language knowledge. These researchers review thousands of sources in over 20 different languages on a regular basis. As over half a million events are now collected in ACLED, millions of articles, reports, local exchanges, and other source materials have been reviewed. Researchers individually read over each report, inputting relevant information around the type of disorder, actors involved, geographic location, source information, and reported fatalities. Coded information goes through three rounds of review each week prior to release — with reviews also involving regional research managers as well as senior staff — checking for inter-coder reliability, intra-coder reliability, and inter-code reliability to allow for cross-context comparability. Duplicate events — those involving the same groups on the same day in the same location — are reviewed and aggregated as necessary; reports of the same event are aggregated together to ensure all relevant information about an event is included. It is important to note that for datasets like ACLED that rely on researchers, some degree of automation may still be used as it can help to increase efficiency. ACLED uses a quasi automated process to crawl through media reports, which account for a significant component of its collection. However, automation is particularly poor for media sources that exist outside of media aggregators, or are in languages other than English or French. For that reason, automated systems tend to prioritize both English sources and national sources, rather than multi-language, local and regional media sources, radio sources, etc. In ACLED, regardless of the multiple systems to source information (as covered below), researchers, reviewers, and experts have the final say in the coding of events, and these practices ensure that the coding process remains carefully overseen.

UCDP, for example, relies on a team of human researchers who read each reports and, from there, input information on events that contains information on organized violence. Duplication is dealt with manually, and approximately 10,000-12,000 events are coded annually. Similarly, GTD too relies on a team of human researchers, splitting events across six separate teams specializing in specific domains based on location, perpetrators, targets, weapons and tactics, casualties and consequences, and general information. Each element of the event is coded by coders with expertise in that domain area. As noted by the program, "for example, the perpetrator domain team will have greater familiarity with active perpetrator organizations, their naming conventions, aliases, spelling variations, factions, and splinter organizations, making them well-suited to systematically record information on the organizations attributed responsibility for an attack."

Machine-based coding platforms (Phoenix, ICEWS, and GDELT as explored here) are collections of events based entirely on automated coding without intervening researchers.

1. Inclusion Criteria for Researcher-Led Collections

ACLED is specifically designed to capture armed, organized political violence and demonstrations as they occur in all countries. Our inclusion criteria for a political violence event is that an episode must

have involved at least one armed, organized group with a political objective.[2] Political objectives can include — but are not limited to — motivations to replace an agent or system of government; the protection, harassment, or repression of identity groups, the general population, and political groups/organizations; the destruction of safe, secure, public spaces; elite competition; contests to wield authority over an area or community, etc. This means that ACLED covers activity in all countries,[3] and captures the presence of armed, organized groups engaging in extrajudicial and/or other forms of political violence. Armed, organized groups include state forces, rebels, political, government and community militias, and external state security forces. Further, ACLED casts a wide net in order to capture demonstrations, allowing any reported riot or protest of more than three people to be included, regardless of the media reported cause of the action. Demonstrations are not considered to be political violence, but a political act. These definitions are important to clarify as **an act of political violence or a demonstration is the sole basis for inclusion within the ACLED dataset.**

Therefore, in ACLED, all acts of political violence and demonstrations are included — those committed by governments, rebels, militias, or others as well as those perpetrated upon other armed groups, the government, civilians, citizens, and demonstrators. However, other data projects have alternative inclusion criteria. In GTD, the inclusion criteria is explicitly related to whether the event was 'terrorism'. In GTD, an event must be "an intentional act of violence or threat of violence by a non-state actor" in addition to two of the three following criteria: "(1) the violent act was aimed at attaining a political, economic, religious, or social goal; (2) the violent act included evidence of an intention to coerce, intimidate, or convey some other message to a larger audience (or audiences) other than the immediate victims; or (3) the violent act was outside the precepts of International Humanitarian Law." An act of terrorism by a non state actor is not the same as capturing the activity of a terrorist group. For example, events involving the Islamic State that conform to terrorism would be included by GTD, but not all acts of violence by the Islamic State are necessarily included[4] because they are not all terrorist acts. GTD does not have other inclusion criteria, and so it operates as a clearly defined, and well-established subset of data on a wide range of activities and groups.

In another human-coded dataset, inclusion criteria can be arbitrary and unsystematically applied. The UCDP GED set is based on events which occur within a predefined contest. The inclusion criteria of this dataset is twofold, and strongly based on whether a fatality was recorded, rather than whether an act of political violence occurred. To begin, armed force must be used by one organized actor against another, or against civilians. Second, the event must result in at least one direct death at a specific location-date for it to be recorded. The dataset has a dyad focus; a conflict dyad is only coded once it crosses the 25 battle-related deaths threshold in a given year.

[2] Or in the cases of active conflict, an unidentified group is presumed to have political intentions when their acts of violence are against state forces, infrastructure, citizens, or another armed, organized group.
[3] ACLED has collected data, and established data programs, for all areas. However, in some cases, such as North America and Australia, these are not yet public but are in pilot form.
[4] Indeed, the definition of terrorism most closely relates to 'violence against civilians' in the ACLED set. This is not a direct overlap; IS suicide bombing against government troops would be included as an event, whereas that would not be catagorized as violence against civilians by ACLED.

"Who is considered a relevant and legitimate actor in conflict is pre-determined by the mandate of the dataset … The rules of inclusion determine how and in which detail we understand violence."

This serves to create several inconsistencies in how the UCDP-GED represents reality. The first is the arbitrary cutoff of at least 25 deaths; this threshold tends to suppress the events of common militias operating across most states (e.g. during elections, elite competition, in urban environments, etc.). It, by design, captures only the events of larger, organized groups who have set similar agendas. This cutoff therefore does not include events perpetrated by actors who purposefully do not kill many individuals yet terrorize populations (such as early Boko Haram activity, or that of present-day militias across Kenya). It also obfuscates those actions perpetrated by 'Unidentified Armed Groups'[5] — a group category that is more often the result of the strategy to be anonymous than a lack of detailed reporting.[6] Further, whole countries with persistent but low-level conflict are missing in certain years of coverage. For example, Madagascar — which held elections in 2018 with reports of election-related violence, such as attacks on poll workers — does not appear in the 2018 subset of UCDP GED. In South Africa, a number of ANC affiliates have been targeted and killed throughout the year, but such activity also does not appear in the 2018 subset of UCDP GED. This violence is political and carried out by anonymous groups, results in limited fatalities, and is a growing and pervasive threat across developing states. But it is not captured by the UCDP dataset. In other cases, countries where conflicts have emerged which do not fit into the criteria designed for civil wars of decades past (e.g. El Salvador or Honduras) fail to appear in the catalogue of UCDP events.[7] Yet, conflict events occurring in South Sudan, Brazil, and Turkey are only partially represented in UCDP. These problems will be magnified in countries or subnational areas where media reporting on the details of conflict is not consistent or reliable, including poorly developed peripheral areas (e.g. Congolese conflict areas), areas with poor media or government access (e.g. parts of Syria, Somalia, and Afghanistan) or those with high levels of violence against journalists (e.g. parts of the Philippines and Myanmar). All these instances, and more, bias a dataset whose inclusion criteria depends on an arbitrary fatality threshold. For example in Turkey during 2018, UCDP does not include any police repression in the west, while only capturing violence involving the Kurdistan Workers' Party (PKK) in the east; as a result, ACLED's reported 2018 Turkish fatality numbers are four times those recorded by UCDP-GED and the event total is nearly five times as many.[8]

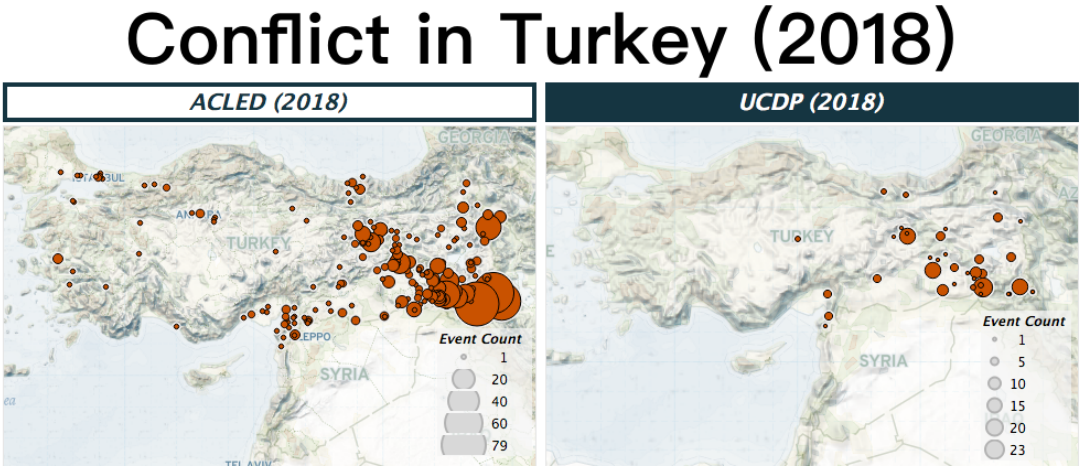
[5] While UCDP notes their coding of informally organized groups — groups without an announced name, but who use armed force against another similarly organized group, where the violent activity indicates a clear pattern of violent incidents that are connected — their coding suggests that these groups can never engage with anyone other than other informally organized groups, which is not reflective of reality.

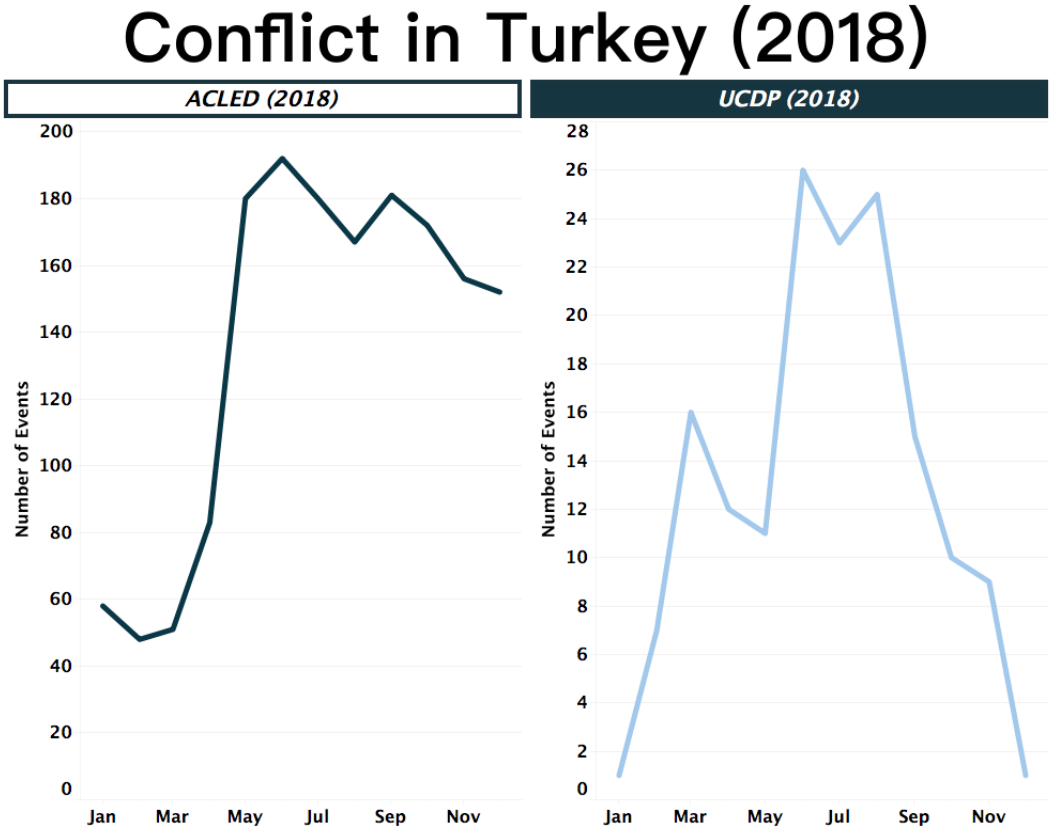
[6] Kishi, Roudabeh. (9 April 2015). "The Strategic Use of Unidentified Armed Groups In Conflict Zones." Armed Conflict Location & Event Data Project (ACLED). <http://www.crisis.acleddata.com/the-strategic-use-of-unidentified-armed-groups-in-conflict-zones/>.

