Expanding the Coverage of Conflict Event Datasets: Three Proofs of Concept[*]

Andrew Shaver[†], Hannah Kazis-Taylor[‡], Claudia Loomis[§], Mia Bartschi[§], Paul Patterson[§], Adrian Vera[§], Kevin Abad[§], Saher Alqarwani[§], Clay Bell[§], Sebastian Bock[§], Kieran Cabezas[§], Heidi Felix[§], Jennifer Gonzalez[§], Christopher Hoeft[§], Aileen Ibarra Martinez[§], Kai Keltner[§], Jessica Moroyoqui[§], Kieko Paman[§], Ethan Ramirez[§], Priscilla Reis[§], Juan Jose Rodriguez jr[§], Jazmin Santos-Perez[§], Katha Komal Sikka[¶], Arjan Singh[§], Cassidy Tao[§], Richard Tirado[§], Aishvari Trivedi[§], Lillian Xu[§], Margaret You[§], Meriam Eskander[§]

September 7, 2023

Many contemporary studies on political violence and social unrest rely on conflict event datasets, primarily derived from major international/national news media reports. These conflict-event data are widely used and funded by governmental and other entities. Yet, a large body of research identifies systematic patterns of 'missingness' in these data, calling into question statistical results drawn from them. In this project, we explore three specific opportunities for additional data collection to help recover systematically excluded events, and to potentially assist in addressing resulting bias. We find that all three approaches result in additional and often systematically different material than that reported in news-based datasets, and we reflect on the advantages and drawbacks of each approach.

Introduction

Academics and policymakers rely on cross-national conflict event datasets derived wholly or partly from news reports; however, a significant body of research has revealed systematic

[*] The authors report there are no competing interests to declare. We thank the seven freelance journalists who collaborated with us and the many additional news media professionals who agreed to be interviewed for this research. For comments on this project, we are grateful to participants of the Empirical Studies of Conflict (ESOC) University of California Faculty Meeting and *Civil Wars*' editorial team and anonymous referees. For their support in executing this project, we thank the following University of California, Merced departments: Customer Relations and Research and Academic Procurement Team; Center for Business Services and Solutions; Institutional Review Board Office; Office of Research and Economic Development; Tax Services; and Accounts Payable. We received approval from the University of California-Merced Institutional Review Board #UCM2021-125.

[†] Assistant Professor, University of California, Merced; ashaver@ucmerced.edu; Corresponding Author

[‡] Doctoral Student; Princeton University; hk2880@princeton.edu

[§] Research Intern, Political Violence Lab

[¶] Research Fellow, Political Violence Lab

issues with such datasets. In this 25th Anniversary Special Issue article, we pilot diverse approaches to supplementing existing datasets, and offer recommendations to address potential sources of systematic ‘missingness’.[1]

Empirical conflict research has shifted away from analyzing broad global patterns in conflicts using country-level data (e.g. conflict onset (Fearon and Laitin, 2003), duration (Collier et al., 2004; De Rouen Jr and Sobek, 2004), and settlement (Walter, 1997)). Recent literature has taken a more fine-grained approach, analyzing patterns in individual violent events or conflict dynamics in smaller regions (see Berman et al. (2018) for a broad review). To do so, scholars across different fields[2] have increasingly relied on conflict event datasets. Datasets on various aspects of conflict, including such conflict-event datasets, have improved by becoming increasingly disaggregated, i.e. reporting more precise details rather than topline statistics (e.g. yearly measures at the country level).[3]. Disaggregated incident-level datasets have been used to examine at a more micro-level not only civil wars and insurgencies (Berman et al., 2011; Crost et al., 2014; Sexton, 2016; Condra et al., 2018), but also terrorism (Hoeffler et al., 2022; Mroszczyk and Abrahms, 2021; Laktabai, 2020; Tin et al., 2021), social unrest including protest activity (Sutton et al., 2014; Klein and Regan, 2018; Bodnaruk Jazayeri, 2016; Ives and Lewis, 2020), and other forms of political violence.

Disaggregated conflict-event datasets are also both used and funded by a variety of public sector actors including United Nations entities, United States governmental agencies, and other national governments.[4] These datasets, which are broadly focused on political violence/social unrest and often global in coverage, include the Armed Conflict Location Event dataset (ACLED) (Raleigh et al., 2010); Global Data on Events, Location, and Tone (GDELT) (Leetaru and Schrodt, 2013); the Georeferenced Events Dataset (GED) (Sundberg and Melander, 2013); Global Terrorism Dataset (GTD) (LaFree and Dugan, 2007); Integrated Crisis Early Warning System database (ICEWS) (Boschee et al., 2015)[5]; and the Social Conflict Analysis Database (SCAD) (Salehyan et al., 2012). Derivative datasets have manipulated the existing media-based datasets to focus on more specific areas of conflict.[6] Finally, the rising collation of news reports into datasets has generated subsequent integration efforts to improve the accuracy of relevant data (Zhukov et al., 2017; Donnay et al., 2019).

These conflict events datasets underpinning increasingly micro-level research on conflict by academics and policymakers are constructed largely or wholly from major international and national news media reports.[7] However, a growing body of research has identified patterns of systematic missing data in media-based conflict event datasets (which we cite at length below). We pilot and test approaches to recovering missing data, and show that our approaches can help correct patterns of systematic missing data in existing conflict event datasets.

Given the prominence of media-based datasets, we do not seek to discourage their use (though, we encourage caution!) but rather suggest ways of enriching these data to benefit academic scholarship and governments' policy and programming. In the discussion, we reflect on the reliability of alternative data sources and how these approaches might help address sources of systematic missingness in existing cross-national datasets. We focus on recommendations that entities funding and developing conflict event datasets might consider adopting.

We pilot three different approaches to expanding existing datasets. Our approach and conclusions are informed by in-depth interviews with media professionals familiar with reporting on political violence/social unrest in countries around the world[8] who work, or have reported as freelance journalists, for major outlets.[9].

Our first effort uses photo and video journalism, rather than written articles, to track previously unidentified incidents of political protest and social unrest. Second, we integrate records of violent incidents from local-language media, NGOs, and local authorities. Third, we contract in-country journalists to log all relevant events they learn about; we then compare the events they identify to those reported in existing conflict-event data.

In brief, we find that all three approaches result in additional and often systematically different material than that reported in news report based data, and we reflect in the discussion on the advantages and drawbacks of each approach.

Through this paper, we make several contributions. First, we help advance the debate over the use of conflict event datasets based on news reports. Rather than a black-or-white recommendation of simply avoiding or using these datasets, we not only encourage actively working to improve them but offer tangible solutions to do so. Second, our collaborations with local journalists through independent contracts provide a model for sustained and potential broader engagement with journalists around the world. Such partnerships between academia and news media, particularly in the IR space, are currently lacking and are therefore urgent and necessary. Finally, if/as the data augmentation processes we recommend are implemented, the newly updated data might be used to retest prominent research findings that rely on the existing conflict event datasets.

This paper proceeds as follows: first, we outline the shortcomings of using news reporting to build protest and violence datasets. Second, we enumerate our three independent approaches to expanding these datasets, namely (1) photo and video journalism, (2) local-language media, NGOs, and local authorities, and (3) the independent contracting of in-country journalists. Third, we present our results, and fourth, we reflect on our results and conclude with further suggestions on additional data sources.

Shortcomings of News Reporting for Building Protest and Violence Datasets

A significant body of research has shown that cross-national datasets tend to omit certain events; but more importantly, they are likely to systematically omit particular *types* of events. Skewed patterns of reporting are particularly problematic as systematic mis-measurement is likely to produce bias in statistical estimates.

Scholars have identified patterns in omissions of violent events based on geography, time, type of violence, and identity of the perpetrator. Geographically, an event in a populous area is more likely to be covered than an event in a remote one (Weidmann, 2015; Dietrick and Eck, 2020; Eck, 2012; Kalyvas, 2004); events in 'Western' countries are also covered at higher rates (Behlendorf et al., 2016). The timing of an event also influences its media coverage: instances of political violence are significantly under-reported prior to elections when compared to post-election reporting (Von Borzyskowski and Wahman, 2021). Furthermore, the type of violence can influence reporting: media outlets disproportionately cover more severe (Croicu and Eck, 2022) and sensational forms of violence (e.g. bombings) (Zhukov and Baum, n.d.; Shaver et al., 2022).

Given that cross-national datasets rely on media reporting, they may reproduce these biases in media coverage. Conflict events may go unreported in international media for a host of reasons. As Shaver et al. (2022) detail, journalists face a variety of restrictions on their ability to report on events, including difficulties accessing remote areas. Media outlets may be influenced or directly controlled by the local government, impacting whether and how they cover events (Miller et al., 2022). For instance, governments can deliberately restrict journalists' access to information with internet blackouts, as likely occurred following protests in Iran in 2022 (Campbell, 2023) and in India in 2018 (Hussain, 2023).[10]

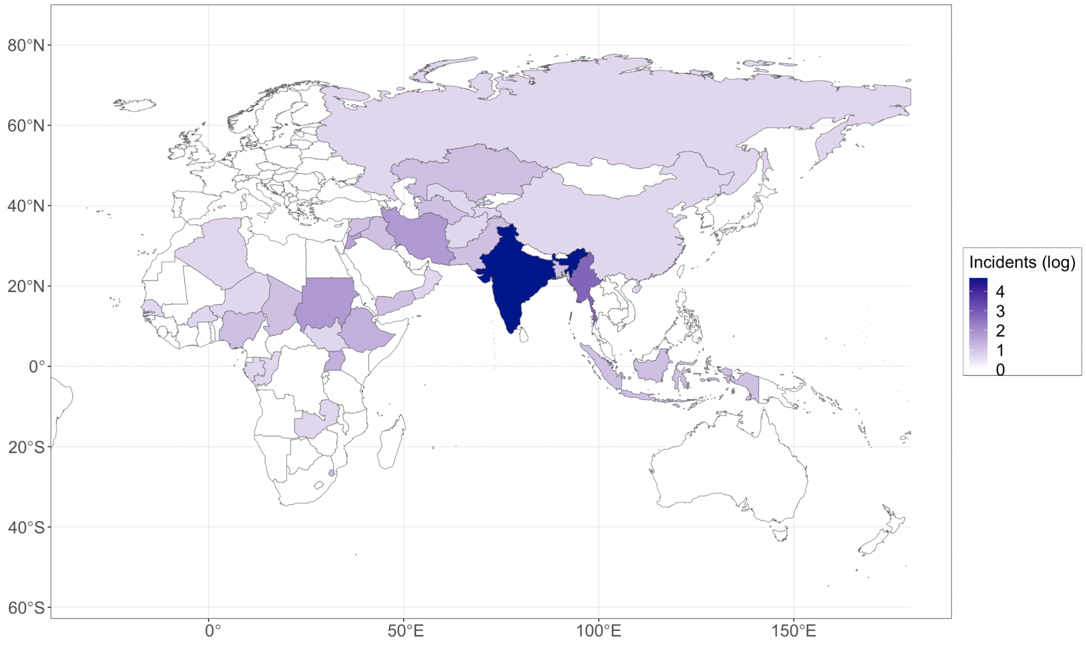


Figure 1: This figure shows many of the countries experiencing political violence/social unrest around the world are the same ones whose governments are shutting down information and communication technologies that likely underpin the news media's ability to report violence in those countries. Sources: AccessNow (2016), Schvitz et al. (2022).

The effects of systematic underreporting are significant. In a large-scale "reverse replication" exercise, Shaver et al. (2022) attempt to recover the results of a large number of articles published in leading economics and political science journals using media-derived conflict data in place of high-quality administrative data originally used in those studies. They find that the majority are irrecoverable.

Paths Forward? Exploring Plausible Supplements to Major News Article-Derived Data

Given the established limitations of conflict event datasets that rely on major news media reports, how can we supplement these datasets to limit missing data problems?

Schutte and Kelling (2022) highlight some solutions for remedying discrepancies in data collection, including recommended statistical analysis and conceptual framings. Similarly, Donnay et al. (2019) put forth MELTT, or Matching Event Data by Location, Time, and Type, as an inexpensive methodology designed to improve data collection accuracy concerning spatially aggregated and machine-coded datasets. Von Borzyskowski and Wahman (2021) suggest employing a survey-based approach for studies with a small scope, and cross-checking the results with dataset findings to reduce errors. When working on a larger scale where multiple datasets contain relevant data, Cook et al. (2017) introduce an estimate model of misclassification and appropriate risk-model probabilities weighting, which they find significantly reduces the impact of biases on data. Otto (2013) offers more broad suggestions, including increasing transparency of coding procedures, using more exact definitions, and ensuring researchers utilize appropriate statistical models.

In the rest of this section, we propose and carry out preliminary tests of three strategies for supplementing media-derived conflict event data.

Identifying Events From Photo- and Video-Journalism

Our first data effort to track the types of incidents systematically overlooked by media and conflict event datasets uses photojournalism, rather than print journalism.