[7] El Salvador and Honduras both yield extremely high numbers of homicides and gang violence in Latin America; this violence is very public, with gangs controlling certain areas, suggesting a challenge to the power of the state. As such, violence involving gangs in these countries is considered political within ACLED methodology (ACLED will be releasing Latin America data by the end of 2019). Conflict is not recorded in either country in the UCDP GED dataset for 2018.

[8] This is counting only organized violence in the ACLED dataset; if including total event count, ACLED's event count is almost ten times as many as UCDP's.

Perhaps most worrying is that the composition and distribution of events is distorted by poor representation. Consider the following maps and graphs of Turkish events in 2018:





The landscapes and patterns appearing here are drastically different despite covering the same period and same forms of conflict. The patterns are distorted by definitional constraints on the right (UCDP GED), and an accurate interpretation of political violence by state and non-state forces on the left (ACLED). The magnitude, locations, intensity, geography, and duration of events are markedly different. In short,

using the dataset on the right would present a view and pattern of Turkish conflict that is not in keeping with reality.

These variations in definitions and scope are important as a dataset will include and categorize an event of political violence only if it falls within its catchment. Who is considered a relevant and legitimate actor in conflict is pre-determined by the mandate of the dataset; the definitions, catchment, and categorization are critical, as they tell a user who and what is likely to be included. In turn, the rules of inclusion determine how and in which detail we understand violence. This can be as simple as the number of events each dataset includes for the same conflict, to the composition of those events and resulting interpretation of that conflict by the user community. Inclusion criteria should allow for accurate representations of political violence, while being flexible to how political violence has changed.

Conflict has changed considerably since the end of the Cold War, and now includes a far higher rate of activity by non-state actors against the state and civilians, and higher rates of state actions against civilians and non-state groups who are *not seeking to replace* the regime. Militias — non-state, armed, organized, politically violent agents — are widespread, existing in contexts as different as India, the United States, and Malawi. These types of agents are particularly important as they can show how different political violence data account for modern conflict. For example, the figure below maps disorder in India in 2018. UCDP, for example, limits disorder in India to violence in the Jammu & Kashmir region and the Red Corridor where Naxalite-Maoists have been active for years, and to a lesser degree the northeast; it largely misses, however, the presence of violence anywhere else in the country. This low-level violence involving mobs, castes, and vigilantes is often the primary form in which violence manifests in India, yet this threat is unaccounted for by the dataset. This is likely a result of the limited lethality of this violence as well as the agents engaged in this violence. **Definitions matter, as do strict and systematic inclusion criteria, as these impact how we understand the boundaries of a conflict and whether they reflect current trends.**

"Definitions matter, as do strict and systematic inclusion criteria, as these impact how we understand the boundaries of a conflict and whether they reflect current trends."

2. INCLUSION CRITERIA FOR MACHINE-LED COLLECTIONS

Machine-based data projects including GDELT, ICEWS, and Phoenix have a wider remit based on an automated search term rather than a bounded definition of political violence. Little documentation exists about the definitions that determine which events are searched for. Keywords are highlighted and articles pulled from within a predefined set of sources, and those potential events are then automatically categorized by the Goldstein scale (as the basis for the Conflict and Mediation Event Observations [CAMEO] coding scheme). Rather than the inclusion criteria creating a potential set of events that are then reviewed according to a methodology and an oversight schedule, machine sets have a wide remit, and little to no oversight. This means that the inclusion criteria become the de facto methodology of the set: the boundaries set at this stage determine the follow-on machine commands for inclusion. This leads to a considerable amount of crime and irrelevant events (e.g. sports 'battles') being included in automated sets (see below for further details).

This also leads to significant problems of false positives (incorrectly included events), inflated numbers due to duplicates, and incorrect patterns due to a lack of review.

False Positives

False positives are a serious problem for machine sets in particular, as they inflate or distort media stories, leading to a conflict event being coded where none has occurred. Consider the following examples, one from each major set:

In **Phoenix,** an event coded on 3 December 2018 suggests conflict between Zimbabwe and the US; the actual report[9] is about a US tourist falling victim to a hippopotamus attack at Victoria Falls. Such coding suggests Zimbabwe's involvement in a number of international conflicts within the span of a year, which is entirely false.

In **GDELT**, a report from Agence France Presse titled "Cows come home to haunt India's Modi"[10] describes state protection of cows in India, given their status under Hinduism. This event is coded as 'conventional military force' involving the government, despite the fact that the report does not involve actions by Indian military forces at all. Similarly, an article about Jammu Martyr's Day,[11] commemorating the 1947 massacre in Jammu & Kashmir, was coded as 'engaging in mass violence' in 2018 between Pakistan and an unidentified army in India within the GDELT dataset. This suggests that mass violence is occurring in a contentious region, which in turn can have quite negative ramifications in analysis using these data.

ICEWS reports that four acts of "abduction, hijacking or hostage" occurred in 2018 in Zimbabwe. One event is recorded in Chikore, Zimbabwe on 24 October 2018 against "Air Zimbabwe." Chikore is not a location in Zimbabwe, and 'Air Zimbabwe' or anyone associated with this business was not hijacked or abducted on this day. Mercury is a Durban-based newspaper, and the story was not retrievable from this source. However, this larger issue refers to how Mr. Simba Chikore, former President Mugabe's son-in-law, used corrupt tactics during his tenure as head of this state enterprise.[12] This story was recorded by ICEWS as a '-9' level event, which is the designation for 'severe repression' and 'assault.' Other examples of abductions as reported by ICEWS include acts against a former Government minister (who was arrested rather than abducted), a UK tourist (without evidence that this occurred), and a 'citizen' on 6 August 2018.[13]

---

[9] News 24. (3 December 2018). "US tourist injured in Vic Falls hippo attack." WN. <https://article.wn.com/view/2018/12/03/US_tourist_injured_in_Vic_Falls_hippo_attack/>.

[10] Agence France Presse (AFP). (5 April 2019). "Cows come home to haunt India's Modi." Radio France Internationale (RFI). <http://en.rfi.fr/wire/20190405-cows-come-home-haunt-indias-modi>.

[11] Kashmir Media Service. (5 November 2018). "Jammu Martyr's Day tomorrow." Kashmir Media Service. <https://kmsnews.org/news/2018/11/05/jammu-martyrs-day-tomorrow/>.

[12] Zindoga, Tichaona. (11 March 2019). "Zim Airways: The untold story." The Herald. <https://www.herald.co.zw/zim-airways-the-untold-story/>.

[13] Additional examples: An event coded on 6 December 2018 is coded as the most serious form of violence, due to an apparent misreading of an article title: "Zimbabwe Battles Influx of Congolese, Mozambique Refugees." This is a false positive. It is not an attack, assault, or serious act of political violence. The original report does not concern violence. In a more extreme example, an event coded on 7 February 2018 notes violence between Afghanistan and Zimbabwe; the event, however, is coded from a report originally published in the Frontier Post — a Pakistani

Perhaps more worrying is that these examples of false positives are easily identifiable, but still remain in sets despite being clearly incorrect. This indicates that no reviews or corrections are occurring, at any stage of the collection. Consider this example from 6 June 2019 which suggests that a conflict event — on par with a nuclear conflict (-10 on the CAMEO scale) — occurred in the United States between the US government and Iran, in Washington, DC (event ID number 34342006) with the designation "fight with artillery and tanks." In fact, Iran had engaged in a "war of words" without any conflict or threats with the US.[14] This still remains in the set after three months, indicating that at no point does a member of ICEWS review the veracity of the coding.

"Rather than the inclusion criteria creating a potential set of events that are then reviewed according to a methodology and an oversight schedule, machine sets have a wide remit, and little to no oversight ...The inclusion criteria become the de-facto methodology of the set."

Many of the examples we have noted above are from low-level conflicts, where aberrations may be more apparent. We also considered events in countries that are in active war, under the assumption that a higher rate of true positives are likely in those cases given the scope and definitions across the datasets. Yemen was one of the worst cases of political violence in 2018, and we hence assume coverage of such a crisis will be consistent across datasets. Yet coding by Phoenix returned a high level of incorrectly coded events. Coding of an event on 30 January 2018 suggests the involvement of China as an actor in the conflict, yet no aspect of the original report[15] makes any mention of the country. This may be a mistake, but as it is automated it is unclear why an unmentioned country would be coded at all. In another event on 5 December 2018, Sweden is coded as engaging in conflict (CAMEO code of -10) with Yemen; coding seems to be solely the result of the UN-brokered peace consultations discussed in the report[16] being held in Sweden; no violence occurred. While these false positives serve to incorrectly internationalize the war, other events have the opposite effect. An event on 13 June 2018 is coded as occurring in Yemen with both the source and target actors also coded as Yemeni; yet, the original report[17] is about Saudi-led forces attacking the Yemeni port city of Hodeidah. This example serves to domesticate a conflict that is highly internationalized. Between the false domestication coupled with false internationalization, it becomes quite difficult to decipher the war in Yemen. Examples from GDELT have been closely covered by

newspaper — about a cricket match between the two countries. This is another example of a false positive, but also a relatively absurd scenario where Afghanistan and Zimbabwe are 'in conflict.' These common false positives have the ability to skew trends greatly. These distortions are important because both false positives and skewed reports indicate that the number and the information from machine coding is incorrect.

[14] Havasi, Amir. (25 June 2019). "Iran vows to ditch more nuclear curbs in war of words with US." Agence France Presse (AFP). <https://news.yahoo.com/iran-warns-us-attack-imperil-interests-across-mideast-000622035.html?guccounter=1>.

[15] Reuters. (30 January 2018). "Yemen separatists capture Aden, govt confined to presidential palace." WN. <https://article.wn.com/view/2018/01/30/Yemen_separatists_capture_Aden_govt_confined_to_presidential/>.

[16] Dettmer, Jamie. (5 December 2018). "Yemen's Warring Sides Agree to Prisoner Swap Ahead of Talks." Voice of America (VOA). <https://www.voanews.com/a/yemen-warring-sides-agree-to-prisoner-swap-ahead-of-talks/4688099.html>.

[17] Gambrell, Jon. (12 June 2018). "Saudi-led forces begin attack on Yemen port city." The Star. <https://www.thestar.com/news/world/2018/06/12/saudi-led-forces-begin-attack-on-yemen-port-city.html>.

others,[18] while ICEWS rarely provides enough information to assess whether the events reported are accurate or real.

Inflated numbers

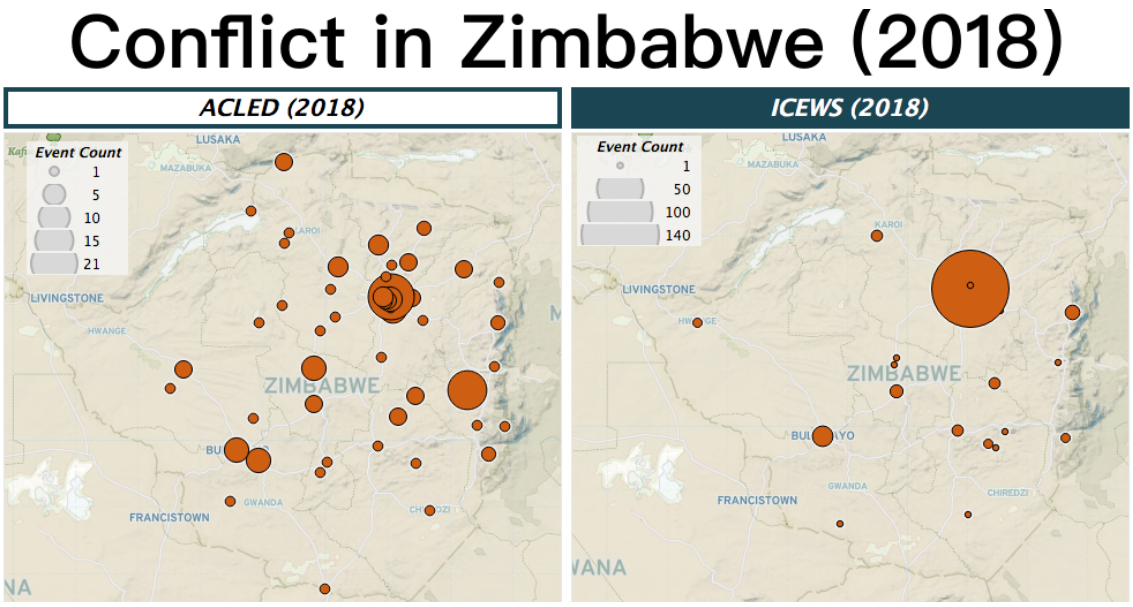
Fully automated datasets come with some advantages in the form of speed and (perceived) efficiency, as removing the role of researchers speeds up the processing of data. ICEWS and Phoenix are able to publish data daily as a result, and GDELT publishes new data every 15 minutes. However, without oversight, this automation can produce a great deal of noise in the form of irrelevant events and inaccurately coded events. Duplicate events are commonplace. While "duplicate stories from the same publisher, same headline, and same date are generally omitted, ...duplicate stories across multiple publishers (as is the case in syndicated new stories) are permitted" within ICEWS coding. Phoenix, meanwhile, admits that "...articles and event-bearing sentences were not de-duplicated. This could inflate event counts in some cases."[19] GDELT's duplication checking can be even more flawed; GDELT has an indicator for the number of times an event was mentioned across all source documents during the 15-minute window in which it was first seen (as well as the number of sources it appeared in and the number of articles). This limits the flagging of duplicate events to only a 15-minute time period around the first mention of the event, so future articles that mention that event will not be flagged. When pulling data from the GDELT API, a user can ask it to drop events that it views as duplicates (not an exact match, but similar events). Given the way in which news stories break and develop, this can have a significant impact on event count. Events that make international headlines will get updated repeatedly as more and more details become available in the aftermath of the event, but if this does not occur within co-occurring 15-minute segments surrounding the original breaking news, the updates will not be flagged. This means such large events will be duplicated repeatedly as the story is updated and picked up by various outlets.

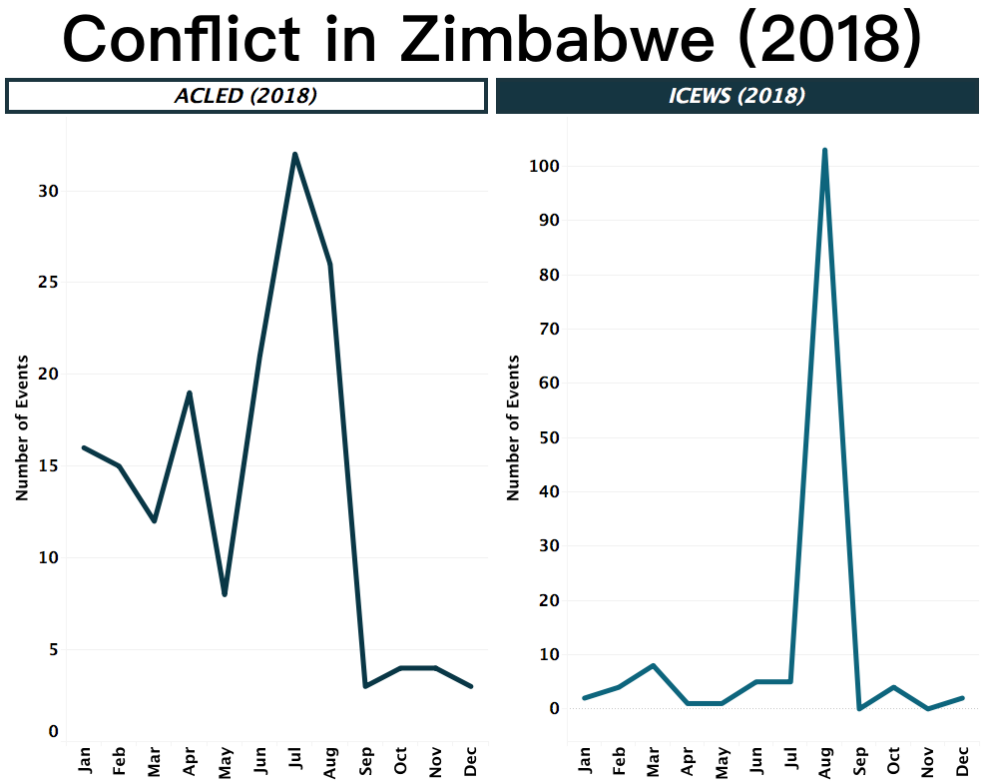
Let us return to the earlier example of 'inflation' of extreme events in reference to Zimbabwe, which is a case where violence is low enough to discern clear trend differences in conflict events. ICEWS records 214 events in an 'armed conflict only' version of the dataset: 107 '-10' acts which include 'assaults and/or nuclear weapons / engage in mass violence;' 19 '-9.5' acts referring to 'Impose blockade, restrict movement and engage in mass expulsion;' and 88 '-9' acts, including 'abductions, unconventional violence and sexual assault.' In one example of the latter, former MDC leader Thokozani Khupe is

[18] In 2014, in the aftermath of the highly publicized Chibok girls abduction in Nigeria, the popular FiveThirtyEight blog published a piece on the sudden, stark rise in kidnappings in Nigeria — using GDELT data. Soon, they were forced to issue a correction noting how "[GDELT] is a repository of media reports, not discrete events. As such, we should only have referred to 'media reports of kidnappings,' not kidnappings." Going on they note, "This mistake led to other problems [in the analysis]" -- affecting both the trends as well as geography of these events. They underline that "while GDELT makes an effort to remove duplicate media reports of the same event, it is not always successful in doing so because media reports often conflict with one another. There were many conflicting reports about the mass kidnapping of hundreds of girls from the Government Girls Secondary School in Chibok in April. This likely accounts for at least some of the rise in media reports of kidnappings." (For more, see the original story: https://fivethirtyeight.com/features/mapping-kidnappings-in-nigeria/). This is one of many cases of issues with use of GDELT data being described. See also pieces in the Political Violence at a Glance blog (https://politicalviolenceataglance.org/2014/02/20/raining-on-the-parade-some-cautions-regarding-the-global-database-of-events-language-and-tone-dataset/) and the Source OpenNews Project (https://source.opennews.org/articles/gdelt-decontextualized-data/), amongst others.

[19] For example, two events coded on 6 December 2018 in Yemen both link to the same article — though one is routed through World News *Politics* while the other is through World News *World.*

recorded as physically assaulting "Men (Zimbabwe)" for giving a speech decrying sexual and gender harassment in the country. More to the point, in 2018, Zimbabwe had one episode of significant organized violence in the aftermath of the elections in early August and a drastic rise in protesting in February of that year. It is not an actively violent country according to several human rights organizations that operate throughout the country. Nevertheless, ICEWS, Phoenix, and GDELT code significant and widespread conflict across the country (ICEWS specifically is mapped and graphed below), suggesting that the vast majority of coding for Zimbabwe by these datasets is made of false positives. Zimbabwe is not a unique case, but one that we have chosen to investigate here. The inflated numbers are a function of international media interest in a state, rather than a reflection of its actual events; there is no reason to suspect that events and media reports are reliably correlated.





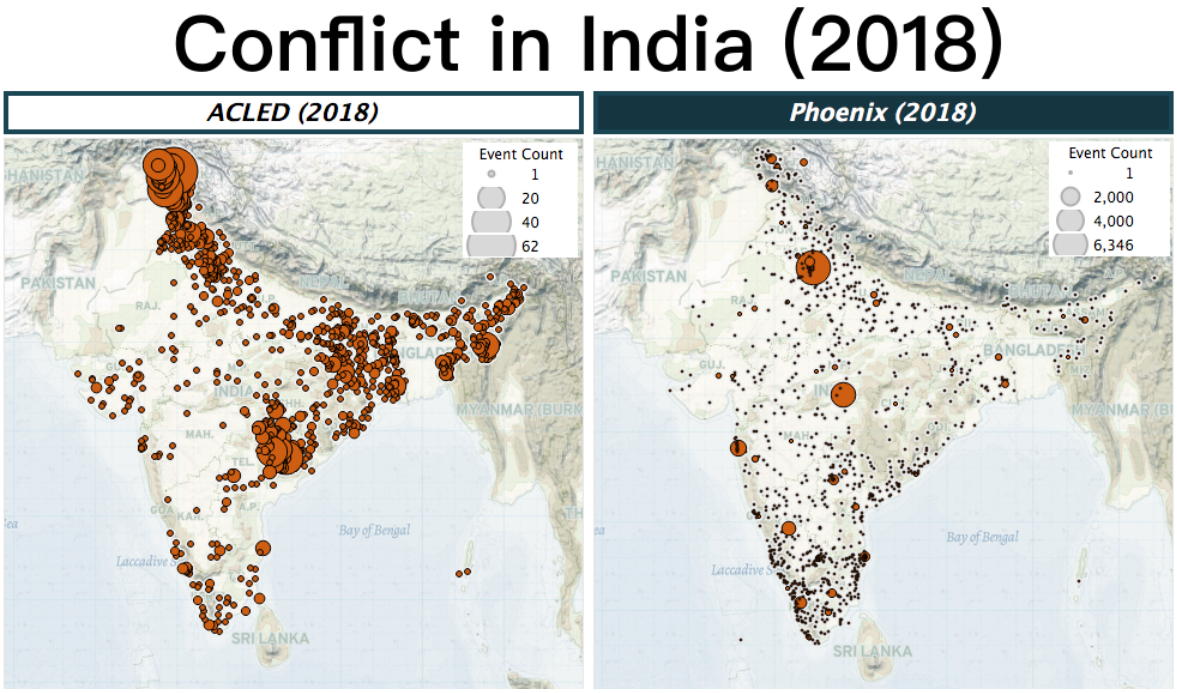
The comparison above suggests that conflict built steadily in Zimbabwe over the course of the year leading to the violence during the election in early August. However, evidence of that building tension, coupled with the extensive activity across the country, is largely absent from the machine based collection (on the right).