Conflict-event datasets generally rely on written news articles. Yet, in addition to producing news articles about incidents of violence and social unrest, major global news organizations like the *Agence-France Presse* (AFP), *Associated Press* (AP) and *Reuters* frequently capture video and photos of events. Crucially, these outlets often do not publish a written story about events captured by photo and/or video. As an interviewee[11] described:

> *There's a considerably lower bar for covering protests and such for video than for text. If we are covering a protest for text... before you write a story about it, you*

> *would want there to be a reason – and the reason is usually that it represents a large portion of society going out to the streets... you want thousands of people, at least, representing the grievances of many more thousands of people. With photo and video especially, they cover protests a lot, even when there is as few as fifteen people there...*[12]

We test whether photo and video archives maintained by the *AFP Forum* (n.d.) and the *AP Newsroom* (n.d.) can help expand conflict event datasets. For this proof of concept, we focused specifically on comparing protest and riot events tracked by ACLED, given that both photo-/video-journalism and that particular dataset both often focus on protests and social unrest.

We examained patterns of reporting across eleven countries, chosen to ensure generally broad geographic representation: Brazil, Colombia, Egypt, France, Haiti, Myanmar, Nicaragua, Pakistan, Peru, South Africa, and Yemen. We describe our video and photo search processes and comparisons with ACLED in detail in A.2.

Through a subscription with the *AP*, we also gained access to the outlet's planned coverage of events by media type. For each news event that the organization plans to cover, the means of coverage (photo, video, and/or text) is indicated, allowing us to validate an interviewee's[13]

description of heterogeneity across reporting types. For a one month period beginning on May 23, 2023, we monitored the planned news coverage of *expected* protest and similar activity. To identify such events, we used search terms consistent with our previous efforts investigating biases in photo and video journalism (again, see A.2 for details). For each identified incident of expected social unrest, we recorded the intended coverage method. For instance, on May 29, 2023, *AP* reported plans to cover the "[p]rotest over Saudi execution of 2 Bahraini men over militant activities," with a video segment but no written article.

During this one month effort, we identified 65 planned coverage events deemed relevant. Of those, 38 events ≈58.46% were to be covered by text. 25 (≈38.46%) were to be covered by photo or video and not with text. More generally, 62 (≈95.38%) were to be covered with photo or video. The extremely high rate of photo/video coverage is informative because, in addition to the events that do not receive text coverage that can be detected through photo/video, these resources might also be used in the identification of additional details that do not appear in the written articles.[14]

Detecting Events From Local Sources

Our second effort involves tracking events covered by local-language news sites, civil society actors, and government to identify systematic patterns of missing data in existing datasets.

While international/national media may tend to cover larger-scale events and those otherwise aligned with particular editorial preferences, local news sources and organizations based on the ground may be more likely to also report smaller scale events, given local populations' interest. A number of scholars have empirically evaluated whether reports by local media and civil society organizations suffer from fewer biases than aggregated event datasets that often draw on international news sources. Demarest and Langer (2018) show that in Nigeria, datasets relying on local as opposed to international news sources log significantly more protest and political violence events. Clarke (2021) also finds that local media sources on protests in Egypt capture many more events than existing datasets, but still reveal some biases of existing datasets in undercounting smaller events outside of the capital, as compared to local activist groups' records. Davenport and Ball (2002) identify different biases of violence reporting by newspapers, human rights organizations, and interviews in Guatemala.

However, local media may not offer better coverage of violent events than current datasets in all cases. The quality of local media reporting depends on local conditions including availability of communications technology (Weidmann, 2016; Croicu and Kreutz, 2016), regime type (Baum and Zhukov, 2015), severity of conflict (Davies and True, 2017) and geographical region, with sparser event coverage in Africa (Dietrick and Eck, 2020). Cultural biases may lead local media to omit certain forms of violence, such as violence against women or ethnic minorities (Davies and True, 2017). Shaver et al. (2022) describe how threats to local journalist safety can skew reporting.

For a proof of concept of improving existing datasets with local-language reporting, we piloted the use of local media, governmental, and NGO sources to identify political violence events in Israel/Palestine. As described in the appendix, we track only extrajudicial violence, and not military activity. To track local-language media coverage of political violence events, we used the Hebrew-language version of the Israeli news site *Yedioth Ahronot* (n.d.), a popular mainstream news site in Israel. We identified all articles including the term "attack", and manually determined whether an article described a case of political violence meeting the criteria of existing datasets. We also logged political violence events covered by local NGO/watchdog organizations and government bodies that were not reported in existing datasets. We drew on catalogs created by civil society organizations aligned with both sides of the conflict. We used catalogs of attacks compiled by the Israeli anti-occupation nonprofit Btselem[15], the American Jewish non-profit Jewish Virtual Library[16], the Foundation for the Defense of Democracy[17], the pro-Palestinian DC think tank the Jerusalem Foundation[18], the Israeli government-linked Meir Amit Intelligence and Terrorism Information Center[19], and a catalog by Dr. Wm. Robert Johnston[20]. Finally, the Israel Defense Forces published a list of political violence events covering September 2016 through October 2016; we logged events recorded here but absent from existing datasets.

Direct Information Sharing From Journalists

Our third approach to collecting data on the types of violent incidents often overlooked by international/national media—and thus conflict event datasets derived from their reporting—takes a different approach: we pilot directly contracting journalists on the ground to report on all incidents of political violence and social unrest about which they hear.

Our in-depth interviews with media professionals make clear that journalists learn about many more events than are ultimately reported by the outlets they write for: “There is a lot of violence that happens and it can’t all be written about...” [21]

Interviewees highlighted various factors that affect ultimate reporting likelihood. Generally, the more people involved or affected by an event; the more novel the event; and the more high-profile individuals or groups involved, the more likely it is to be covered. Below, we draw from the interviews, highlighting some of the exclusionary criteria and examples of reporting bias they shared.

In conflict settings, often only fatal events (and particularly attacks with many fatalities) are covered. “If people do not die, [there is] much less chance that we are going to be writing about it.” [22]

Another interviewee who reported extensively on conflict in Colombia, made a similar observation: “Sadly and tragically, news events that involve... casualties, deaths often rise in importance and how they are viewed. That's kind of just the nuts and bolts of our business. It elevates a news event in a way that it wouldn't otherwise."[23]

Reflecting on reporting on violence in Burkina Faso, an interviewee described how attacks "against security forces get absolutely no traction... it happens too regularly unfortunately... Sometimes it's six [military personnel killed]; sometimes it's ten... but, in terms of international news, it never makes headlines anymore."[24]

Identity of the people affected can also determine coverage. "...I hate to be blunt about it, but all lives are not considered equal in the eyes of journalists from the lowest level to the highest."[25]

The international wire services "put a lot more emphasis on reporting who is harmed if that person is American or Western."[26]

Another interviewee offered examples from the Iraq War: "If ten Iraqi people get killed, that's nothing. That is not even worth a story in the *New York Times*. [And even more so] if they are killed in remote parts."[27]

Concerning the perpetrators, an interviewee described "one of the more frustrating aspects of reporting [on conflict in Colombia is] that international news organizations would often pay much more attention to atrocities that were carried out by the guerillas... than atrocities carried out by paramilitaries or by government forces."[28]

In settings where protests and social unrest are more common, the focus is often instead often on the numbers of individuals involved. Describing coverage of protest activity, an interviewee described their organzation "limit[ing] the number and type [of protests] we report on because of their newsworthiness"[29] which two interviewees[30] described in terms of participant numbers: protests typically would not be covered unless they reached thousands.

One interviewee further highlighted that even major protests are less likely to be covered the longer they last given the "repetitive" nature of the events and that "stories end up looking alike."[31] Another interviewee echoed this: "There is a calculus of whether something is newsworthy: supposing there is a protest in a place where there is always protests, you wouldn't necessarily write about that."[32]

Several interviewees all described a lack of editorial interest in reports of violence in particular countries or during particular periods.[33] For instance, an interviewee described the difficulty they faced in placing stories of events they uncovered during the 2016 through 2021 period of conflict preceding Russia's invasion of Ukraine.[34]

These factors influencing event coverage clearly highlight significant differences between events journalists learn about and what they publish; further, this list of editorial pressures is certainly not exhaustive. Ideally, accessing the complete set of incidents that journalists learn about in the course of their reporting (not just those that are reported) can help circumvent editorial bias.[35] This is precisely what we seek to do through a series of collaborations with freelance reporters in our third proof of concept.

To better understand the set of events that journalists learn about in the course of their reporting, we entered into partnerships with seven freelance journalists who have all written for major international news media outlets. They reported to us all incidents of political violence and social unrest that they learned about in the course of their work, regardless of whether or not they considered the events newsworthy or likely to be published. Collectively, they covered events in Mozambique, Pakistan, Peru, South Africa, and Zimbabwe. We then compared the events that these journalists identified with the events reported by ACLED for these same countries over the same respective time periods. Additional details of these arrangements and our approach to comparing events appears in A.3.

These countries were chosen on the basis of 1) the success of our efforts in identifying journalists to collaborate with and 2) the nature of ongoing or expected unrest in these countries. These countries vary significantly in terms of the type of political violence they are experiencing. They span different regions of the world. For instance, one journalist described how ongoing violence in Pakistan's Balochistan province is underreported by the news media relative to other parts of the country. Peru recently experienced significant unrest in its Puno region, where protesters established roadblocks and temporarily shut down several airports (*Al Jazeera*, 2023). Despite a recent respite in unrest, violence is expected to resume. Zimbabwe has experienced some violence as the 2023 elections approach, with more potential political violence and government repression expected as the date is confirmed and approaches.

Results

Our three efforts each identified new incidents that were not previously tracked by news-based datasets. While the efforts varied in the number of new incidents they uncovered, each revealed incidents that differed systematically from those that were tracked by existing media-based datasets.

Identifying Events From Photo- and Video-Journalism

We find that the news-based datasets did not cover a significant portion of events tracked by photo- and video-journalism outlets, and that the events not tracked by the news-based datasets often systematically differed from those included in the datasets.

However, we found that a relatively small number of events overall were reported only in photo/video journalism, but not in print news reporting. Photographed/video recorded events not captured by the news-report based data made up a small proportion of total events related to social unrest – from just shy of 1% in one case (Myanmar) to approximately 4% in another (Nicaragua).[36] Relative to the other two proofs of concept, this approach thus generated much less overall new material. (Though, as we discuss later, these results are based on limited set of source materials, which future efforts might expand.)

Although the volume of newly identified events is relatively limited, we find that previously and newly identified events vary significantly. Newly identified incidents may serve not only to expand overall content but to help specifically in expanding the set of activities that are systematically underreported in existing event datasets. Using our estimates of whether the photographed/video recorded events involved violence, we calculate the predicted probability of inclusion in ACLED.[37] We find that violent events were ≈16.48 percentage points more likely be included in the media based data (≈42.66% vs. ≈26.18%). (See Figure 2.) Furthermore, when we compare the set of newly identified events with the overall body of ACLED events involving social unrest (that is, all comparable events from the the same countries and time periods), we again find that non-violent events were significantly less likely to be captured.[38]

We also observe significant cross-country differences. Of the photos and videos depicting social unrest in Myanmar, for instance, only 14.96% were not tracked by ACLED. In contrast,

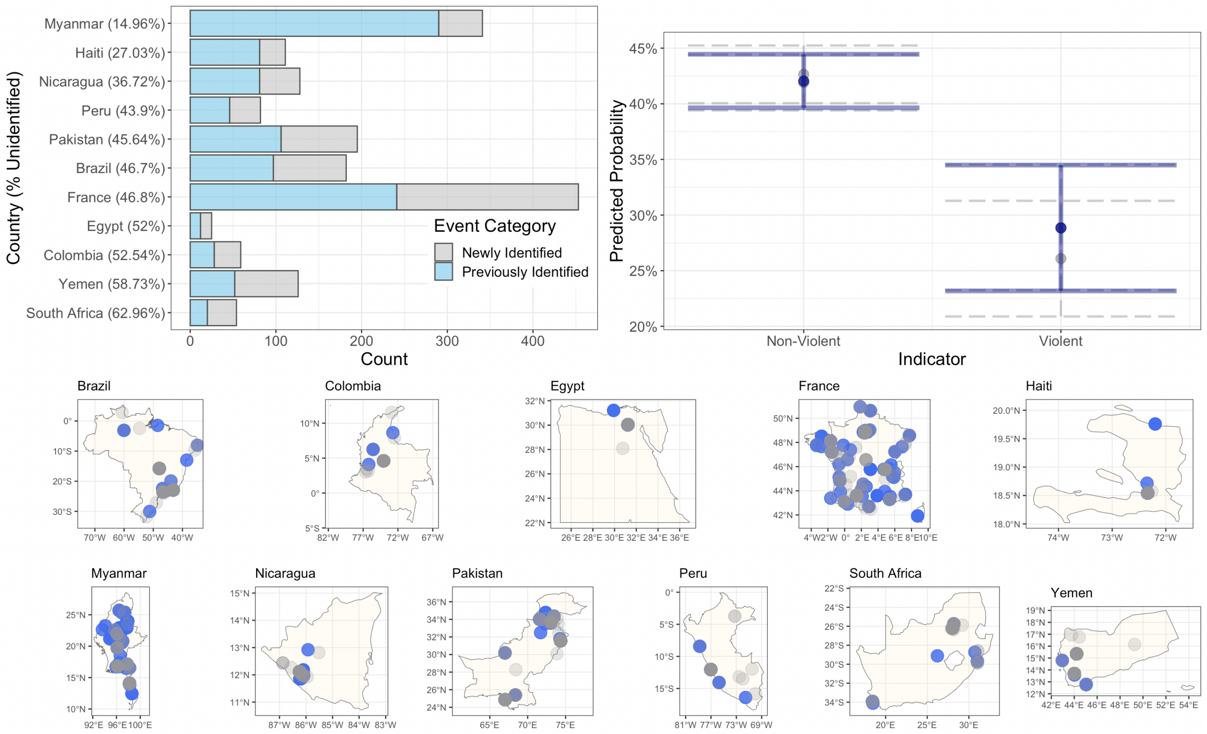


Figure 2: This figure displays the results of comparing photo/video content tracked by the news media based data (blue) vs. those newly identified from the materials (gray). The upper-left figure depicts differences across countries. The upper-right figure displays the predicted probability of not being previously identified when violence is and is not assessed to have been associated with the event. Finally, the bottom figure displays the locations of newly and previously identified events. (These plots involve, but are not limited to, the use of data from acleddata.com.)

of the incidents identified in South Africa, we estimate that 62.96% were not captured.

So, given that the photo/video-only events often differed systematically from those recorded in existing datasets, there are compelling reasons to augment existing conflict event datasets with photo and video based incidents; we discuss these recommendations further in the conclusion.

Detecting Events From Local Sources

Our second data collection effort, which draws on local media, nonprofit, and government sources, approximately doubled the number of attacks identified in Israel/Palestine by existing cross-national datasets from 2000-2023 (we identified 4,516 relevant attacks across the ACLED, RAND, and GTD datasets, and uncovered an additional 4,191 incidents).[39]

We find that the three existing datasets we tested systematically omit certain kinds of

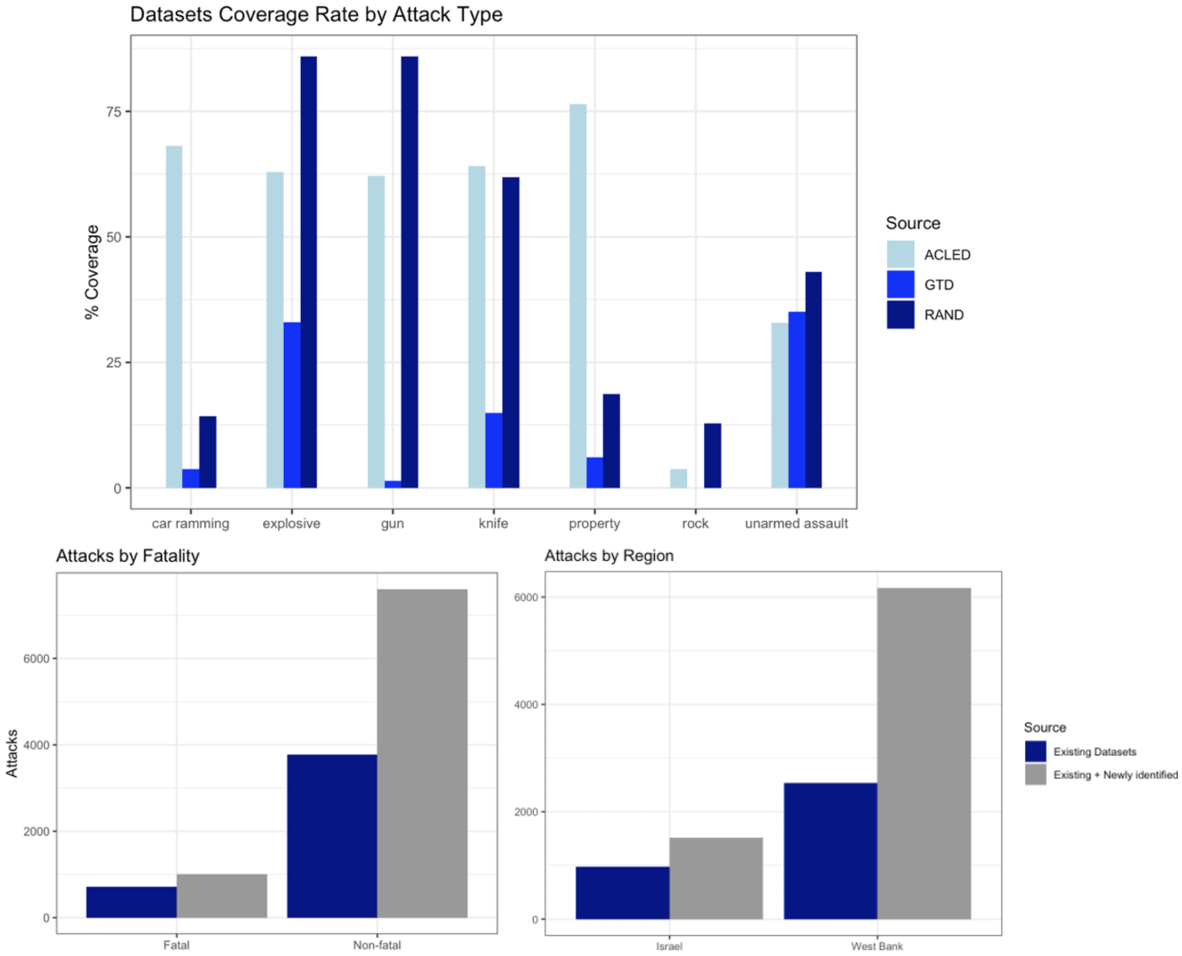


Figure 3: This figure displays the results of our data collection efforts on extrajudicial violence in Israel/Palestine from 2000-March 2023, based on nonprofits, local governments, and local-language media. The first figure shows the percentage of all identified attacks (pooling existing datasets and our original dataset) logged within the time period each dataset was active that each dataset had included, by attack type. The second and third graphs compare all three datasets' coverage of attacks by fatality and region, respectively, to the complete set of attacks identified by any of the datasets or our original efforts. (These plots involve, but are not limited to, the use of data from acleddata.com.)

attacks; omissions follow the patterns identified by (Shaver et al., 2022). We find that non-fatal attacks are systematically omitted, with existing datasets identifying only 50% of non-fatal attacks we have logged, as compared to 72% of all identified fatal attacks.

We also find that unarmed attacks, or attacks using homemade weapons, are disproportionately omitted from existing datasets (see Figure 3). While almost all attacks using guns have previously been identified by existing cross-national datasets, existing datasets have identified less than half of unarmed attacks and attacks targeting property, and almost no attacks using rocks.

As suggested by our interviews, existing datasets disproportionately miss attacks in more dangerous regions. Existing datasets cover 23% more of the identified attacks within Israel as compared to attacks within the occupied West Bank, which has suffered approximately four times as many attacks in the period under study.

The identity of the perpetrator also predicts omission from cross-national datasets. Existing datasets have significantly more complete coverage of attacks perpetrated by Palestinians as compared attacks perpetrated by Israelis.

Direct Information Sharing From Journalists

Our third data collection effort, contracting local journalists to report on violent incidents, also substantially expand the set of events reported by the media-based data. We believe that such efforts can help mitigate patterns in missing data in existing conflict-event datasets. We provide the broad descriptive statistics for each country in turn:

In the most modest case, in South Africa, the journalist reported 19 events, ≈36.84 to ≈52.63% of which were newly identified.[40] We estimate that these newly identified events make up ≈3.61 to ≈5.15% of the 194 total comparable events tracked by the news report based data during this period.

In contrast, in Pakistan, the two journalists reported 184 events, ≈61.41 to ≈76.09% of which we estimate to be newly identified events. Again, for comparison, the newly identified events are estimated to make up between ≈19.25 to ≈23.85% of the total 587 events tracked by the news report based dataset during the same period. Of particular note, a substantial number of these newly identified incidents depict extreme levels of violence in the country's Balochistan province that are virtually invisible in the news-report based data, which we discuss in more detail below.

Results from Zimbabwe are also stark: the two journalists reporting from that country reported 31 events, of which ≈62.07 to ≈68.97% were newly identified. The newly identified events account for 75.00 to ≈83.33% of the 24 total events tracked by the news media data during this period.

Finally, in Mozambique the journalist reported 18 events, of which we estimate ≈22.22 to 50.00% were newly identified (making up between ≈13.33 to 30% of the total number of news report based dataset entries). And in Peru, the journalist reported 34 incidents. ≈55.88 to ≈61.76% were newly identified, making up ≈11.24 to ≈12.43% of 169 incidents tracked by the new report based data.

Importantly, the journalists often picked up classes of events that systematically differed from those that appear in the news report based data. The heterogeneity of results across countries makes generalizing difficult. Thus, we instead remark on a few prominent findings across country cases, which reveal the potential power of involving journalists directly in the reporting/data collection process.

In Pakistan, we make two observations. First, the journalists reported a substantial number of deaths associated with armed conflict and social unrest beyond those reported in the news report based data. Over the one month of reporting, we estimate that they tracked between 59 and 83 additional fatalities (between ≈33.91 and ≈47.70% of the total number of fatalities reported by the news report based data). Second, they tracked a substantial number of armed attacks involving insurgent and separatist forces that were not captured by the media-based data. We estimate that the journalists captured between 71 and 91 additional attacks during their one month of reporting (resulting in many dozens of previously untracked injuries and deaths).[41]

As Figure 4 reveals, much of this fighting occurs in Pakistan's Balochistan region, which was clearly highly undercovered relative to the provinces of Khyber Pakhtunkhwa and Sindh. As one of the reporting journalists described to us, "[l]awlessness, Balochistan's remote location, strict army control, and inadequate communication and infrastructure were the dominant factors" that limit reporting in the area.

In Zimbabwe, we note the broad geographic coverage of the two journalists' activities. Nearly one third of incidents of political violence and social unrest occurred across six districts[42] of the country's 81 (≈7.41%) in which there was no recorded activity in the media based data.

Finally, we note that relative to the percentage of newly identified incidents of social unrest (relative to captured incidents of social unrest), the percentage of newly identified violent attacks (relative to detected attacks) was larger in three of the five countries. While our sample is small, this result may point to important cross-country heterogeneities in marginal returns to journalist engagement across different forms of violence/unrest. For instance, in Peru, we estimate that between 50 to 60% of incidents of social unrest captured by the journalists were newly identified. In contrast, more than 70% of attacks were newly identified. Similarly, in Zimbabwe, whereas ≈33.33% of incidents of social unrest were newly identified, we estimate that between ≈66.67 and 75% of attacks were newly identified. We observe a similar pattern in South Africa.

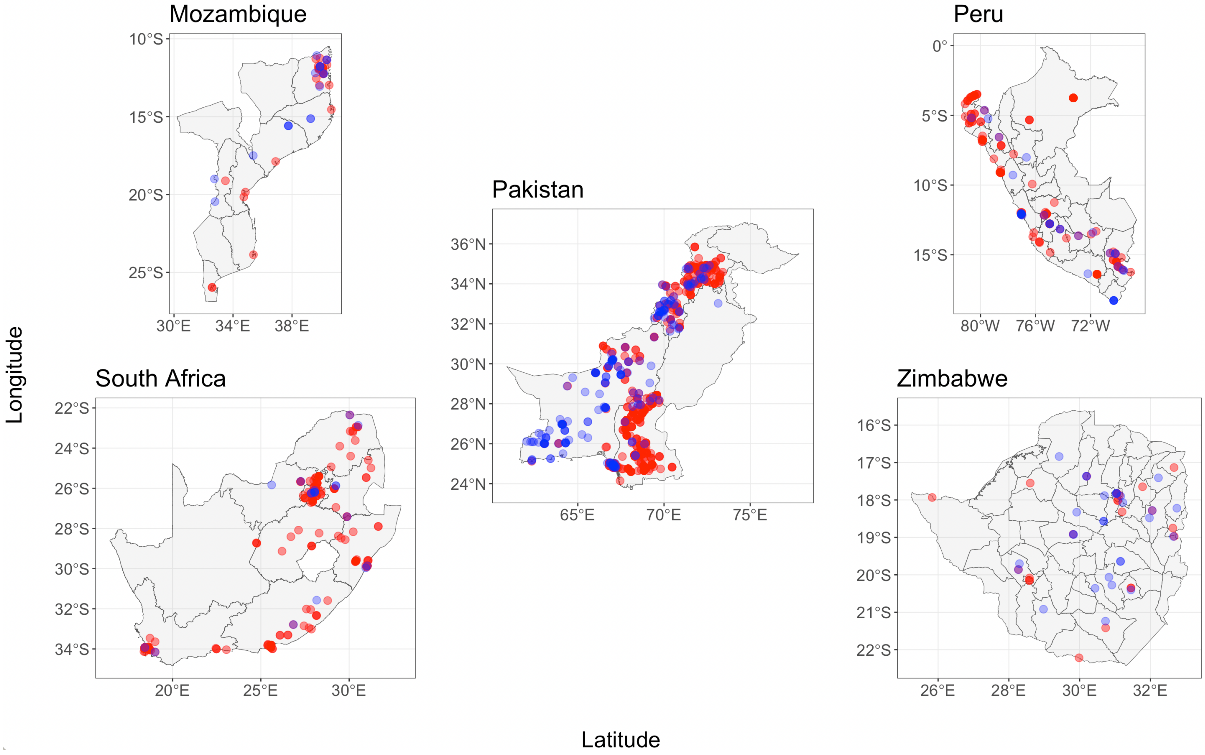


Figure 4: This figure displays the distribution of incidents of political violence and social unrest tracked by the journalists with whom we contracted (blue) alongside the distribution of comparable events tracked by news report based conflict event data (red) over the same time periods. (These plots involve, but are not limited to, the use of data from acleddata.com.)