In short, in automated machine-based data collection systems, the number of included events is often extremely high — and this is not because human-based coding is missing large numbers of reports. With sufficiently sized research teams, human-coded data projects are fully able to collect and review the same (and often more) sources. Rather, it is because duplicate reports and stories about the same event are not removed in machine-based data collection systems, therefore resulting in numbers that can be vastly inflated.

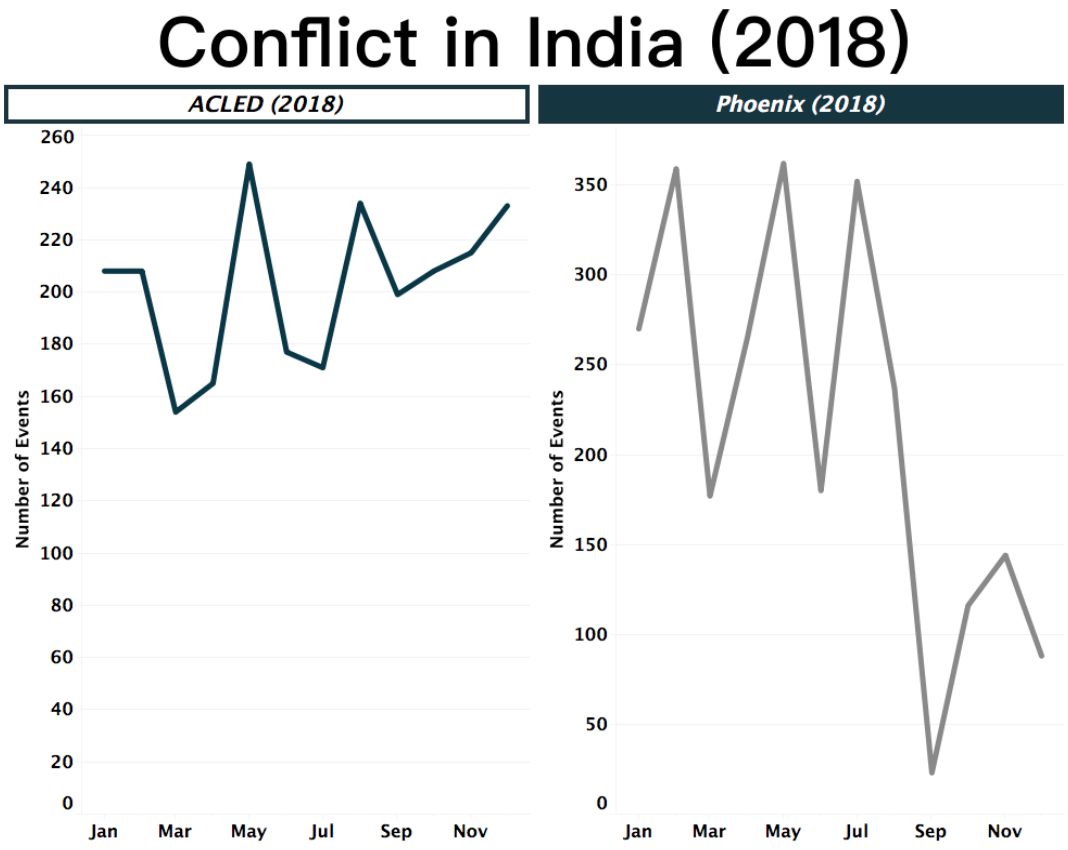
Incorrect Patterns

The 'inclusion' problem for automated machine-coded datasets like ICEWS, Phoenix, and GDELT is whether what is included allows for a valid representation of conflict. Our examples above indicate that distortions in conflict information by machine-coded sets are due to false positives or inflated events. But the overall patterns of conflict are also incorrect. Some users may believe that machine sets provide a relatively accurate 'big picture' of aggregated patterns, while getting smaller details wrong. But due to inflated numbers, incorrect geographies, false positives, and a reliance on media stories rather than information, the aggregated patterns are also likely wrong. For example, India as coded by Phoenix (mapped and graphed below) seems to miss the violence in Jammu & Kashmir region in 2018 considerably — a region the international community agrees is the hotspot for violence in the country —

while seemingly *entirely* missing violence in India's northeast — a region where the Indian state has been dealing with insurgent groups seeking independence and/or autonomy since 1950. Meanwhile, some articles with non-conflict event information are miscoded as conventional military actions, including an article[20] about a helicopter crash that killed five employees of an energy company. In the article, the CEO of the company referred to the employees as "in a way...energy soldiers," which may be why the algorithm coded the event as conventional military action.



[20] Live Mint. (15 January 2018). "ONGC reviews offshore logistics operations after chopper crash." Live Mint. <https://www.livemint.com/Industry/UY1O3GJUhajb1bR4gVTyAO/ONGC-reviews-offshore-logistics-operations-after-chopper-cra.html>.



In the India graphic comparison above, drastically different geographies and temporal patterns are evident. While the left displays a pattern over the year reflecting repeated surges correlated with elections, seasonal patterns and other consistent instigators, on the right, the temporal patterns are the opposite of that reported from domestic sources. The geographies of only armed, organized conflict are diametrically different, with the data on the right reflecting a central cluster due to limitations of accurate georeferencing. Further examples of distortion are highlighted below.

B. Event Types, Frequency, and Geography

A number of differences in datasets can stem from how events are categorized once selected for inclusion. For example, ACLED categorizes events in three ways, asking "is this event a violent episode, a demonstration, or a 'non violent' case"? From there, each event is categorized further into an event type and a sub-event type. Users can, if needed, then select specific events or sub-events. All definitions of each event and sub-event are noted in ACLED's public methodology section, along with examples and decisions for complex cases (see below).

Other datasets, such as GTD, engage with a specific event type as their limited mandate, and it is pre-determined by their inclusion criteria above. For that reason, datasets like GTD are not explored in further detail in this subsection.

Event types in the automated datasets — ICEWS, Phoenix, and GDELT — are categorized with the Conflict and Mediation Event Observations (CAMEO) system. The codebook for CAMEO includes "codes" for 300+ types of events that include both cooperation and conflict situations: they range from making a public statement or consulting, to threatening or exhibiting military posturing, to engaging in unconventional mass violence. Each event is also assigned a "Goldstein score" which represents its intensity. These scores are assigned based on the event type rather than the circumstances of the event. In other words, a riot with 10 deaths gets the same score as a riot with 100 deaths because they are both riot events. (For more on the actual application of this scoring system, see below.) These categorizations mean that a great deal of non-conflict events are captured as active conflicts; yet, a great deal of actual conflict can also be missed. For example, the figure below depicts conflict in Nigeria in 2018. While ICEWS captures the Boko Haram threat occurring in the northeast of the country, all other violence is seemingly missed across the northwest, Middle Belt, and Niger Delta regions. This is despite a spike in intercommunal violence involving Fulani militias that occurred in 2018. Phoenix, meanwhile, captures the Boko Haram threat, yet misses many of the other dynamics. Both Phoenix, and to a lesser degree ICEWS, code a number of events seemingly to the centroid of the country, producing a 'hotspot' of violence in central Nigeria that does not reflect reality. For example, an article clearly referencing an attack in the outskirts of Maiduguri in Borno State was coded by Phoenix as occurring in Abuja State in the center of the country,[21] and, in another case, an airstrike in Borno State

General	Event Type	Sub-Event Type
Violent events	Battles	*Armed clash*
		Government regains territory
		Non-state actor overtakes territory
	Explosions/Remote violence	*Chemical weapon*
		Air/drone strike
		Suicide bomb
		Shelling/artillery/missile attack
		Remote explosive/landmine/IED
		Grenade
	Violence against civilians	*Sexual violence*
		Attack
		Abduction/forced disappearance
Demonstrations	Protests	*Peaceful protest*
		Protest with intervention
		Excessive force against protesters
	Riots	*Violent demonstration*
		Mob violence
Non-violent actions	Strategic developments	*Agreement*
		Arrests
		Change to group/activity
		Disrupted weapons use
		Headquarters or base established
		Looting/property destruction
		Non-violent transfer of territory
		Other

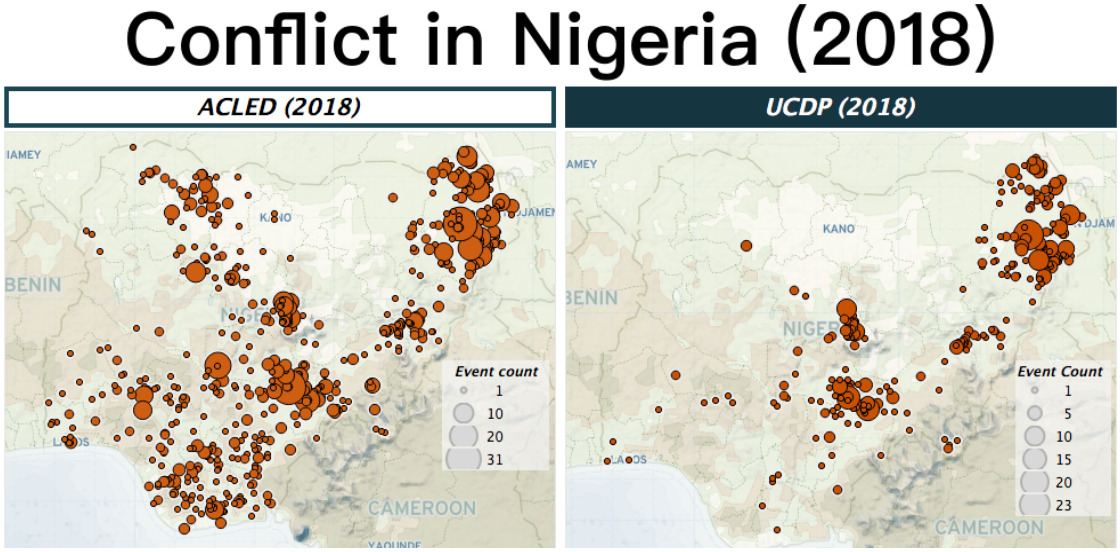
[21] APO Group. (2 April 2018). "UN Deputy Humanitarian Coordinator in Nigeria strongly condemns the continued targeting of civilians in north-east Nigeria." WN. <https://article.wn.com/view/2018/04/02/UN_Deputy_Humanitarian_Coordinator_in_Nigeria_strongly_conde/>.

was coded as occurring in Abuja State instead.[22] Similarly, a Boko Haram attack in Maiduguri was georeferenced to Abuja State by GDELT.[23]

Our conclusion is that, despite the wide variation in event types captured by the Goldstein scale used by machine sets, in practice, these automated systems distort the conflicts of cases like Nigeria. Many articles are miscategorized, and these mistakes are consistently in the direction of excessively violent designations: there is an overinflation of '-10' classifications, which is the most serious forms of conflict. The geography and composition of the events are also consistently flawed.