Discussion & Conclusion

In this article, we explore a series of methods that curators of conflict event datasets might engage in to supplement existing efforts. We find that all three efforts can be used to identify incidents not tracked in news report-based conflict data. Below, we reflect on the advantages and disadvantages of each approach, and on additional possible avenues for collecting incident-level data on political violence.

Reflections on the Photo/Video Effort

Although the number of events that we newly identify is modest, we first note that we did not consult the universe of professional news photo and video media for this effort. This paper's analysis relies only on entries from the *AP* (photo and video) and *AFP* (photo only). However, other major databases exist (e.g. *AFP* video, Bloomberg (n.d.), Reuters (n.d.), and EPA Images (n.d.)), which if collectively consulted would result in a wider and potentially more substantial set of newly identified events.

Furthermore, incorporating photo and video materials into the conflict event datasets may be relatively straightforward and, perhaps more importantly, sustainable. Just as the curators of existing conflict event datasets have established data streams consisting of written news article content, they might also establish subscriptions and collection methods with news organizations regularly augment their materials with details extracted from photos and videos.

AI language models might be used to increase the efficiency of collecting conflict event data from photo- and video-journalism records. For instance, data extracted from photo- and video-journalism metadata through existing application programming interfaces might then be filtered through an AI language model to classify incidents of political violence and social unrest.

Limitations of this effort are similar to those associated with relying on written news articles to identify and describe incidents of political violence and social unrest. Specifically, only those details reported in the photo/video's title or caption or that can be gleaned from the photo/video itself can be translated into the rows of datasets with incident-level details. For instance, event locations associated with photos/videos are often general (e.g. city name), limiting the precise spatial identification of events.[43] Differences across news media platforms sometimes produce discrepancies in how dates are reported—creating the need for deeper critical analysis of existing creation, arrival, and event dates to determine actual incident dates. Additionally, for a given event (e.g. for a particular protest), multiple photos/videos may be produced capturing that event. While there may be advantages to this approach[44], it also complicates efforts to identify unique events. Furthermore, when individual photos or videos are used to extract details about an event, there is a risk that particular details related to the event may be missed. For instance, one event photo may depict violence while another does not.

Reflections on the Effect to Detect Events From Local Sources

Our pilot data collection effort tracking incidents of political violence in Israel/Palestine shows that local civil society and government authorities organizations have tracked a significant number of violent events ommitted in existing cross-national datasets. We have found that including events from these government or NGO actors in datasets is significantly easier than reviewing local media records. Logging attacks from local media sources is labor intensive and often requires language skills. Further, research assistants often cannot code local media reports without an understanding of a country's geography, factions of a conflict, and ability to recognize the perpetrator's background based on a name. On the other hand, government and NGO reports are typically easier for foreign researchers to code.

Yet, researchers must be cognizant of governments' and NGOs' incentives to exaggerate or underplay certain forms of violence based on their political interests. We omitted a significant number of NGO reports that did not meet our threshold for a violent incident (eg. involving only verbal exchange of insults), or that did not provide sufficient information about an attack. Researchers must verify the quality of the NGO and its reporting, and perhaps triangulate different NGO/media reports.

Despite the challenges related to using local-language media sources, we have found that local-language media often covers more incidents than English-language national media, and may be a useful data sources. Conflict-event datasets and academics might consider using automated translation of local-language media in order to review publications in languages that are not widely spoken.[45]

Reflections on Contracting with Local Journalists

Contracting with local journalists comes with a wide variety of benefits. Chief amongst these is the ability to work with them to obtain the specific details associated with each event of interest to researchers.

Journalists provided us with a more precise location of incidents than would typically be reported in a news article. Per agreement, they provided specific details related to the weapons employed in attacks and the precise coordinates at which an event took place. While journalists writing news articles that form the basis of much of the existing conflict event datasets likely also have access to such details, the news report writing process typically does not provide a mechanism for conveying that information.

Furthermore, only a small number of journalists are required to achieve substantial increases in reporting. In particular, in those countries in which we hired two journalists, increases in newly identified events relative to levels of events reported in the media based data were substantial – though, of course, returns to additional journalists are likely to vary substantially across countries given differences in their sizes (geographic and population), government restrictions, and levels of ongoing political violence. In short, direct and continuing engagement with journalists may be more feasible than expected given high marginal returns at low numbers.

Nevertheless, there are limitations as well. While the collaborations with journalists overcome substantial editorial biases, they do not disappear entirely. For instance, the journalist with whom we collaborated in Peru remarked after working with us on the difficulty of learning about events when many in the country were eager to move past recent unrest: "[A]fter a couple extremely tumultuous months, the country seems to be exhausted and wanting to avoid anything related to political violence." Broader industry editorial pressures may simply make identifying particular content difficult even where the individual journalist is not themself bound by the editorial constraints of their principal employer(s).

Furthermore, freedom from editorial constraints does not necessarily address impediments that make learning about events difficult in the first place. For instance, as one of our journalists reported to us after completing the assignment, "[a]nother issue was that most of the cases of political violence were/are happening in remote rural areas where victims are not even reporting the cases." One of the Pakistan-based journalists described the limitations of reporting given governmental restrictions: "To control the narrative, the Pakistani military has imposed restrictions on media outlets... As a result, mainstream media sources do not cover all militant attacks, except major attacks that occur in cities, like Peshawar." Direct collaborations with journalists may partially circumvent some of these issues (as our analysis shows); but they are likely to persist, continuing to produce some degree of systematic 'missingness' in the data in the process (though attenuated relative to reliance on news articles alone).

Steps Forward

The set of efforts we pilot are by no means comprehensive, and scholars seeking to incorporate conflict event data in their own work might consider parallel efforts. For instance, various high-quality administrative records have been released from government sources, and other similar records may exist.

Researchers might pursue available channels for requesting administrative data from relevant governments and international organizations to potentially acquire non-media data on political violence. For instance, to the best of our knowledge, the U.S. military has not released "SIGACTs" type data related to its engagements in countries like Libya and Yemen. Given the comprehensive nature of data released by the U.S. Defense Department related to its Operations Iraqi Freedom, Enduring Freedom, and Inherent Resolve, it stands to reason that similar records may exist. Wartime records may also be accessible in archives. For instance, both the U.S. Department of Veterans Affairs, through its Official Military Activities Report (OMAR) (Aragao, 2019) database, and Shaver et al. (2023) have extracted fine grained conflict details from the Vietnam War from the National Archives and Records Administration archival base camp data (NARA, 2023), which are electronically available for download. (See Figure 5.)

We also note past and ongoing efforts worldwide to track political violence through other means. For instance, we note the recent efforts of Solstad (2023) to track wartime activity in Ukraine through the use of satellite data on temperature anomalies. Another example comes from the various United Nations missions that have collected high quality civilian casuality data. An excellent example of this is the United Nations Mission in South Sudan (UNMISS)—the UN's peacekeeping mission for the country—which is engaged in a large-scale effort to track civilian casualties across that country. UNMISS collects a large quantity of fine-grained data on violence against civilians. The UNMISS initiative itself provides an important proof of concept for supplementary collection methods to media-based datasets. Indeed, in comparing the number of civilian casualties that UNMISS tracked in South Sudan between the years of 2019 through 2021 with GED, we estimate that UNMISS tracked 981 (2019), 2,336 (2020), and 1,856 (2021) additional fatalities (UNMISS, 2021, 2022, 2023*a*,*b*). In percentage terms, GED's civilian fatality numbers make up ≈13.26, ≈3.67, and ≈2.67% of UNMISS' totals. Similarly,

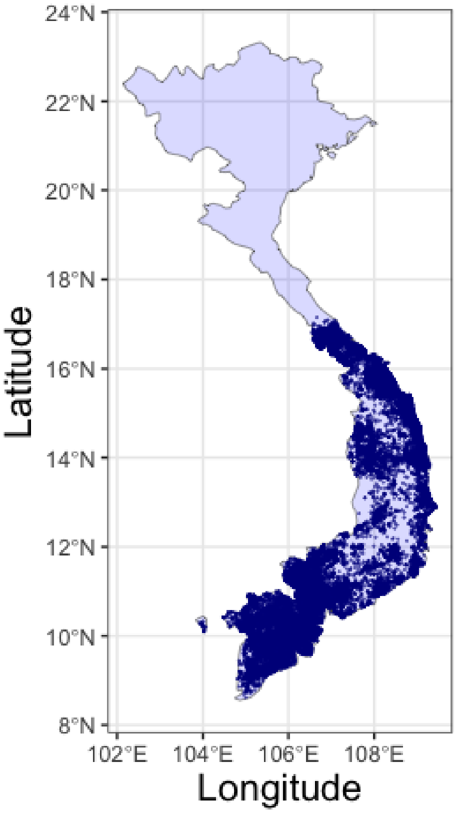


Figure 5: This figure displays the distribution of wartime incidents tracked by U.S. forces during the Vietnam War for the year 1969. Source: Shaver et al. (2023); NARA (2023)

we estimate that UNMISS tracked an additional 160 (2019), 1,763 (2020), and 966 (2021) civilian fatalities compared to ACLED, which equates to around ≈85.85, ≈27.30, and ≈49.34% of UNMISS' totals for those years.[46]

As discussed above, future efforts to collect data on political violence and social unrest are likely to be increasingly augmented with AI technologies—to potentially include utilizing computer vision, enabling machines to analyze and extract information from visual inputs including images, videos, graphics, and text. Indeed, the existing literature already includes some proofs of concept for applying such technologies to conflict analyses. For example, Mueller et al. (2021) trained an AI model to identify structural damage in satellite imagery in Syrian cities, and Aronson (2018) developed an AI model to classify objects in citizen video to identify human rights violations in Aleppo. Further, Radeva (2021) demonstrated the use of visual AI in analyzing documentary evidence through analyzing text, document format, graphics, and predefined objects.

We close, however, with a focus on data generating processes. As AI methods enable us to collect increasingly large conflict datasets, the importance of deep, ongoing collaborations with experts familiar with the data generating processes (e.g. journalists) will be essential to ensure that such future efforts do not fall victim to the same patterns of selection that have skewed existing datasets. As David Hand (2020), emeritus professor of mathematics of Imperial College London, writes, “while it helps to have lots of data—that is, ‘big data’—size is not everything. And what you don’t know, the data you don’t have, may be even more important in understanding what’s going on than the data you do have... [T]he problems of dark [missing] data... are ubiquitous.” We hope that our proofs of concept not only provide specific paths forward but serve to encourage greater and sustained attention to those processes.

Notes

[1] As described below, this effort involves various comparisons with existing conflict event data. We have sought to use these data responsibly and in good faith. The overall goal of this exercise is to identify means by which these existing datasets might be further improved to the collective benefit of the dataset curators and their users, including potential governmental funders. As such, this effort is in no way intended to aid in the development of datasets (or other products) that serve as competitors for these existing conflict event datasets. Instead, the intention is to provide their curators insights about the nature of missing or likely missing incidents from their previous data collection efforts that might inform future collection efforts to their benefit. This research is not intended to negatively depict these conflict datasets or their curators in any manner. Indeed, we have invested a substantial number of work hours in this project precisely because we consider news report based conflict event datasets to be such a critical resource to academic (and potentially other) communities seeking to understand, forecast, and otherwise engage conceptually with political violence and social unrest globally. To the best of our knowledge, there are presently no viable alternatives to the existing news report based conflict event datasets that track conflict and/or social unrest on a global basis. As such, and given how extensively these data are used within academia and government/intergovernmental entities, understanding how these datasets might be further improved is an important public good.

[2] In economics, see Voors et al. (2012); Minoiu and Shemyakina (2014); Manacorda and Tesei (2020); political science, see Choi (2010); Fortna (2015); Steinert-Threlkeld (2017); climate, atmospheric sciences, and oceanography, see O'Loughlin et al. (2014); Hoffmann et al. (2020); and ecology and evolutionary biology, see Daskin and Pringle (2018).

[3] Gleditsch et al. (2014) show that datasets include increasingly disaggregated statistics on (1) the actors in conflict such as ethnic minorities (including the "All Minorities at Risk" (Birnir et al., 2018) and "Ethnic Power Relations" datasets (Vogt et al., 2015); (2) strategies and tactics of conflict such as improvised explosives and terrorist attacks; and (3) conflict beyond violence such as nonviolent protests (see the "Nonviolent and Violent Campaigns and Outcomes Data" dataset (Chenoweth and Lewis, 2013)).