UCDP previously limited their earlier collections around state-based conflict — only including conflicts in which a rebel group challenged the state. More recently, UCDP has included non-state conflict defined as conflict "not involving the state," between two non-state actors, as well as one-sided violence, where an armed actor targets vulnerable populations. As conflicts are pre-defined dyadically, the dynamics and alliances within multi-party conflicts cannot be captured. Large-scale, dyadic conflicts (i.e. more conventional conflicts) are occurring less frequently than in the near past, and because more countries experience high levels of multifaceted conflict, these event forms can fail to account for ongoing dynamics. The evolution of even dyadic conflicts becomes difficult to detect as conflicts are not differentiated by time; conflicts over the same incompatibility are given the same ID regardless of the time that has passed between them, suggesting they are continuations of the same incompatibility, when in fact they might not be. Finally, smaller conflicts may be missed if the fatality threshold is not met; this results in more peripheral conflicts which do not yield mass fatalities being largely missed. For example, the figure below maps and graphs conflict in Nigeria in 2018; here, UCDP seemingly misses conflict in the northwest involving communal militias and state forces entirely, as well as violence in the Niger Delta region in the south involving non-state actors targeting civilians. Lulls in conflict are markedly lower in the temporal graph on the right, possibly due to missing entire conflicts in the northwest and southwest of the state. Nigeria's conflict patterns are remarkably cyclical when events are sourced from local and national programs and media, but are marked by inconsistencies when not done so (right).

[22] Adepegba, Adelani. (21 August 2018). "NAF neutralizes regrouping Boko Haram insurgents." Punch. <https://punchng.com/naf-neutralizes-regrouping-boko-haram-insurgents/>.
[23] Lanre, Ola and Ahmed Kingimi. (27 April 2018). "At least four killed in Boko Haram attack on Nigerian city." Reuters.
<http://news.trust.org/item/20180427092931-5qfi6>.



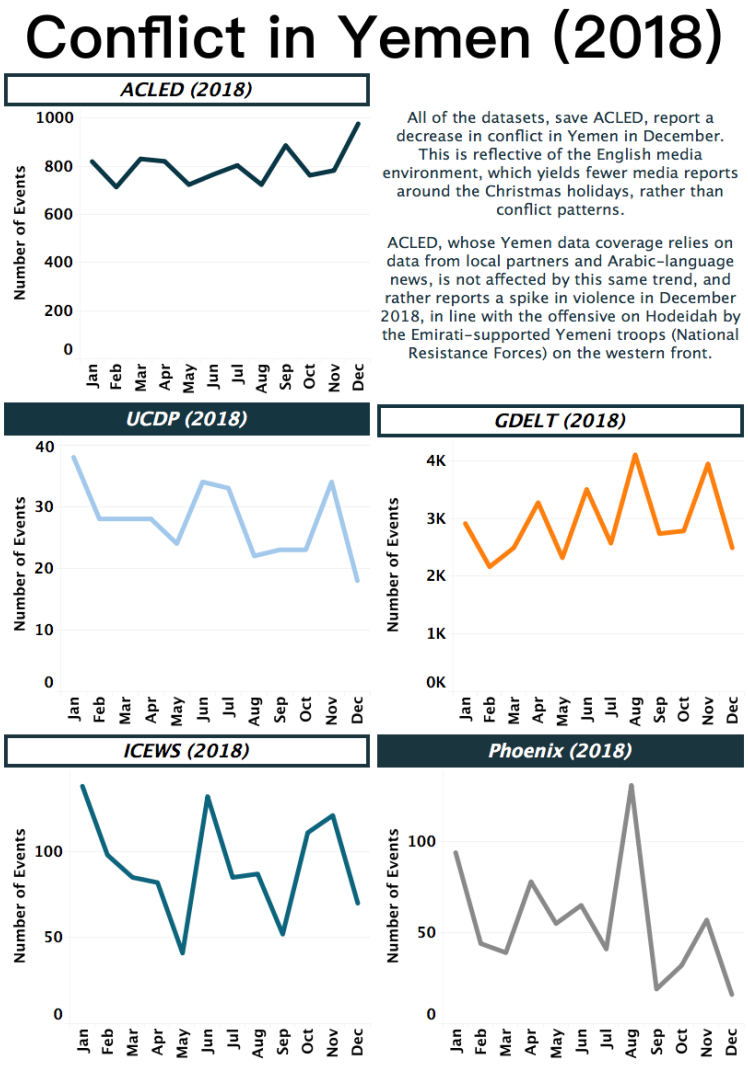


The reason that clear, coherent, and correct classifications are important for users is that conflicts are not homogenous: the events that suggest risk and instability differ in their frequency, sequences, and intensity. Event types that reflect the variation of modalities common across conflicts and periods of disorder are basic, central components of insightful and useful analysis. Without capturing the spectrum

of conflict, datasets can distort the truth about the distribution of instability and impact on civilians and state security.

> *"The reason that clear, coherent, and correct classifications are important for users is that conflicts are not homogenous: the events that suggest risk and instability differ in their frequency, sequences, and intensity."*

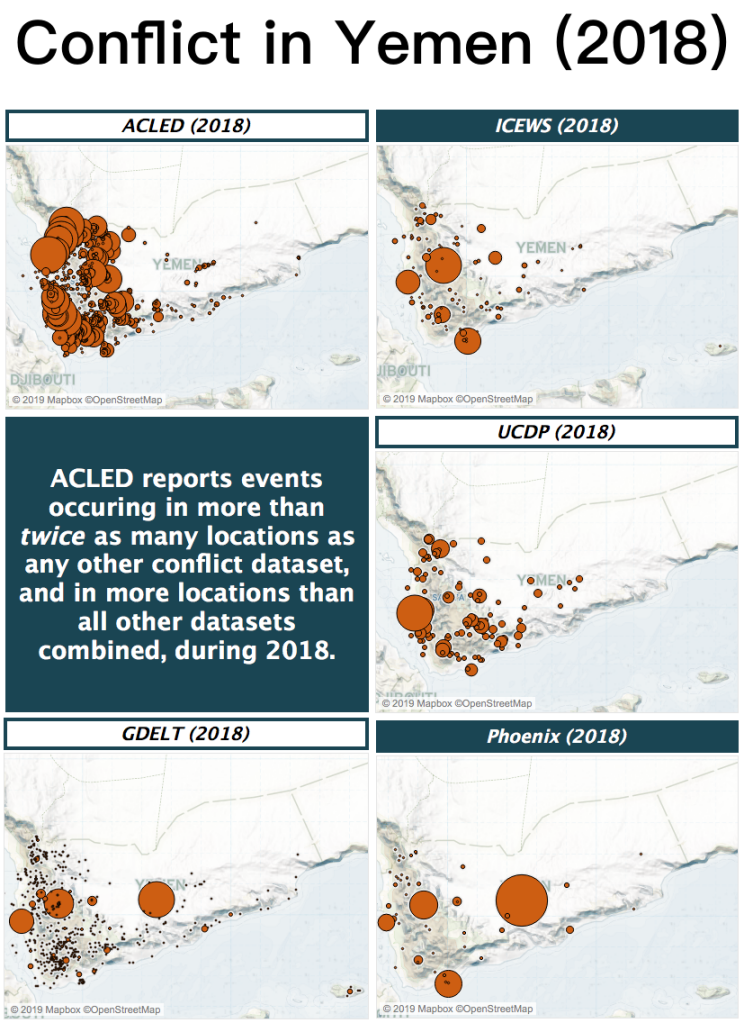
In addition to the inclusion of irrelevant events or the inaccurate coding of relevant events, there are serious distortions in the frequency, geography, and intensity of conflict based on machine coding. Further, important details as to who is involved and how different groups are interacting are missed. **Details are important. And getting the details correct — including the counts, agents, dates, locations, and type of violence — are vital to understanding conflict.** The figure below, for example, depicts the frequency, or 'pulse', of conflict in Yemen in 2018 for each dataset, and helps to underline how this frequency is not standard across them. The datasets do not agree on when spikes in violence occurred throughout the course of the year. Most damning is the general trend in the conflict as the year comes to a close: all of the datasets, save ACLED, report a decrease in conflict in Yemen in December. This is reflective of the English media environment, which tends to yield fewer media reports across all countries in December, around the Christmas holidays. Meanwhile, ACLED, whose Yemen data coverage relies on data from local partners and Arabic-language news, is not affected by this same trend, and rather reports a spike in violence in December 2018, in line with the offensive on Hodeidah by the Emirati-supported Yemeni troops (National Resistance Forces) on the western front.



The figure below maps the war in Yemen in 2018 as seen by a number of datasets;[24] the geographic scope of the war varies greatly depending on whose eyes the war is seen through. GDELT and Phoenix suggest a 'hotspot' of conflict in the center of Yemen — evidently a centroid point for events in which the datasets could not determine where in Yemen to code the event. Incidentally, that area (Shabwah) is not an active area of conflict in 2018. While ICEWS and UCDP do identify the west as the primary arena of violence in the country, ICEWS misses the presence of violence along the border with Saudi Arabia almost entirely. This is despite the severity of violence in the Sa'dah governorate — the historical homeland of the Houthi movement; in fact, US Green Berets were deployed to the border region to assist Saudi forces in targeting missile launch sites in Yemen.[25] UCDP too, while identifying more events

[24] GTD is not included in the comparison here as 2018 data are not yet available.
[25] Cooper, Helene, Thomas Gibbons-Neff, and Eric Schmitt. (3 May 2018). "Army Special Forces Secretly Help Saudis Combat Threat From Yemen Rebels." *The New York Times.* <https://www.nytimes.com/2018/05/03/us/politics/green-berets-saudi-yemen-border-houthi.html>.

occurring in the border region than ICEWS, also minimizes the strategic importance of conflict in the area.

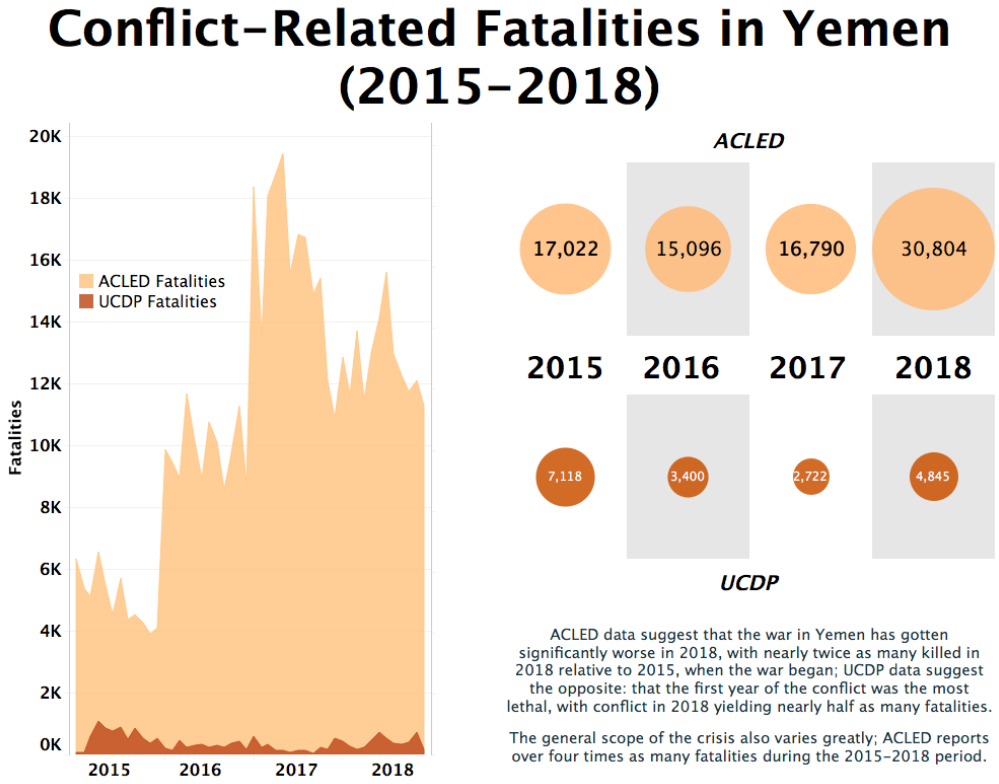


Even the intensity of conflict can vary immensely. The figure below depicts fatality counts in Yemen over the past few years, comparing ACLED to UCDP. Not only are the numbers vastly different — with ACLED reporting over six times as many fatalities in 2018, for example — the trajectories of violence also vary. ACLED data suggest that the war in Yemen worsened significantly in 2018, with nearly twice as many killed in 2018 relative to 2015, when this Saudi led coalition began its assaults in Yemen. UCDP data suggest that the first year of the conflict (2015) was the most lethal, with conflict in 2018 yielding nearly half as many fatalities as that period. The UCDP information is largely out of step with international organizations which have emphasized the intensity and complexity of this war, with the United Nations (UN) referring recently to Yemen as 'the worst humanitarian crisis in the world'.[26] In early 2019, the UN Development Programme (UNDP) commissioned a study on the conflict in Yemen that explicitly addressed the different fatality estimates provided by ACLED and UCDP:

[26] See 19 August 2019 statement from the Office of the Resident Coordinator and Humanitarian Coordinator for Yemen: https://reliefweb.int/sites/reliefweb.int/files/resources/HC%20Statement_Sanaa_19%20August.pdf

> ACLED publishes a host of methodological documents and codebooks on their website, as well as a Yemen-specific methodological overview. The methodology is like that used by the UCDP as far as triangulation of death counts, source bias consideration and filtering, etc. However, ACLED does not have the same criteria that UCDP does concerning the need to know both actors in a conflict event to code that event. Thus, a meaningful stream of conflict-related deaths is potentially left out of UCDP official estimates, due both to the traditional obstacles associated with reporting casualties in conflict zones as well as challenges that are specific to Yemen.[27]

The UNDP-commissioned study ultimately concluded that "because ACLED has less restrictive exclusion criteria than UCDP, we determined that ACLED's data are more representative of the totality of human life lost in Yemen during the conflict, from 2016 forward."[28]



C. Sourcing

How and where projects source information is possibly the most important factor in explaining why coverage and data differ. Any data coding project can only ever be as reliable and accurate as its methodology and sourcing allow.

---

[27] Moyer, Jonathan D., David Bohl, Taylor Hanna, Brendan R. Mapes, and Mickey Rafa. (23 April 2019). "Assessing the impact of war on development in Yemen." United Nations Development Programme (UNDP). <http://www.arabstates.undp.org/content/rbas/en/home/library/crisis-response0/assessing-the-impact-of-war-on-development-in-yemen-.html>.

[28] Ibid.

ACLED privileges a wide and deep net of sourcing when gathering information, which comes from: (1) traditional media, (2) select new media, (3) public and private reporting, and (4) partnerships with local conflict observatories. Traditional media includes news media ranging from national to local outlets, newspapers to radio. Thousands of media are reviewed in over 20 different local languages; English-language media is not privileged. ACLED's remote institutional structure means that researchers based around the world with specific local language skills can be hired to cover specific states in order to ensure the most relevant information is integrated. Yet traditional media alone can have limitations. Hence, select new media is also integrated; this includes, for example, the Twitter accounts of vetted journalists or activists, as well as trusted Telegram channels or WhatsApp groups. Importantly, not all social media is included given the lack of oversight and high levels of noise that entails. Information from public and private reports are also included; this includes in-depth qualitative reporting by Amnesty International, for example, as well as 'official' state-reported events, such as ceasefire violations. Lastly, yet importantly, information sourced from local conflict observatories with whom ACLED has established partnerships is integrated. Often such local organizations gather primary data, and have coverage that is most reflective of local-level realities. However, such initiatives are often limited in scope — capturing only a certain type of violence, or conflict in only a specific sub-region of a country — hence making comparability difficult. ACLED translates the information gathered by these local organizations into a standardized ACLED methodology to allow for integration into the larger ACLED dataset — thereby allowing for comparability.[29] **The information for each unique context is therefore a particular combination of these sourcing types, suited to the specific reporting environment in each state.[30]**

UCDP uses Factiva searches to yield the first tranche of their information. News aggregators like Factiva emphasize national and international news sources; as such, about 60% of UCDP's sources are from global newswires, with a strong reliance on English-language reports.[31] The use of reports originally produced in other languages is only available in cases where the aggregators have made them so. Such reliance on these types of media sources also suggests that where there are media blackouts (e.g. Burundi in 2015; Ethiopia since 2014), Factiva and, by extension, a user like UCDP, will not collect an accurate assessment of conflicts. Tools like Factiva can be helpful in bringing together numerous disparate sources, but cannot be relied upon for a complete picture of difficult, unstable, or local contexts: that is not the role they are filling in the media environment. UCDP notes that after their 'first pass', a project leader or UCDP coder may elect to conduct a 'second pass' only if detail about the conflict is seen as insufficient; this 'second pass' is said to include sources such as "local monitoring of various local media," "local monitoring and research organizations," "global NGO reports," as well as reports from international organizations, "governmental publications," and "research articles or books." However, no information (name, use, accessibility, bias, coverage, etc.) is provided on any of these sources. This list of 'in depth, possible second-pass' source forms is similar to ACLED's primary source materials; this means that ACLED's consistent, weekly source checks are much broader, deeper and wider than others as a standard practice. If similar sources are used for cases such as Ethiopia, then similar data should be collected. However, ACLED collects far more material on actors that both ACLED and UCDP collect,

[29] ACLED has numerous partnerships with local conflict observatories based in countries around the world. Partnerships often involve data sharing between ACLED and the partner organization. Additionally, ACLED supports capacity building for partners around data collection methodology, analysis, and visualization — offering trainings around these points both remotely and in the field.

[30] For more on this, see this Clingendael report written by ACLED staff: https://www.clingendael.org/sites/default/files/2018-01/Report_Conflict_environments_and_coverage.pdf

[31] See page 12 of the UCDP GED Codebook: https://ucdp.uu.se/downloads/ged/ged191.pdf

and on the same events both claim to collect. Suffice to say, conflict environments require flexible, in-depth, expansive, and adaptable event sourcing techniques, which ACLED has developed for every state it operates in.[32]

GTD, meanwhile, relies upon "media articles and electronic news archives, and to a lesser extent, existing data sets, secondary source materials such as books and journals, and legal documents" to gather information. Similar to UCDP's use of Factiva, GTD relies upon Lexis Nexis to identify articles for review; as such, the articles are in English primarily, and foriegn language articles are made available only when translated by the aggregator. Similar to ACLED, they too make use of Natural Language Processing and machine learning to identify relevant articles and to remove duplicates before articles are then reviewed by coders.

Both GTD and UCDP GED rely on media sources, yet do not share the details of the source for each event so the user does not know where information for each event comes from. ACLED has extensively documented its sourcing arrangements and plans; has written extensively about these details as they apply to several difficult sourcing scenarios, such as in Turkey and Tunisia;[33] and notes the source of every event. ICEWS, Phoenix, and GDELT also note the source of information for each event (discussed in further detail below).

"Conflict environments require flexible, in-depth, expansive, and adaptable event sourcing techniques, which ACLED has developed for every state it operates in."

Similar to UCDP, ICEWS sources information via Factiva, as well as through Open Source Enterprise. It relies upon "...about 6000 sources...many of these are aggregators of hundreds of other sources" (Ward et al, 2013) in English, Spanish, Portuguese, and Arabic. GDELT notes that they rely on "hundreds of thousands of broadcast, print, and online news sources from every corner of the globe in more than 100 languages and its source list grows daily." Without a consistent source list, the inclusion of new sources can introduce artificial changes to trends in the data; this issue does not seem accounted for within GDELT. Similar to ICEWS, only information from traditional media is used, meaning the coverage is susceptible to the limitations of such media (e.g. prioritizing large-scale, 'sensationalized' events, missing peripheral events or events involving smaller groups, etc.). Phoenix has the most limited sourcing of the automated datasets; they note their reliance on the New York Times ("based on its status as the national paper of record and as an authoritative source for events occurring in the US or involving the US government") as well as the BBC Summary of World Broadcasts (SWB) and the CIA Foreign Broadcast Information Service (FBIS) ("[their] content was chosen to ensure broad international coverage—these are two open-source information projects generated by government agencies aggregating many foreign-language sources and using highest-quality human translators, covering thousands of broadcast, print, and Internet news outlets"). As a result of this sourcing, information comes in English (or has been translated already into English). Given this very limited sourcing net, there is much Phoenix is likely to miss in its coverage.