[4] Some of these governmental actors are identified publicly. Though, we have learned about the identities of various other government/intergovermental users through interviews with foreign affairs professionals. See A.1 for a description of these interviews.

[5] ICEWS incorporates the data formerly included in the WITS database (Bowie, 2017).

[6] Such datasets track terrorism (EDTG) (Hou et al., 2020), one-sided ethnic attacks (EOSV) (Fjelde et al., 2021), violence against refugees (POSVAR) (Gineste and Savun, 2019), electoral violence (DECO and CREV) (Fjelde and Höglund, 2022; Birch and Muchlinski, 2020), and violence against peacekeepers (PAR) (Lindberg Bromley, 2018). Others have created new media-based datasets on specific regions or topics: these include country-specific conflict measurements (BFRS and OCVED) (Bueno de Mesquita et al., 2015; Osorio and Beltr´an, 2019); data on suicide attacks (CPOST) (Pape et al., 2021); violent and non-violent electoral contestation (ECAV) (Daxecker et al., 2019); water-related conflict (WARICC) (Bernauer et al., 2012); non-violent

resistance in conflict setting Chenoweth et al. (2019).

[7] As of August 26th 2023, Google Scholar citations of the articles introducing/describing these datasets are: ACLED (Raleigh et al., 2010): 1,975; GDELT (Leetaru and Schrodt, 2013): 843; GED (Sundberg and Melander, 2013): 1,233; GTD (LaFree, 2010): 861; ICEWS (Obrien, 2010): 350; and SCAD (Salehyan et al., 2012): 541.

[8] These include Afghanistan, Burkina Faso, Burundi, China, Colombia, El Salvador, Iraq, Israel, Libya, Mali, Mexico, Pakistan, the Palestinian Territories, the Philippines, South Korea, Sudan, Syria, Rwanda, Ukraine, Venezuela, Yemen, and Zimbabwe.

[9] These include, but are not limited to, *Al Jazeera*, the *British Broadcasting Corporation* (*BBC*), *BuzzFeed*, *Der Spiegel*, *France 24*, *The Guardian*, *The HuffPost The New York Times*, *Public Radio International*, *Reuters*, and *The Wall Street Journal*. In some cases, outlets are not listed here following interviewee requests for anonymity.

[10] In Figure 1, we plot the global distribution of government-imposed internet outages from 2016 through 2022. Journalists operating in conflict zones may also engage in self-censorship and under-report events due to safety concerns (Larreguy et al., 2020). Outlets sympathetic to one side of a conflict may selectively report events favorable to their cause, while withholding unfavorable information (Gibilisco and Steinberg, 2022). Not surprisingly, countries experiencing political violence/social unrest around the world are often those whose governments seek to restrict information, particularly during crucial periods of political violence.

[11] Interviewee 1, 2022. Reporter from a major wire service.

[12] However, it is also important to note that some events are only covered with a written article. Another interviewee reported that budget constraints led their outlet to cover protests in Sudan in a written article, but no photojournalism: "Ever since the Ukraine crisis, we have not had the ability to cover [most] of the protests... in Sudan with video and photos because the budget just isn't there. We've still, as text reporters, been able to cover them [in writing]". (Interviewee 1, 2022. Reporter from a major wire service.) Thus, while editorial considerations appear to drive greater coverage of smaller events with photo and video only, budgetary considerations can limit this.

[13] Interview 2, 2022. Sub-regional News Director for a major wire service.

[14] For instance, as technologies develop, video content might be used to estimate crowd sizes when they are not reported/estimated in news reports. Or they might serve as an alternative estimate.

[15] *Fatalities: All Data*, Btselem: The Israeli Information Center for Human Rights in the Occupied Territories

[16] *Comprehensive Listing of Terrorism Victims in Israel (September 1993 - Present)*, www.jewishvirtuallibrary.org/comprehensive-listing-of-terrorism-victims-in-israel

[17] *Mapping Terrorism in the West Bank*, FDD Visuals, www.fdd.org/analysis/2022/12/12/mapping-terrorism-in-the-west-bank/

[18] *Israeli Settler Violence Database*, Palestine Center, docs.google.com/spreadsheets/d/1yixx$_{C}P94IKfmC5Z9qgoVvWrny9CkSodaoyQuRuE8Z0/edit\#gid1980104107$

[19] *Periodical Studies*, www.terrorism-info.org.il/en/c/periodical-studies/

[20] *Johnston's Archive*, "Chronology of Terrorist Attacks in Israel Introduction",

www.johnstonsarchive.net/terrorism/terrisrael.html

[21] Interviewee 5, 2022. Staff Writer at the *New York Times Magazine* (former *Wall Street Journal* writer in the Middle East).

[22] Interviewee 2, 2022. Sub-regional News Director for a major wire service.

[23] Interviewee 10, 2022. Journalist with *The New York Times.*

[24] Interviewee 7, 2022. Freelance journalist who worked with major North American outlets including *BBC, PRI, France 24,* and *Canadian Public Broadcasting.*

[25] Interviewee 4, 2022. Former cable news executive.

[26] Interviewee 8, 2022. *Buzzfeed News* Reporter / former *Reuters* reporter

[27] Interviewee 3, 2022. Freelance journalist/ former *New York Times* reporter.

[28] Interviewee 10, 2022. Journalist with *The New York Times.*

[29] Interviewee 6, 2022. *Reuters* reporter in Latin America.

[30] Interviewee 1, 2022. Reporter for a major wire service; Interviewee 2, 2022. Sub-regional News Director for a major wire service.

[31] Interviewee 2, 2022. Sub-regional News Director for a major wire service.

[32] Interviewee 8, 2022. *Buzzfeed News* Reporter / former *Reuters* reporter

[33] Interviewee 9, 2023. Freelance human rights journalist; Interviewee 10, 2022. Journalist with *The New York Times*; Interviewee 11, 2022. Staff Writer at *The New York Times Magazine.*

[34] Interviewee 9, 2023. French Freelance Journalist.

[35] Such approach may not entirely eliminate this source of bias, as editorial pressures surely influence where and how journalists focus their time and efforts in the first place. Yet, significant mismatches in what journalists learn about vs. what they report would provide insight into the nature of editorial bias and potentially provide the direction for measuring/estimating it and potentially using such inferences in statistical analyses (e.g. in establishing upper/lower bounds).

[36] Future work might compare patterns to SCAD as well if/when that dataset has been updated to include recent events.

[37] We estimate with Bayesian logistic regression $P(U_i = 1|\mathbb{1}[i \in \mathbf{V}], \nu_c, \tau_t) = logit^{-1}(\gamma\mathbb{1}[i \in \mathbf{V}] + \nu_c + \tau_t)$, where U_i indicates whether a given incident i was not previously captured by ACLED and the indicator variable captures whether that event is estimated to have involved violence. Country and year fixed effects are given by ν_c, τ_t, respectively. Predicted probabilities from alternative models with either country or year fixed effects are displayed in gray in Figure 2 and are effectively unchanged. We generate uncertainty estimates using quasi-Bayesian Monte Carlo simulation. Linear probability model results are consistent (see the accompanying R code), which we generate given possible incidental parameter biases that fixed effects can introduce in logistic regression.

[38] See the accompanying R code. This speaks to a more general possible use of the photo/video records: they might be used not only to reduce the number of systematically undercovered events, they might be incorporated into imputation efforts intended to more generally estimate overall levels of underreporting.

[39] We compare only the attacks logged by each dataset meeting our inclusion criteria. We exclude datasets' attacks from analysis based on their classification of actors involved and other key terms. In a limited number of cases, ambiguities in event details could potentially lead to events that were indeed captured by these datasets being dropped. However, we do not believe that we systematically under-count dataset event coverage of any particular type of event.

[40] As described in A.3, to determine whether a given event captured by the journalists was included in ACLED, members of our research team manually inspected each event. In some cases, differences in coordinates, dates, or description of the cause(s) and nature of the event between the events reported by the journalists and those included in ACLED made it difficult to determine whether an event reported by the journalists was indeed newly identified. In these cases, we create two datasets: one in which such cases are assumed to be newly identified and one in which they are not. We then calculated the statistics reported in this section using both datasets in order to produce plausible upper and lower bounds.

[41] These numbers are conservative as we subset only to those attacks reported by the journalist in which insurgent and separatist force involvement is described. Other potentially responsive cases are dropped from this calculation.

[42] Beitbridge, Gutu, Hwedza, Marondera, Mutoko, Nyanga.

[43] Nevertheless, this may change. We understand from internal discussions with one news agency that efforts are underway to explore providing specific spatial details extracted from photos/videos.

[44] For instance, with greater photo/video coverage, efforts to estimate crowd sizes, participants, whether or not violence occurred, etc. may benefit from the multiple resources

[45] For instance, ACLED reports using Arabic, but not Hebrew language sources to collect data on Israel/Palestine ACLED (2020).

[46] Please see the accompanying R code for a description of the UNMISS-GED and UNMISS-ACLED comparisons.

References

AccessNow, 2016. Keep It On Global Internet Outage Data, 2016-2022, Technical report.

ACLED, 2020. Israel/Palestine Sourcing Profile, `https://acleddata.com/acleddatanew/wp-content/uploads/2021/11/ACLED_Israel-Palestine-Sourcing-Profile_April-2020.pdf`.

AFP Forum, n.d..
URL: *www.afpforum.com*

AP Newsroom, n.d..
URL: *https://newsroom.ap.org/editorial-photos-videos*

Aragao, C., 2019. War on a Thumb Drive: How SigAct Data Could Transform Care for Veterans.

Aronson, J. D., 2018. Computer Vision and Machine Learning for Human Rights Video Analysis: Case Studies, Possibilities, Concerns, and Limitations, *Law & Social Inquiry* **43**(4), 1188–1209.

Baum, M. A. and Zhukov, Y. M., 2015. Filtering Revolution: Reporting Bias in International Newspaper Coverage of the Libyan Civil War, *Journal of Peace Research* **52**(3), 384–400.

Behlendorf, B., Belur, J. and Kumar, S., 2016. Peering through the Kaleidoscope: Variation and Validity in Data Collection on Terrorist Attacks, *Studies in Conflict & Terrorism* **39**(7-8), 641–667.

Berman, E., Felter, J. H. and Shapiro, J. N., 2018. Small Wars, Big Data, *in* Small Wars, Big Data, Princeton University Press.

Berman, E., Shapiro, J. N. and Felter, J. H., 2011. Can Hearts and Minds be Bought? The Economics of Counterinsurgency in Iraq, *Journal of Political Economy* **119**(4), 766–819.

Bernauer, T., Böhmelt, T., Buhaug, H., Gleditsch, N. P., Tribaldos, T., Weibust, E. B. and Wischnath, G., 2012. Water-Related Intrastate Conflict and Cooperation (WARICC): a New Event Dataset, *International Interactions* .

Birch, S. and Muchlinski, D., 2020. The Dataset of Countries at Risk of Electoral Violence, *Terrorism and Political Violence* **32**(2), 217–236.

Birnir, J. K., Laitin, D. D., Wilkenfeld, J., Waguespack, D. M., Hultquist, A. S. and Gurr, T. R., 2018. Introducing the AMAR (All Minorities at Risk) Data, *Journal of Conflict Resolution* **62**(1), 203–226.

Bloomberg, n.d.. Bloomberg Media Distribution. [Accessed 18-Aug-2023].
URL: *https://www.bloomberg.com/distribution*

Bodnaruk Jazayeri, K., 2016. Identity-Based Political Inequality and Protest: The Dynamic Relationship Between Political Power and Protest in the Middle East and North Africa, *Conflict Management and Peace Science* **33**(4), 400–422.

Boschee, E., Lautenschlager, J., OBrien, S., Shellman, S., Starz, J. and Ward, M., 2015. ICEWS Coded Event Data.

Bowie, N. G., 2017. Terrorism Events Data: An Inventory of Databases and Data Sets, 1968-2017, *Perspectives on Terrorism* **11**(4), 50–72.

Bueno de Mesquita, E., Fair, C. C., Jordan, J., Rais, R. B. and Shapiro, J. N., 2015. Measuring Political Violence in Pakistan: Insights from the BFRS Dataset, *Conflict Management and Peace Science* **32**(5), 536–558.

Campbell, E., 2023. Mahsa Amini and the Future of Internet Repression in Iran.

Chenoweth, E., Hendrix, C. S. and Hunter, K., 2019. Introducing the Nonviolent Action in Violent Contexts (NVAVC) Dataset, *Journal of Peace Research* **56**(2), 295–305.

Chenoweth, E. and Lewis, O. A., 2013. Nonviolent and Violent Campaigns and Outcomes (NAVCO) Data project, Version 2.0, Campaign Year Data, Codebook, *Josef Korbel School of International Studies. University of Denver* .

Choi, S.-W., 2010. Fighting terrorism through the rule of law?, *Journal of Conflict Resolution* **54**(6), 940–966.

Clarke, K., 2021. Which Protests Count? Coverage Bias in Middle East Event Datasets, *Mediterranean Politics* pp. 1–27.

Collier, P., Hoeffler, A. and Söderbom, M., 2004. On the Duration of Civil War, *Journal of Peace Research* **41**(3), 253–273.

Condra, L. N., Long, J. D., Shaver, A. C. and Wright, A. L., 2018. The Logic of Insurgent Electoral Violence, *American Economic Review* **108**(11), 3199–3231.

Cook, S. J., Blas, B., Carroll, R. J. and Sinha, S., 2017. Two Wrongs Make a Right: Addressing Underreporting in Binary data from Multiple Sources, *Political Analysis* **25**(2), 223–240.

Croicu, M. and Eck, K., 2022. Reporting of Non-Fatal Conflict Events, *International Interactions* **48**(3), 450–470.

Croicu, M. and Kreutz, J., 2016. Communication Technology and Reports on Political Violence: Cross-National Evidence Using African Events Data, *Political Research Quarterly* **70**, 19–31.

Crost, B., Felter, J. and Johnston, P., 2014. Aid Under Fire: Development Projects and Civil Conflict, *American Economic Review* **104**(6), 1833–1856.

Daskin, J. H. and Pringle, R. M., 2018. Warfare and wildlife declines in africa's protected areas, *Nature* **553**(7688), 328–332.

Davenport, C. and Ball, P., 2002. Views to a Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977-1995, *Journal of Conflict Resolution* **46**(1), 427–450.

Davies, S. E. and True, J., 2017. The Politics of Counting and Reporting Conflict-Related Sexual and Gender-based Violence: the Case of Myanmar, *International Feminist Journal of Politics* **19**(1), 4–21.

Daxecker, U., Amicarelli, E. and Jung, A., 2019. Electoral Contention and Violence (ECAV): A New Dataset, *Journal of Peace Research* **56**(5), 714–723.

De Rouen Jr, K. R. and Sobek, D., 2004. The Dynamics of Civil War Duration and Outcome, *Journal of Peace Research* **41**(3), 303–320.

Demarest, L. and Langer, A., 2018. The Study of Violence and Social Unrest in Africa: A Comparative Analysis of Three Conflict Event Datasets, *African Affairs* **117**(467), 310–325.

Dietrick, N. and Eck, K., 2020. Known Unknowns: Media Bias in the Reporting of Political Violence, *International Interactions* **46**(6), 1043–1060.

Donnay, K., Dunford, E. T., McGrath, E. C., Backer, D. and Cunningham, D. E., 2019. Integrating Conflict Event Data, *Journal of Conflict Resolution* **63**(5), 1337–1364.

Eck, K., 2012. In Data we Trust? A Comparison of UCDP GED and ACLED Conflict Events Datasets, *Cooperation and Conflict* **47**(1), 124–141.

Al Jazeera, 2023. Peru Anti-Government Protesters Clash with Police in Puno, Al Jazeera (January).

EPA Images, n.d.. Epa. [Accessed 18-Aug-2023].
URL: *https://epaimages.com/search.pp0*

Fearon, J. D. and Laitin, D. D., 2003. Ethnicity, Insurgency, and Civil War, *American Political Science Review* **97**(1), 75–90.

Fjelde, H. and Höglund, K., 2022. Introducing the Deadly Electoral Conflict Dataset (DECO), *Journal of Conflict Resolution* **66**(1), 162–185.

Fjelde, H., Hultman, L., Schubiger, L., Cederman, L.-E., Hug, S. and Sollenberg, M., 2021. Introducing the Ethnic One-Sided Violence Dataset, *Conflict Management and Peace Science* **38**(1), 109–126.

Fortna, V. P., 2015. Do terrorists win? rebels use of terrorism and civil war outcomes, *International Organization* **69**(3), 519–556.

Gibilisco, M. and Steinberg, J., 2022. Strategic Reporting: A Formal Model of Biases in Conflict Data, *American Political Science Review* pp. 1–17.

Gineste, C. and Savun, B., 2019. Introducing POSVAR: A Dataset on Refugee-Related Violence, *Journal of Peace Research* **56**(1), 134–145.

Gleditsch, K. S., Metternich, N. W. and Ruggeri, A., 2014. Data and Progress in Peace and Conflict Research, *Journal of Peace Research* **51**(2), 301–314.

Hand, D. J., 2020. *Dark Data: Why What You Don't Know Matters*, Princeton University Press.

Hoeffler, A., Kaiser, F., Pfeifle, B., Risse, F. et al., 2022. Tracking the SDGs: A Methodological Note on Measuring Deaths Caused by Collective Violence, *Economics of Peace and Security Journal* **17**(2), 32–46.

Hoffmann, R., Dimitrova, A., Muttarak, R., Crespo Cuaresma, J. and Peisker, J., 2020. A meta-analysis of country-level studies on environmental change and migration, *Nature Climate Change* **10**(10), 904–912.

Hou, D., Gaibulloev, K. and Sandler, T., 2020. Introducing Extended Data on Terrorist Groups (EDTG), 1970 to 2016, *Journal of Conflict Resolution* **64**(1), 199–225.

Hussain, B., 2023. Kashmir Registers Highest Number of Internet Restrictions Globally.

Ives, B. and Lewis, J. S., 2020. From Rallies to Riots: Why Some Protests Become Violent, *Journal of Conflict Resolution* **64**(5), 958–986.

Kalyvas, S. N., 2004. The Urban Bias in Research on Civil Wars, *Security Studies* **13**(3), 160–190.

Klein, G. R. and Regan, P. M., 2018. Dynamics of Political Protests, *International Organization* **72**(2), 485–521.

LaFree, G., 2010. The Global Terrorism Database (GTD) Accomplishments and Challenges, *Perspectives on Terrorism* **4**(1), 24–46.

LaFree, G. and Dugan, L., 2007. Introducing the Global Terrorism Database, *Terrorism and Political Violence* **19**(2), 181–204.

Laktabai, V. K., 2020. Using GIS to Assess the Risk of Terrorism: a Case Study of Garissa County, PhD thesis, University of Nairobi.

Larreguy, H., Lucas, C., Marshall, J. and Riaz, Z., 2020. Dont Read All About It: Drug Trafficking Organizations and Media Reporting of Violence in Mexico.

Leetaru, K. and Schrodt, P. A., 2013. GDELT: Global Data on Events, Location, and Tone, 1979–2012, *in* ISA Annual Convention, Vol. 2, Citeseer, pp. 1–49.

Lindberg Bromley, S., 2018. Introducing the UCDP Peacemakers at Risk Dataset, Sub-Saharan Africa, 1989–2009, *Journal of Peace Research* **55**(1), 122–131.

Manacorda, M. and Tesei, A., 2020. Liberation technology: Mobile phones and political mobilization in africa, *Econometrica* **88**(2), 533–567.

Miller, E., Kishi, R., Raleigh, C. and Dowd, C., 2022. An Agenda for Addressing Bias in Conflict Data, *Scientific Data* **9**(1), 593.

Minoiu, C. and Shemyakina, O. N., 2014. Armed conflict, household victimization, and child health in côte divoire, *Journal of Development Economics* **108**, 237–255.

Mroszczyk, J. and Abrahms, M., 2021. Terrorism in Africa: Explaining the Rise of Extremist Violence Against Civilians, *International Relations* .

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.

NARA, 2023. Electronic Records Relating to the Vietnam War — archives.gov, `https://www.archives.gov/research/military/vietnam-war/electronic-data-files#:~:text=During%20the%20Vietnam%20War%2C%20the,(IBM)%20developed%20the%20system`. [Accessed 30-Jun-2023].

Obrien, S. P., 2010. Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research, *International Studies Review* **12**(1), 87–104.

Osorio, J. and Beltr´an, A., 2019. Enhancing the Detection of Criminal Organizations in Mexico Using ML and NL, pp. 1–7.

Otto, S., 2013. Coding One-Sided Violence from Media Reports, *Cooperation and Conflict* **48**(4), 556–566.

O'Loughlin, J., Linke, A. M. and Witmer, F. D., 2014. Effects of Temperature and Precipitation Variability on the Risk of Violence in Sub-Saharan Africa, 1980–2012, *Proceedings of the National Academy of Sciences* **111**(47), 16712–16717.

Pape, R. A., Rivas, A. A. and Chinchilla, A. C., 2021. Introducing the New CPOST Dataset on Suicide Attacks, *Journal of Peace Research* **58**(4), 826–838.

Radeva, E., 2021. The Potential for Computer Vision to Advance Accountability in the Syrian Crisis, *Journal of International Criminal Justice* **19**(1), 131–146.

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.

Reuters, n.d.. Reuters. [Accessed 18-Aug-2023].
URL: *https://www.reutersagency.com/en/content-types/pictures/?gad=1amp;gclid=Cj0KCQjw4s-kBhDqARIsAN-ipH0n7_lZy9naHHKczJDCVFOCCUmn4aULvDwvrlRWAbrQ1PoHL8B25ucaAt6XEALw$_w$cBamp;gclsrc =*
aw.ds

Salehyan, I., Hendrix, C. S., Hamner, J., Case, C., Linebarger, C., Stull, E. and Williams, J., 2012. Social Conflict in Africa: A New Database, *International Interactions* **38**(4), 503–511.

Schutte, S. and Kelling, C., 2022. A Monte Carlo Analysis of False Inference in Spatial Conflict Event Studies, *PLoS one* **17**(4), 1–22.

Schvitz, G., Rüegger, S., Girardin, L., Cederman, L.-E., Weidmann, N. and Gleditsch, K. S., 2022. Mapping the international system, 1886-2017: The CShapes 2.0 dataset, *Journal of Conflict Resolution* **66**(1), 144–161.

Sexton, R., 2016. Aid as a Tool Against Insurgency: Evidence from Contested and Controlled Territory in Afghanistan, *American Political Science Review* **110**(4), 731–749.

Shaver, A., Komal Sikka, K., Keltner, K., Azimioara, N., Manly, J., Marin, R., Milinic, S., Green, B., Shah, D., Barbour-Berson, S., Creach, C., Baradwaj, A., Fan, S., Cabezas, K., Sanders, H. and Goldstein, L., 2023. Global Patterns of Political Violence and Social Unrest, *Working Paper* .

Shaver, A. et al., 2022. News Media Reporting Patterns and our Biased Understanding of Global Unrest, Technical report, Empirical Studies of Conflict Project.

Solstad, S. U., 2023. Using satellite data to track Ukraine's counter-offensive. see also: https://github.com/TheEconomist/the-economist-war-fire-model.

Steinert-Threlkeld, Z. C., 2017. Spontaneous collective action: Peripheral mobilization during the arab spring, *American Political Science Review* **111**(2), 379–403.

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

Sutton, J., Butcher, C. R. and Svensson, I., 2014. Explaining Political Jiu-Jitsu: Institution-Building and the Outcomes of Regime Violence Against Unarmed Protests, *Journal of Peace Research* **51**(5), 559–573.

Tin, D., Granholm, F., Hart, A. and Ciottone, G. R., 2021. Terrorism-Related Chemical, Biological, Radiation, and Nuclear Attacks: a Historical Global Comparison Influencing the Emergence of Counter-Terrorism Medicine, *Prehospital and Disaster Medicine* **36**(4), 399–402.

UNMISS, 2021. Annual Brief on Violence Affecting Civilians, January – December 2020, `https://www.ohchr.org/sites/default/files/Documents/Countries/SS/unmiss_annual_brief_violence_against_civilians_2020.pdf`. [Accessed 30-Jun-2023].

UNMISS, 2022. Annual Brief on Violence Affecting Civilians, January – December 2021, `https://www.ohchr.org/sites/default/files/2022-03/unmiss_hrd_annual_brief_2021.pdf`. [Accessed 30-Jun-2023].

UNMISS, 2023*a*. Annual Brief on Violence Affecting Civilians, January – December 2022, `https://reliefweb.int/report/south-sudan/unmiss-annual-brief-violence-affecting-civilians-january-december-2022#:~:text=The%20brief%20reveals%20that%20while,2021%20to%201%2C674%20in%202022`. [Accessed 30-Jun-2023].

UNMISS, 2023*b*. Annual Brief on Violence Affecting Civilians, January – March 2023, `https://reliefweb.int/report/south-sudan/unmiss-brief-violence-affecting-civilians-january-march-2023`. [Accessed 30-Jun-2023].

Vogt, M., Bormann, N.-C., Rüegger, S., Cederman, L.-E., Hunziker, P. and Girardin, L., 2015. Integrating Data on Ethnicity, Geography, and Conflict: The Ethnic Power Relations Data Set Family, *Journal of Conflict Resolution* **59**(7), 1327–1342.

Von Borzyskowski, I. and Wahman, M., 2021. Systematic measurement error in election violence data: Causes and consequences, *British Journal of Political Science* **51**(1), 230–252.

Voors, M. J., Nillesen, E. E. M., Verwimp, P., Bulte, E. H., Lensink, R. and Soest, D. P. V., 2012. Violent conflict and behavior: a field experiment in burundi, *American Economic Review* **102**(2), 941–964.

Walter, B. F., 1997. The Critical Barrier to Civil War Settlement, *International Organization* **51**(3), 335–364.

Weidmann, N. B., 2015. On the Accuracy of Media-based Conflict Event Data, *Journal of Conflict Resolution* **59**(6), 1129–1149.