The inclusion of events from a range of agents is only possible where sourcing is dedicated to capturing all reports — specifically emphasizing local partnerships and local data and reporting. A reliance on traditional media, especially English-based media, across the datasets means that information

[32] For more on this, see this Clingendael report written by ACLED staff: https://www.clingendael.org/sites/default/files/2018-01/Report_Conflict_environments_and_coverage.pdf
[33] Ibid.

around coverage of certain types of violence are privileged over others. Small groups, and events that are not sensationalized or of interest beyond the country, are extremely unlikely to be reported by large national services. Media sources from international and national sources are far more likely to report on high-fatality events, sensationalized events, and those involving large groups with an international profile (e.g. ISIS, Boko Haram). In some cases, national newspapers and official state reports are more likely to underreport or mis-report government activity (e.g. Afghanistan). For these reasons, an extensive array of sourcing forms — partnerships, media, and private reports — coupled with a dedication to relaying the extent and form of all political violence, is the best solution.

"The inclusion of events from a range of agents is only possible where sourcing is dedicated to capturing all reports — specifically emphasizing local partnerships and local data and reporting."

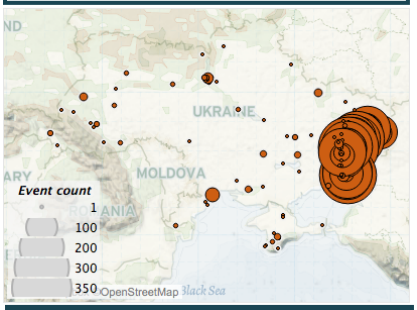
For example, the figure below maps and graphs conflict in the Ukraine in 2018; the distinction in sourcing across the datasets and its impact on the conflict landscape is evident. The automated datasets — ICEWS, Phoenix, and GDELT — present Kiev, the capital, as a hotspot of conflict, despite conflict in the capital being quite limited relative to elsewhere in the country. GDELT and Phoenix also suggest a hotspot in the center of the country — a result of coding events at the country-level — despite limited conflict actually occurring in that region given its limited strategic importance. The Ukraine conflict through the eyes of these datasets suggests that conflict in the Donbass region is a relatively smaller threat than violence in Kiev or in the center of the country; this is a distortion of reality.[34] UCDP, meanwhile, limits conflict in the Ukraine to *only* the Donbass region in the east with no other conflict reported anywhere else in the country. This is despite reports of explosions in Kharkiv and violence against civilians in Odessa, amongst other incidents. ACLED reports conflict across the country, with the vast majority of events occurring in the Donbass region. ACLED's use of Russian-, Ukrainian-, and English-language sources in the country — unlike the other datasets noted here — contributes to its deeper coverage of conflict in the country. In addition, reports from the Organization for Security and Cooperation in Europe (OSCE) are crucial for coverage of the hundreds of ceasefire violations that are reported weekly across the 'no man's land' that separates Ukraine and Russia.[35] Given the frequency of these violations, not all are covered by traditional media every week. A non-media source like the OSCE reports must be used to capture these trends.

[34] BBC. (18 February 2015). "Ukraine crisis in maps." BBC. <https://www.bbc.com/news/world-europe-27308526>.

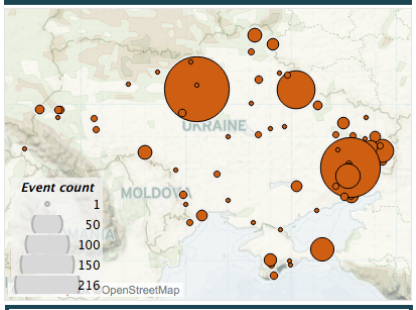
[35] ACLED offers an extensive methodological explainer around how we translate these OSCE reports into ACLED methodology: https://www.acleddata.com/wp-content/uploads/dlm_uploads/2019/04/ACLED-Ukraine-Methodology-Brief_4.2019.pdf.

Conflict in Ukraine (2018)

ACLED (2018)

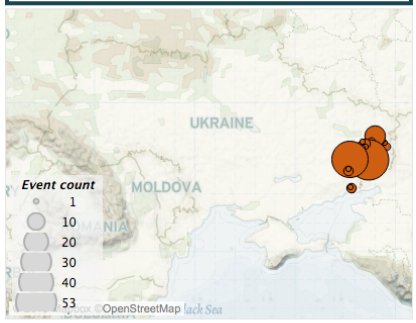


ICEWS (2018)

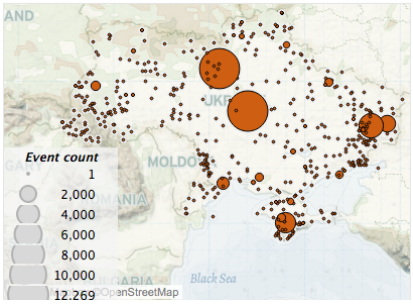


In 2018, ACLED presents the Donbass region as a hotspot of conflict, yet also reports events occuring elsewhere in the country as well.

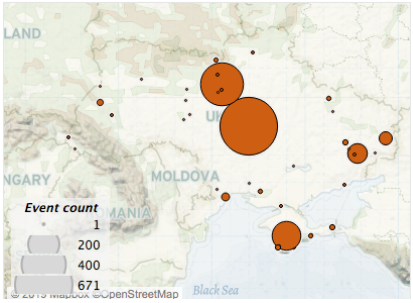
UCDP (2018)

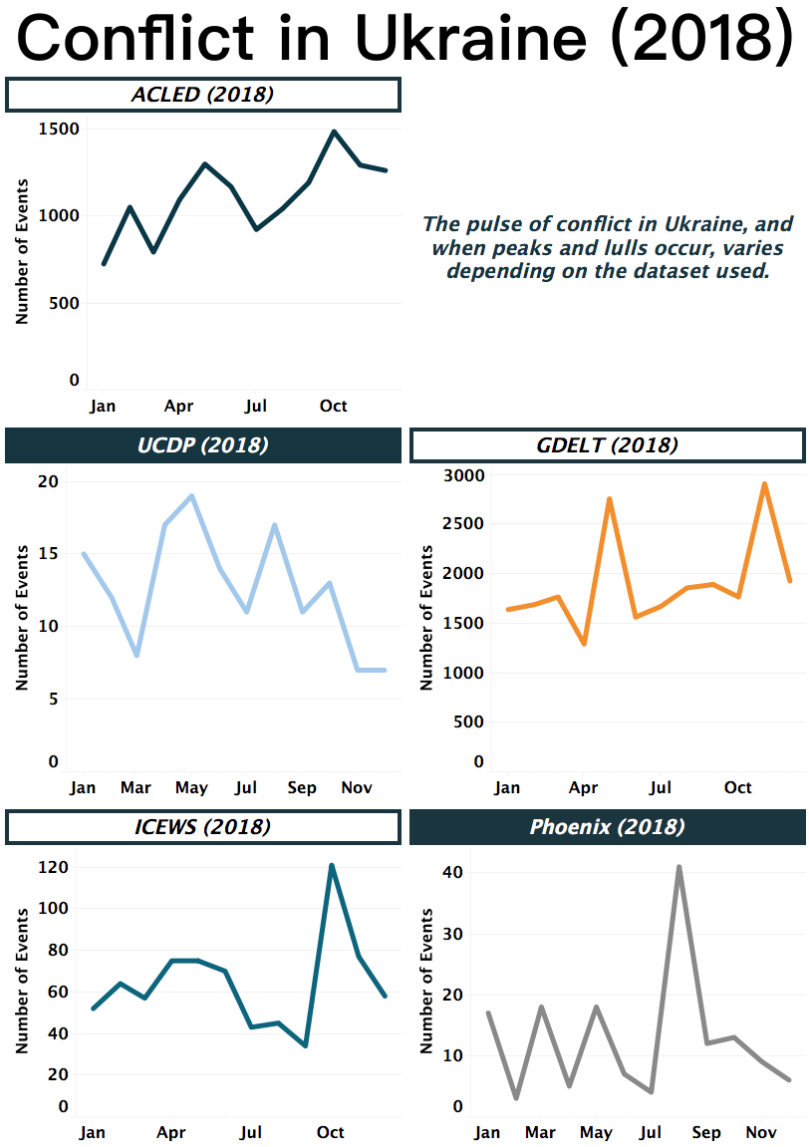


GDELT (2018)



Phoenix (2018)





D. USEABILITY

Datasets must be useful and useable if they are to be relied upon for regular analysis. Release schedules, updating/correction processes, ease of download and access to data, and methodological transparency, are all crucial components in making a dataset useable.

Many of the datasets reviewed have intermittent coding schedules and release schedules. ICEWS, Phoenix, and GDELT have machine coding running constantly. With limited oversight by researchers built into their release schedule, new data are released[36] on a daily system for ICEWS and Phoenix and a 15-minute interval system for GDELT. ACLED releases new data on a weekly schedule, with new data

[36] Tuesday releases for Asia are due to geographical and time differences in review schedules for managers and reviewers.

published every Monday and Tuesday. UCDP GED and GTD release annual datasets once a year on a variable schedule (i.e. an annual release at some point in the following or subsequent years).

In addition to releasing new data, datasets can alter or amend previously published data when new information becomes available. This is especially crucial for near real-time data (ACLED and the machine-coded sets), as in the immediate aftermath of an event, reports might get some of the details of an event wrong. In other cases, the information may change in the aftermath of an event: a group may claim responsibility for an attack, or fatality counts might increase as injuries become deaths. In other cases, the UN Security Council (or similar investigative bodies) may release a report after in-depth inquiry into trends, or another secondary account might contribute new information surrounding events previously unknown or unreported. These kinds of details are crucial to account for and such reporting timelines are common in event data coverage. As new information comes to light and previously published information is corrected, it is imperative to alter and amend as necessary. However, the process of including such corrections and updates is only clearly specified in ACLED.[37]

Ease of downloading and access to data is a 'frontline' experience for users. ACLED, GTD, UCDP, and GDELT all have dedicated websites for their data projects that clearly delineate where and how users can download data and codebooks/documentation. Both ACLED and UCDP also provide dashboard features to allow users to work with the data without having to download files if they choose not to. GDELT allows users to visualize their data using the GDELT Analysis Service or with Google BigQuery.[38] Downloading and working with the data can prove quite difficult for the average user given the sheer size of files if coverage is global or continuously updated. For these reasons, ACLED allows users to download all the data, or a section based on country, event, actor name, geography, or intensity.

Phoenix and ICEWS can be a bit more difficult to access. A search for Phoenix data, for example, leads one to a UT Dallas page labeled as "Real Time Event Data / Phoenix", though with no link to actually download the data or see documentation. The website notes "Phoenix_RT: Real-time data from Oct. 2017 to today, please see more details at the Open Event Data Alliance." Upon heading to the linked page, there is *still* no access to the data. Within the text on the page, there can eventually be found a sentence that says "Maintenance of a set of reliable, internet-based open access, multi-sourced social and political event data sets that are updated on at least a daily or weekly basis." Following this link leads to a page with a number of links to specific file names flagged as 'test files'; the user is left wondering which file(s) to download making it difficult to know when data are updated and when the link on the web page may have last been. The most straightforward way to download Phoenix events is to download from their API directly, which — despite their written directions — can be difficult and non-intuitive for users without previous experience working with APIs. A search for ICEWS leads one to a Harvard dataverse page. There, access to 'automated event data' or 'coded event data' is available (neither the difference nor the method of selection is particularly clear). Clicking on the first leads one to a page with over 200

[37] ACLED details the review process in an explanatory document hosted on the website: https://www.acleddata.com/wp-content/uploads/2017/12/Coding-Review-Process_FINAL.pdf

[38] But it must be noted that it was extremely difficult to use GDELT: For example, researchers at ACLED had to download 61 separate files in order to be able to append together a full file covering India in 2018 alone, even when limiting the download to only conflict and demonstration events.

files, which are again difficult to know which ought to be downloaded. For both Phoenix and ICEWS, the user must be quite familiar with the internal structure of the dataset and webpage in order to be able to find the data they need; this makes the barrier to entry for new users quite high.