Weidmann, N. B., 2016. A Closer Look at Reporting Bias in Conflict Event Data, *American Journal of Political Science* **60**(1), 206–218.

Yedioth Ahronot, n.d.. [Accessed 18-Aug-2023].
URL: *https://www.ynet.co.il/home/0,7340,L-8,00.html*

Zhukov, Y. M. and Baum, M. A., n.d.. How Selective Reporting Shapes Inferences About Conflict, *Unpublished working paper. pages = 1–7, year = 2019, month = July, .*

Zhukov, Y. M., Davenport, C. and Kostyuk, N., 2017. Introducing xSub: A New Portal for Cross-National Data on Subnational Violence, *Journal of Peace Research* p. 0022343319836697.

A Appendix

Contents

A.1 Interviews with Foreign Affairs Professionals

For this and related research, we engage in a series of in-depth semi-structured interviews with foreign affairs professionals employed or previously employed by/within the United Nations system and other intergovernmental organizations; many of the world's largest international non-governmental organizations; major philanthropic organizations; and foreign affairs think tanks, amongst other organizations. During these interviews, we frequently asked whether they or other members of their organizations use conflict-event datasets derived from news reports.

A.2 Photo- and Video-Journalism Event Detection Specifics

To probe the plausibility of supplementing conflict event data with events detected through photo- and video-journalism, we accessed photo and video archives maintained by the *AP Newsroom* (n.d.) and photo archives maintained by the *AFP Forum* (n.d.). For this proof of concept, we focused specifically on comparing unique events identified through the photos and videos with the events tracked by ACLED (Raleigh et al., 2010) given that particular dataset's focus on protests and social unrest. We examined patterns of reporting across eleven countries: Brazil, Colombia, Egypt, France, Haiti, Myanmar, Nicaragua, Pakistan, Peru, South Africa, and Yemen.

We used Boolean search terms (e.g., "Riot" OR "Protest" OR "Demonstration") to identify relevant photos in the *AFP* Forum, a multimedia database and stored links to photos, dates, and event descriptions in a spreadsheet corresponding to each country case study.[1] We then cross-referenced our search results with *AP Newsroom*, adding events that *AFP* did not include to each spreadsheet.

With respect to the use of photos, we make several notes about our process. As described in the paper, we identified details related to responsive events from photos that clearly captured protests, demonstrations, riots, or other form of social gathering or unrest. Due to the site-specific differences between *AFP* and *AP*, we evaluated *AFP's* photo database to gather initial event photos before examining *AP's* for non-duplicate photos for that particular dataset. Specifically, we used the short description of the event in each platform's photo description to determine the nature of the event. We included only those events relevant to national (or international) politics. Though, we note our choice to define COVID-19 protests and demonstrations as politically motivated.

The incident location was derived from the city/country format that is associated with each photo. Typically, the event date could be determined from the date associated with the photo description.

To identify incidence from video, we followed a very similar method. Using the same Boolean search terms, we surveyed *AP Newsroom*, filtered for video content, and stored video ID numbers, dates, and event descriptions in the spreadsheet corresponding to each country case study. We then cross-referenced our search results with the photo effort, including only videos that did not capture an event previously identified by a photo entry.

In many cases, single events were photographed multiple times. Given that our focus is on identifying individual events, in such cases we arbitrarily chose one of the several possible photos, using that to confirm the event. We classified a photo as capturing an independent event as one with a unique date, location, and central issue. Subsequently, we declared a duplicate photo as one with all three exact identifiers. For example, there was a case in Nicaragua where a pro-government rally[2] took place on the same day and in the same location as an anti-government rally[3] leading us to count these as two discrete events. In cases in which protests spanned multiple days (or at least appeared to), we treated each day as a separate event. For instance, in Brazil, protests of the government's health policies appear in photos from October 08, 15, and 18 of 2021. Each of these protests were treated as unique events although to corresponded to a broader coordinated focus during that period. This methodology is consistent with how ACLED disaggregates their events.[4]

Our subsequent effort with video mirrored our photo-journalism effort, using the same Boolean search terms coupled with a country identifier (e.g. "Brazil" AND ("protest" OR "riot" OR "demonstration")) through *AP*'s archive. Each video provided us with a location, source information, and an ID. Additionally, there were shotlists and storylines provided which served as our event description from which the event date came. These details were all recorded into the master sheet containing photo entries, with the caveat that a video entry was only included if the event was not previously captured by a photo. Once we comprehensively captured these events, we then cross-checked against ACLED using the same methodology as above.[5]

After our data was gathered across each country's time series, we assigned a violence indicator to each identified incident. Using the photo/video descriptions, photos and videos themselves, *Google* searching the event for additional details, and/or, for cases included in ACLED, using ACLED's details related to the presence or absence of violence, we assigned a value of "1" to events in which violence was known or suspected to have been involve, and we assigned a "0" otherwise. Our violence threshold was low; including anything that could reasonably cause bodily harm and/or death (e.g. arson, homemade mortars, rock throwing, etc.) and the presence of weapons from non-state actors which indicated intent of violence.

A.3 Direct Information Sharing From Journalists

In an effort to identify the set of events that journalists learn about in the course of their reporting (but which may or may not ultimately be published as news stories), we entered into independent contracts with seven freelance journalists all of whom have written for major international news media outlets. These journalists focused on tracking events within Mozambique, Pakistan, Peru, South Africa, and Zimbabwe.

Per the terms of agreement, the journalists reported to us all incidents of political violence and social unrest they learned about in the course of their work regardless of whether or not they considered the events newsworthy or likely to be published.

Below, we first detail the specific sub-national areas and time periods theses journalists covered with respect to their country of coverage. We then provide the general scope of work that guided engagement with the journalists with whom we contracted. The scope of work outlines the role of journalist during the reporting period and the specific types of incident details to track as they are able to. It also defines qualifying and non-qualifying events and provides clear directions that all qualifying events should be reported regardless of any characteristics that might otherwise render them less likely to be published.

A.3.1 Temporal and Geographic Specifics

1. **Mozambique:** One journalist reported on this country. Specifically, they reported on events that took place within the Manica, Maputo, and Cabo Delgado provinces. They covered events that took place between 2023-04-15 and 2023-06-05.
2. **Pakistan:** Two journalists reported on this country. The first journalist reported on events that took place within the Balochistan province. The second covered events within the Khyber Pakhtunkhwa and Sindh provinces. They covered events that took place between 2023-05-01 and 2023-05-31 (Balochistan) and 2023-04-25 and 2023-05-24 (Khyber Pakhtunkhwa and Sindh).
3. **Peru:** One journalist reported on this country. They reported on events affecting the entire country. They covered events that took place between 2023-04-20 and 2023-06-02.
4. **South Africa:** One journalist reported on this country. They reported on events affecting the entire country. They covered events that took place between 2023-05-10 and 2023-06-08.
5. **Zimbabwe:** two journalists reported on this country. The first journalist reported on events that took place within the Masvingo, Mashonaland East, Mashonaland West, Mashonaland Central, and Manicaland provinces. The second covered events within the Harare, Bulawayo, Midlands, Matabeleland South, and Matabeleland North provinces. They covered events that took place between 2023-04-20 and 2023-05-19.

A.3.2 Scope of Work

The following language was supplied to journalists in advance of their reporting efforts –

The Political Violence Lab seeks to collaborate with journalists working across the world in countries experiencing political violence/social unrest to learn more about the number, type, and nature of discrete instances of political violence/social unrest occurring in the countries/regions covered by those journalists. Below, we define more formally what type of incidents qualify as political violence and/or social unrest.

The role of the independent contractor is to document and report all incidents of political violence and/or social unrest that they learn about during the period under which they are under contract, which shall run for the equivalent of one complete month (30 to 31 consecutive days) to fall sometime between the period of March 15, 2023 and July 1, 2023.

The independent contractor will specifically report individual incidents of political violence and social unrest that 1) occur during this period and/or 2) that they learn about during this period even if they occurred before the start of the reporting period. Specifically, they will enter, to the greatest extent reasonably possible, details relating to each individual incident into a spreadsheet containing various columns specifying the specific data types to be provided for each event. The geographic focus of reporting is [DETAILS SPECIFIC TO EACH JOURNALIST ENTERED HERE]. The base spreadsheet "[JOURNALIST COUNTRY] Incident Tracking" that the contractor will use is also attached to this "Scope of Work" description.

[JOURNALIST NAME] will attempt to identify to the best of their ability the following details associated with each relevant incident:

1. The date of the incident.
2. The end date of the incident (in cases in which the event (e.g. a protest) lasts for more than one day).
3. Its approximate time of occurrence (expressed in local time).
4. Its location (as accurately identified as possible). Please see the "Latitude and Longitude" sheet of the spreadsheet for additional details on how location should be entered. When latitude-longitude coordinate estimates can be supplied, this is strongly preferred, and a description of how to generate these coordinates quickly and easily with Google Maps is described in that sheet.
5. Confidence level that the event actually occurred. In cases where the contractor learns

about an incident but is unable to verify its authenticity, they should indicate to the extent reasonably possible their confidence level. (If the contractor is confident that the event occurred, but unclear about particular details, they may express this in the notes section of the spreadsheet.)

6. In the case of violent attacks (these columns are list in the color blue in the spreadsheet)–

   (a) The perpetrator of the attack:

      i. Perpetrator Identity: This refers to the broad category to which the perpetrator belongs (e.g. police forces, sub-state militants)

      ii. Perpetrator Name: This refers to the specific identity of the the perpetrator (e.g. name of the sub-state militant organization)

   (b) The target(s) of the attack:

      i. Target Identity: This refers to the broad category to which the target belongs https://www.overleaf.com/project/63bf54088def5d7a52f46318(e.g. civilians; government forces). When there are multiple targets (e.g. a bombing that harms both security forces and civilians), all targets should be listed. If it is clear that one or more targets were the intended targets and the other targets were bystanders, then indications of this would be extremely helpful as well.

      ii. Target Name: This refers to the specific identity of the the target (e.g. name of the sub-state militant organization; if civilians, do they have a specific affiliation? – e.g. students of a particular university or members of a particular political movement).

      iii. Target Type: The target(s) of the attack by type (e.g. check point; power line; government building).

   (c) Method(s) of Attack. This refers to the method or type of attack carried out – e.g. shooting; shelling.; grenade attack, to include the perpetration of sexual violence/attack when done in a political context.

   (d) Estimates of the number of individuals harmed:

      i. Number estimated killed.

ii. Number estimated wounded.

7. In the case of social unrest, protest activity, riot, activity, etc. (these columns are list in the color red in the spreadsheet)–

   (a) Type: This indicates whether the social unrest is a protest/demonstration/rally; a march; or a riot.

   (b) An estimate of the number of individuals involved in the protest, riot, etc.

   (c) Target: In the case of riots where violence is directed against individuals/infrastructure, an indication of what was targeted (e.g. automobiles; government offices)

   (d) The purpose of the protest, demonstration, etc. (e.g. protesting a new government policy).

   (e) The actor(s) involved in the unrest. In cases in which counter protests are also held, both parties should be listed. In cases in which protests, demonstrations, etc. lead to violence between multiple parties (e.g. government forces attack protesters), all relevant parties should be listed.

   (f) Whether the activity involved violence or not.

   (g) An estimate of the number of individuals involved wounded (if relevant).

   (h) An estimate of the number of individuals involved killed (if relevant).

The contractor will identify the details of such events through their usual channels/processes for acquiring information as a professional journalist. At the end of the 30 to 31 day consecutive period, the contractor will deliver the spreadsheet, completed as comprehensively as reasonably possible across the relevant columns, to Professor Shaver by e-mail at ashaver@ucmerced.edu.

Qualifying Incidents: What type of incidents should be reported and which should not?

Political Violence: Relevant incidents include, but are not limited to, attacks carried out by government forces, sub-state militants, civilians, or other parties that relate directly or indirectly to political affairs, broadly defined, in the country/region. Relevant incidents include attacks related to the country's politics, including tensions related to national, ethnic, racial, religious, sectarian or other cleavages. They may also relate to other countries' politics (e.g. fighting amongst refugee communities reflecting political conditions in the countries from which the refugees fled) as well as international politics (e.g. violence intended to influence or otherwise motivated by or connected to the political affairs of other countries, intergovernmental organizations, international non-governmental organizations, etc.). Examples of event types include, but are not limited to, small arms fire (e.g. shootings with rifles, pistols); indirect fire (e.g. rocket fire, mortar fire); bombings (e.g. improvised explosive device attacks; mine explosions); non-firearm assaults (e.g. stabbings, rock throwing).

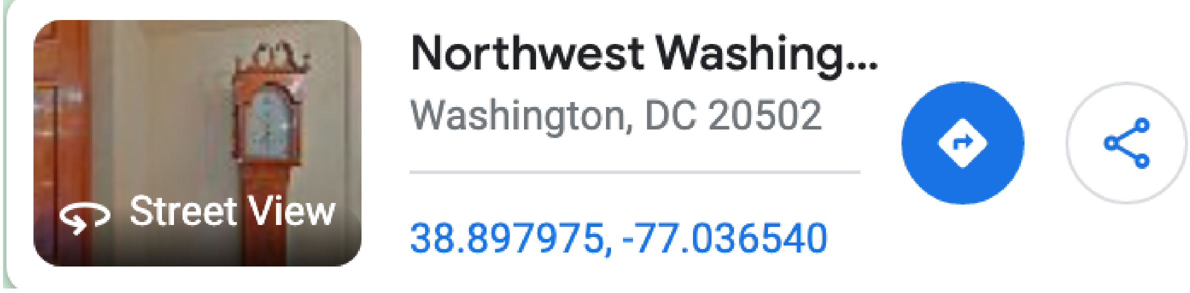
Social Unrest: Relevant incidents also include violent and non-violent incidents of social unrest related to politics (following the same broad definition offered above). Examples of social unrest include, but are not limited to, both peaceful and violent protests, peaceful and violent demonstrations, and rioting.

Please note that actual violence and/or harm to individuals does not need to occur in order for an event to be included. For instance, a protest that does not involve violence qualifies. Similarly, a shooting or bombing that misses its target, resulting in no injuries or fatalities, also qualifies. Events that do not cause injury or death are as relevant as those that do, and equal attention should be given to both classes of events. Attacks on automobiles, property, infrastructure, etc. that do not harm individuals are also relevant.

Of critical importance, all relevant events should be reported regardless of whether or not the journalist would consider them newsworthy or likely to be published. For instance, events should be reported regardless of where they occur. Events should be reported regardless of who carries them out and regardless of who is targeted (or was the intended target in cases of failed attacks), etc.

Non-Qualifying Incidents: Instances of violence/social unrest that **are not** relevant include personal disputes between individuals (e.g. business disputes leading to violence that are not related to politics; national, ethnic, racial, religious, sectarian or other cleavages; etc.); domestic abuse; violence between fans at sporting events (unless there are political undercurrents/motivations to that violence); etc. (If in doubt, the contractor should include the incident and leave a note clearly indicating any concern about its inclusion in the notes section.)

Continuous Incidents: In some cases, incidents may occur over multiple days. For instance, a protest may last for several days. A building seized by rebel forces may be under



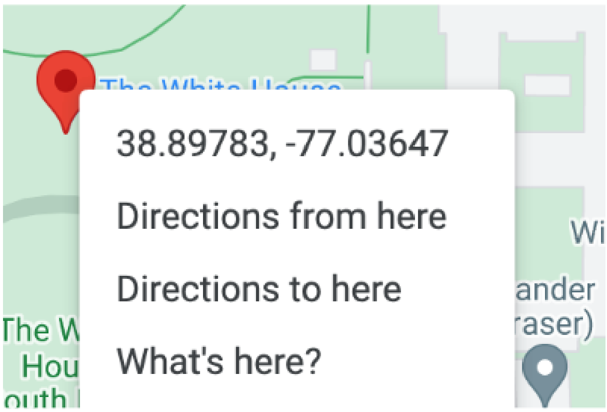
Northwest Washing...
Washington, DC 20502
38.897975, -77.036540

siege for days. In such cases, the contractor should use the end date spreadsheet column to clarify when the event ends and use the notes column to clarify the details of the event to the extent possible. When multiple protests, demonstrations, etc. take place in coordination (e.g. protests in separate cities are planned for the same date/time), all such protests/demonstrations should be listed as separate incidents.

Location Specification: To the extent possible, the contractor is asked to supply latitude-longitude coordinates (or their best estimate thereof) to specify the location of events. Below, we supply instructions for how latitude-longitude coordinate estimates can be quickly and easily pulled from Google Maps. If the contractor is unable to provide latitude-longitude coordinate estimates, they are asked to supply the location as accurately and closely as possible (e.g. address if available; otherwise neighborhood (if in town/city); otherwise town/city name; etc. If the event occurs in a rural location to which no town/city name can be given AND if latitude-longitude coordinates can not be estimated, then we kindly ask the contractor to specify the location in terms of estimated distance and direction from the nearest town/city – e.g. “30 km directly south of [town name]”.

Latitude and Longitude Instructions: Whenever possible, latitude-longitude coordinate estimates are strongly preferred to location name. You may generate latitude-longitude coordinate estimates quickly and easily on Google Maps. To do so, simply search for your desired location on Google Maps. Once you have found the location on the map, use the right click button on your mouse or trackpad, or press the control button on your keyboard while you click on the map.

A single click on the mouse will yield a small gray pin and a white box with blue coordinates, which will appear at the bottom of the screen. These coordinates are listed (latitude, longitude). These coordinates may be easily highlighted and copied.



38.89783, -77.03647

Directions from here

Directions to here

What's here?

Using the control button and clicking on the map, the coordinates will appear at the site where you clicked. Again, these coordinates are listed (latitude, longitude). Clicking on the coordinates as they appear on the list will copy them to your clipboard and then enter them into the spreadsheet.

A.3.3 Matching Process between Journalists' and ACLED's Datasets

Our methodology to determine whether the local journalists recorded novel events of violence and social unrest not captured by ACLED (Raleigh et al., 2010) data in Mozambique, Pakistan, Peru, South Africa, and Zimbabwe is as follows.[6] First, to ensure that we did not miss events that were mutually covered – events that ACLED may have recorded a few days before or after the local journalists did so – we took note of the first and last events' dates that the local journalists recorded for each of the five countries. For each country, we used ACLED data ranging from three days before each journalist's recorded first event date to three days after each journalist's recorded last event date.[7] Next, two authors on the project manually inspected each event, assigning a "0" if the local journalists recorded a novel event of violence and social unrest not captured by ACLED and "1" if ACLED recorded the journalists' non-unique incident. An assignment of "0.5" indicated events ACLED may or may not have recorded: the uncertainty in this hand-coding arose from small yet considerable differences in coordinates, dates, or description of the cause(s) and nature of the event. After evenly dividing the datasets and completing this hand-coding, the two authors then checked each other's work as an additional check.

Throughout this process, we sought to use the ACLED data responsibly and in good faith, referring to ACLED's codebook for "Event Types and Sub-Event Types," "Event Aggregation," "Civilian Targeting," "Actors," "Event Geography," "Event Time," "Reported Fatalities," and "Notes." We then compared these variables to the local journalists' records including, but not limited to, each event's type, method of attack, perpetrator and target identities, types, and names, location, latitude and longitude coordinates, start and end dates, time, number of injuries and fatalities, and notes.

The instructions to the journalists focused on reporting incidents of political violence and social unrest. Thus, from the ACLED data, we removed all cases of "strategic developments" except for those cases that potentially overlapped with our definition of political violence/social unrest. Specifically, if cases of property destruction; looting; or strikes were tracked by ACLED for these countries over the reporting period, we included them as well.[8] We did not compare arrests and weapons discoveries as we did not ask the journalists to systematically track such events. Thus, when such events were nevertheless reported by journalists, we dropped these incidents from the journalists' data. In some instances, journalists also reported other details not pertaining to specific incidents of political violence/social unrest. We dropped these instances from the journalists' data as well to maintain clear comparisons of like events. Readers interested in the specific set of inclusions/exclusions made for the purposes of establishing final sets of comparable events should consult the accompanying R code for those details.

For the spatial plots shown in the paper, we matched the relevant incidents from the journalist and ACLED datasets to sub-national administrative unit boundaries using the latitude-longitude coordinate pairs available in each of these datasets. For sub-national country boundaries, we used HDX (2023*a*,*b*,*c*,*d*,*e*).

A.4 Local Data Sources

We limited our effort piloting the use of local data sources to tracking extrajudicial violence in Israel/Palestine from 2000-March 2023.

Our definition of political violence matches those used by three existing datasets, which all include instances of extrajudicial violence with political aims, regardless of whether the incident resulted in fatalities ACLED (2021); GTD (2021); RAND (n.d.). We included incidents of political violence that could realistically cause injury or meaningful property damage. With this definition, we excluded cases including only hate speech and/or graffiti. For the purposes of our research, assumed inter-ethnic attacks had political motives, unless otherwise specified. We omitted any intra-ethnic violence that did not have clear political motivators (domestic violence, tribal/gang violence). We also excluded any instances of state-sponsored violence or attacks on military targets, such as attacks on soldiers on patrol. However, we included all instances of non-state violence, such as organized militant groups. Accordingly, we included attacks perpetrated by the Hamas militant group before it became the elected political authority in Gaza, but omitted attacks perpetrated after its rise to power in 2006.

Please refer to replication files to review precise keywords used to determine inclusion and classify types of attacks. For each existing dataset, we downloaded the entire set of available data from Israel/Palestine, and filtered for relevant attacks using available R code.

Notes

[1] Country names were not included in the search terms as *AFP Forum* has a separate option to filter content by country.

[2] 000_15$H7ON$. (Accessible via http://u.afp.com/iBMV as of April 7, 2023.)

[3] 000_15$H7NV$. (Accessible via http://u.afp.com/iBM9 as of April 7, 2023.)

[4] ACLED outlines their definitions of unique events here, under the "How are events disaggregated?" section, accessible here: https://acleddata.com/resources/quick-guide-to-acled-data/s7

[5] ALCED data used in these analyses was accessed on or about June 12, 2023 for the countries of Brazil, Colombia, Egypt, France, Haiti, Myanmar, Nicaragua, Pakistan, Peru, South Africa, and Yemen, covering event types: "Protests", "Riots", and "Violence against civilians", and subsetting to cases in which the disorder_type variable was equal to either "Demonstrations" or "Political violence; Demonstrations" or "Political violence" in the case of event type "Riots". Although these categories cover the vast majority of social unrest tracked by ACLED, on or about August 25, 2023, we also accessed ACLED data for event type "Strategic developments" for the purposes of identifying any other incidents of social unrest that may have been tracked under this event category across these same countries. ACLED data (in both instances) was pulled for the dates from January 01, 2017 through December 31, 2022. Though, the specific range of dates used in our comparisons varies across countries. Given the diverse set of activities that ACLED tracks under "Strategic developments", we subsetted these events, focusing specifically on those involving "Looting/property destruction" or "Other" (based on ACLED's "sub_event_type" variable) for these countries across their respective time periods. From there, we subsetted further, removing those events we determined not to involve social unrest until reaching a set of ≈400 potentially relevant events. Members of the research team then inspected these to determine whether

1) they involved social unrest and, if so, 2) the use of violence. We concluded that 32 of these events involved social unrest, and these were added to the broader set of ACLED events involving social unrest for the purposes of our analysis. To determine whether each of these 32 events involved violence, we consulted the “notes” section of ACLED’s data and assessed based on the details presented there. For the set of events identified from the “Protests”, “Riots”, and “Violence against civilians” event types, we relied on whether an event was listed as “peaceful” based on the “Peaceful protest” sub_event_type variable. For additional details of our process, we refer readers to the accompanying R code.

[6] ALCED data for this exercise used in the analyses of Mozambique, Pakistan, Peru, South Africa, and Zimbabwe was accessed on or about June 12, 2023. ACLED data was pulled for the countries of Mozambique, Pakistan, Peru, South Africa, and Zimbabwe, covering event types: “Battles”, “Explosions/Remote violence”, “Protests”, “‘Riots”, “Strategic developments”, and “Violence against civilians”. (Though, as we note below, we removed many instances of “Strategic developments” for the purposes of establishing comparability across datasets.) ACLED data was pulled for the dates from March 01, 2023 through June 09, 2023. However, the exact dates within this range used for each country and varies based on the respective coverage by the journalist(s) reporting on each country. See the accompanying R code for the precise details.

[7] The exception were cases in which the last day of journalist coverage coincided with the last day of ACLED coverage.

[8] These specific categories come from www.acleddata.com/wp-content/uploads/2017/12/Strategic-Developments_FINAL.pdf

References

ACLED, 2021. ‘Armed Conflict Location Event Data Project (ACLED) Codebook’, `https://acleddata.com/acleddatanew/wp-content/uploads/2021/11/ACLED_Codebook_v1_January-2021.pdf`.

AFP Forum, n.d..

URL: *www.afpforum.com*

AP Newsroom, n.d..

URL: *https://newsroom.ap.org/editorial-photos-videos*

GTD, 2021. ‘CODEBOOK: METHODOLOGY, INCLUSION CRITERIA, AND VARIABLES’, `https://www.start.umd.edu/gtd/downloads/Codebook.pdf#page12`.

HDX, 2023*a*. ‘Mozambique - Subnational Administrative Boundaries’, `https://data.humdata.org/dataset/cod-ab-moz`.

HDX, 2023*b*. 'Pakistan – Subnational Administrative Boundaries', `https://data.humdata.org/dataset/cod-ab-pak`.

HDX, 2023*c*. 'Peru – Subnational Administrative Boundaries', `https://data.humdata.org/dataset/cod-ab-per`.

HDX, 2023*d*. 'South Africa – Subnational Administrative Boundaries', `https://data.humdata.org/dataset/cod-ab-zaf`.

HDX, 2023*e*. 'Zimbabwe – Subnational Administrative Boundaries', `https://data.humdata.org/dataset/cod-ab-zwe`.

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.

RAND, n.d.. 'Database Scope', `https://www.rand.org/nsrd/projects/terrorism-incidents/about/definitions.html`.