Once data are accessed, it is important for a dataset's methodology to be clear and transparent. Not every user will require fine-grained details about how decisions are made, but any user should be able to find or ask for those details if she requires them. This allows users to know exactly what assumptions, and caveats are necessary when working with the data, which in turn impacts how findings from the data ought to be interpreted. Similar to the obscurity around how to access the data, transparency behind both the Phoenix and ICEWS datasets is quite limited. While Phoenix provides access to a single documentation file, drafted in 2014, little sourcing information is offered. Phoenix explains their reasoning for the sources they use, but do not go into detail about how (if at all) the sources aggregated by the SWB and FBIS are vetted. ICEWS similarly mentions where their sources come from, but does not provide links to the sources used for each coded event, leaving the user with no way to verify the accuracy of the coding. In short, both ICEWS and Phoenix do not publicize in-depth codebooks, definitions, coding criteria, coding method, or any other details (including explanations and summaries of variables). GDELT does provide access to a single documentation page on their website. Both GTD and UCDP do provide codebooks with the key components of their methodology laid out. For the latter, while UCDP GED has some information publicly available on various components of their coding structure, there is little provided on the justification, definitions, or details of their collection and sourcing process.

ACLED believes that the user should be able to access every detail of how conflict data are coded and collected. Decisions made by the research team relating to violence must be consistent with known rules, and such rules should be explained in terms of how they create clear, consistent, relevant, and robust conflict data. To fulfill this principle, ACLED provides open access to methodology documents on a range of significant and specific issues, in addition to the general dataset codebook outlining methodology decisions. These additional methodology documents help users to understand the specific ways in which general ACLED methodology may be applied to contentious contexts or circumstances. They includes dsicussions of how ceasefire violations reported by the OSCE are coded in Ukraine; how ACLED is dealing with the various obstacles around coverage of the conflict in Yemen; how ACLED codes drug-related violence in the Philippines given the nature of reporting around such events; how various factions of Boko Haram are coded in Africa; and how sexual violence is coded and analyzed across the dataset, to name a few.

II. CONCLUSIONS

Not all conflict datasets offer equal levels of coverage, depth, useability, and content. A review of the inclusion criteria, methodology, and sourcing of leading publicly available conflict datasets — GTD, ICEWS, Phoenix, GDELT, UCDP GED, and ACLED — demonstrates that there are significant discrepancies in the output produced by ostensibly similar projects. Rather than finding substantial overlap between datasets, the review identifies a number of important gaps left by deficiencies across core criteria for effective conflict data collection and analysis, listed below. ACLED — through its unique combination of innovative and dynamic methodology, human oversight and expertise, and extensive source network — is the only dataset to consistently meet these key standards.

Data Collection and Oversight: *A rigorous, human coder is the best way to ensure reliable, consistent, and accurate events that are not false positives.* ACLED and other human-coded datasets generate fewer

false positives and maintain oversight mechanisms to review potential events according to an established methodology. In ACLED — regardless of the multiple systems used to source information — researchers, reviewers, and experts have the final say in the coding of events, and these practices ensure that the coding process remains carefully regulated and updated as necessary. Automated event data projects are still being refined and are not yet at the point where they can be used as accurate representations of reality. In this current stage, the rate of false positives prevents their use for analysis of dynamics, agents, impacts, and spaces of violence. Due to the rate of duplicates, automated datasets like Phoenix, GDELT, and ICEWS are more accurately interpreted as a measure of media interest in a story and the frequency of reporting on various conflicts, rather than a source for the details of the event themselves. It is not appropriate to use these event datasets to present trends, maps, or distributions of violence in a state.

Inclusion: *Inclusion criteria should allow for accurate representations of political violence, while being flexible to how political violence has changed.* Who is considered a relevant and legitimate actor in conflict is pre-determined by the mandate of the dataset; the definitions, catchment, and categorization are critical, as they tell a user who and what is likely to be included. In turn, the rules of inclusion determine how and in which detail we understand violence. ACLED is specifically designed to capture armed, organized political violence and demonstrations as they occur in all countries. It is not limited by arbitrary thresholds that obscure low-level violence, violence linked to unidentified perpetrators, and other forms of disorder, which can distort the results of other human-coded datasets. Likewise, it is not subject to the shortcomings of machine-based datasets, in which the inclusion criteria become the de facto methodology of the set, in the absence of regular oversight and review.

Coverage and Classification: *Clear, coherent, and correct classifications are important for users because conflicts are not homogenous: disorder events differ in their frequency, sequences, and intensity.* Event types that reflect the variation of modalities common across conflicts and periods of disorder are basic, central components of insightful and useful analysis. Without capturing the spectrum of conflict, datasets can distort the truth about the distribution of instability and impact on civilians and state security. Global coverage should include political violence in each country, and should represent how political violence occurs within that state. Modern conflict is complicated, multifaceted, and adapts to the domestic politics of each state. ACLED is the sole dataset that is sufficiently nuanced and flexible to account for political violence in contexts as different as India and Ukraine.

Useability and Transparency: *Datasets must be useful and useable if they are to be relied upon for regular analysis, and users should be able to access every detail of how conflict data are coded and collected.* Useability is closely tied to straightforward, consistent inclusion criteria and clear methodology. ACLED is unique across all datasets for the level of transparency, user interaction, and user and methodology documents. Near real-time data requires the ability to capture low-level events, seek accurate sources of triangulation, and have a robust checking and updating system. It allows for evaluation and monitoring of ongoing and past projects, politics, and strategies. ACLED publishes weekly updates, and updates events based on new, relevant information in real-time, always making use of newly available information. Raw conflict data can be downloaded from each of the reviewed datasets, but ACLED alone provides weekly updates and constant analysis of its own data for public use. These offer a further check on the robustness of the information and the utility of these data for users.

Methodological transparency is the key to trust between users and data providers. Only then can a user truly understand the assumptions they must incur when working with the data.

Sourcing: *Extensive sourcing — including from local partners and media in local languages — provides the most thorough and accurate information on political violence and demonstrations, as well as the most accurate presentation of the risks that citizens and civilians experience in their homes and communities.* ACLED privileges a wide and deep net of sourcing when gathering information, which comes from: (1) traditional media, (2) select new media, (3) public and private reporting, and (4) partnerships with local conflict observatories — all across a wide range of languages. Of the datasets reviewed, only ACLED prioritizes, integrates, and acknowledges local partners and the information they provide, and builds the capacity of local researchers to represent the risks within their own states.

ANNEX: COMPARISON TABLE

	ACLED	GTD	GDELT	ICEWS	UCDP GED	Phoenix
Mandate	Disaggregated, real time, coverage of political violence and protest globally	Terrorism	Reported occurrences and tone around the world	Politically relevant events, including some incidents of crime	'Conflicts' based on a minimum fatality threshold	Unrest, civil conflict, regime change and incidents of accidents and crime
Spatial Coverage	Near global (data collection in progress)	Countries where a terrorist event has occurred	Global	Global	Countries where a conflict has at one point exceeded 25 deaths	Global
Temporal Coverage	Africa 1997-> Asia 2010 -> Middle East 2016 -> Other areas 2018 ->	1970-2017	From 2013->	1995->	1946-2017	1945-
Publication schedule	Weekly	Annual	15 minutes	Daily	Annual	Unknown
Event Definition	A politically violent event is a single altercation where often force is used by one or more groups for a political end. Some instances including protests and strategic development are included to capture the potential precursors or critical junctures of a conflict.	A terrorist act must satisfy three criteria: a) be intentional; b) incident must entail some level of violence or immediate threat of violence; and c) the perpetrators of the incidents must be sub□national actors. The database does not include acts of state terrorism	CAMEO methodology	CAMEO methodology	The incidence of the use of armed force by an organized actor against another organized actor, or against civilians, resulting in at least 1 direct death in either the best, low or high estimate categories	CAMEO methodology
Geographic specificity	Georeferenced, with 8 geography indicators	Georeferenced	Georeferenced[39]	Georeferenced	Georeferenced	Georeferenced
Covered Agents	Government Rebels Political militias Communal/ Identity militias Rioters Protesters Civilians External forces	Non-state groups	Specific Role (Government, Rebels, Political Opposition, etc.) Broad role (Education, Elites, Media, Refugees, etc.)	Individual Group Country Location	Governments Organized groups	Government Opposition groups Military, Political party Rebels/insurgents Criminal actors Business Media-related actors/organizations Civilians
Sourcing	Local partners, NGO and humanitarian agency reporting;	Open-source materials include electronic news	Open source news media available through websites in numerous	Generated from news aggregator in English, Spanish, Portuguese,	Primarily, media aggregator in English	New York Times, BBC, Keesing, CIA's Foreign

[39] The georeferenced location for an actor may not always match the coded location, such as in a case where the President of Russia is visiting Washington, DC in the United States, in which case the Actor1_CountryCode would contain the code for Russia, while the georeferencing fields below would contain a match for Washington, DC. It may not always be possible for the system to locate a match for each actor or location, in which case one or more of the fields may be blank.

	conflict observatories, local, regional, national and continental media in over 20 different local languages	archives in English (or translated into English), existing data sets, secondary source materials.	languages	and Arabic		Broadcast Information Service (up to 2015) in English.
Useability	Easy to download, accessible, data Clear methodology and training, format in excel, API, analysis or dashboard	Data available on project website	Massive data loads with no accessible methodology	Massive data loads with no accessible methodology	Website form to download	Massive data loads with no accessible methodology
Summary	Wide and unique coverage to capture the reality of political violence and protest as it occurs in all countries in real time. Integrate and build the capacity of local groups to measure and relay risk.	Wide coverage, extensive methodology but limited capture of events that fall within definition of terrorism	Inconsistent, poorly maintained, rife with error, subject to enormous degree of false positives, not checked or reviewed for consistency or accuracy	Based on limited, sources with little oversight and a focus on extreme reported classifications and events[40]	Predominantly English-language; limited and strictly defined definition of what a 'conflict' is. Country coverage based on whether a state is believed to host a pre-defined 'conflict'	Based on limited, English sources with little oversight and a focus on extreme reported classifications and events
Coding process (machine or human)	Human	Human	Machine	Machine	Human	Machine

[40] Steinert-Threlkeld, Zachary. (12 April 2015). "What Sources Are in ICEWS?" Zachary Steinert-Threlkeld, WordPress.
<https://zacharyst.com/2015/04/12/what-sources-are-in-icews/>.