18

# Conflict and Forced Migration: Social Media as Event Data

H. AKIN UNVER, AHMET KURNAZ



## Introduction: Defining the Violence-Migration Nexus

War has always been synonymous with large-scale human displacement. Although inter-state wars and large-scale global conflicts have been the main culprits of forced migration in history, more recent cases of human displacement have been triggered either by environmental problems (famine, drought, floods) or by civil wars and subnational violence, or an interaction of both. Since civil wars are increasingly being fought over the control of civilian populations, their impact on forced migration is often more pronounced compared to inter-state military disputes. Perhaps the worst such calamity of modern times, the Syrian Civil War, has produced upwards of 5.6 million refugees and 6.2 million internally displaced people.

Conflicts generate population displacement both because of their immediate violent effect, and due to their secondary after-effects that generate infrastructure destruction, homelessness, poverty, and lack of access to food, water, and security. Contemporary responses to mitigate forced migration have taken on two major forms. The first of those is the 'root causes' literature that explores social, political, and economic drivers of forced migration and seeks to build more sustainable and long-term improvements in target societies to minimise the likelihood of mass migration (Schmeidl, 2001). The second line of response is the protection, assistance, and aid camp, which seeks to create emergency response mechanisms to protect populations from violence, and deploy safeguards and urgent aid to prevent displacement. This second 'urgent response' camp, namely peacekeepers, aid workers, and emergency assistance professionals, have increasingly relied on situational awareness data, which would supply them with patterns of violence (Thobane et al., 2007; Harrington, 2005).

*Proceedings of the British Academy*, **251**, 381–400, © The British Academy 2022.

Although not all conflict-exposed populations migrate (Ibáñez and Vélez, 2008), organised violence, especially recurring violence that incurs sustained damage to the livelihood of inhabitants, is among the primary predictors of forced migration (Czaika and Kis-Katos, 2009). Although war and sustained conflict, without direct violence, may also generate the groundwork for forced migration due to its aftershock effects on poverty, displacement, and grievances, acute violence often serves as a substantial variable that affects the locals' calculus to stay or leave (Moore and Shellman, 2004). In the presence of indiscriminate violence, and exchange of territory where the replacing power has no immediate plans to establish law and order, incentives to leave towards an unknown outcome may outweigh the more explicit benefits of staying (De Mesquita et al., 2005). Earlier studies have demonstrated that conflict-related forced migration creates an outflow towards regions with better security, law and order, as well as better access to employment, social networks, and pre-existing contacts (Dustmann and Kirchkamp, 2002; Pellegrini and Fotheringham, 2002; Klabunde and Willekens, 2016).

However, conflict and violence serve as a more important variable in the study of forced migration than simply being a cause of it. Such violence may not only affect the migration route and tempo of the migrants but may also significantly affect the level of further violence that the migrants suffer at the hands of the warring parties (Salt and Stein, 1997). The tendency to migrate is more significant with populations that have a longer planning horizon, as younger people may become a particular target for combatants when a particular area changes from the jurisdiction of one side to another (Stark and Levhari, 1982; Katz and Stark, 1986; Todaro, 1969). In other words, direct violence or threat of violence both mitigates the effect of uncertainty on populations making a migration choice and strengthens the risk-taking behaviour of populations living close to flashpoint areas.

Because armed conflicts significantly influence not only the decision to migrate, but also the trajectory, direction, tempo, and volume of migration, conflict monitoring and data collection are important tools that assist in the study of force migration (Schon, 2019). Organised violence can also lead to further cases of secondary violence such as deliberate migrant targeting, massacres, and other atrocities, which is why the empirical study of the conflict–migration nexus is important both to understand some of the causes of forced migration and also to develop mechanisms to protect displaced persons during their mobility (Kaiser and Hagan, 2015). This growing need to study more detailed mechanisms and patterns of violence has brought about the need for more granular data of 'events', namely who creates violence, who receives violence, and how the act of violence is being transmitted across actors.

Out of both ground operatives' and policy-makers' need for greater situational awareness came the idea of event datasets. While earlier examples of event datasets relied on aggregated data that provided a higher-level view of disputes and crises, over time, event datasets evolved into highly disaggregated, granular, and daily collections of important conflict events (Chojnacki et al., 2012). As conflict event

datasets became more granular, they yielded higher-quality data to undertake forecasting, and also enabled researchers to cross-utilise them with other forms of social, econometric, and political data to explain the onset of armed disputes better (Weidmann, 2013). As a result, a very rich and rigorous field emerged that leveraged conflict event datasets to explore the relationship between poverty, environmental degradation, social hostilities, government quality, and level of repression, and the main drivers of armed conflict.

This chapter explores some of the current debates on conflict event data creation and analysis, how to use social media data as a form of conflict data, and how both rapidly emerging fields can assist forced migration scholars. It starts out with a comparison of the most commonly used conflict event datasets in the field, including their comparative advantages and disadvantages in forced migration research. Then, it explores the methodological advances in harvesting both conventional and social media data as conflict event data sources, paying specific attention to media and availability biases that limit the extent to which we can rely on them. Finally, the chapter ends with a demonstration of how to harvest Twitter data to study a violent conflict (Operation Olive Branch in northern Syria), paying specific attention to how this method compares to the leading datasets, ACLED and UCDP/~~PDIO~~.

## Problems with social media as data: A brief warning about availability bias

Although traditionally scientists worked with data provided by the state security agencies (Kalyvas and Kocher, 2009; Berman et al., 2011) or local security forces (Bowsher et al., 2018), with the increasing presence of aid organisations and reporters, their frontline workers and even civilian reports have turned into valuable sources of event data (Lyall, 2010; Nettelfield, 2010). The multiplicity of data sources that come from different phases of a conflict (combat, aid, reconstruction), as well as different parties (combatants, workers, reporters, civilians), cause data availability asymmetries. Such asymmetries have been a chronic flaw of conflict event datasets, rendering micro-level comparative analyses of multiple conflicts difficult from an availability bias point of view.

In order to remedy some of this mismatch, major dataset projects have been using newspaper reports along with official or journalistic field reports in order to bring some uniformity to different conflict event data collection efforts. Weidmann (2015, 2016) discussed the benefits and pitfalls of such media-based data creation processes by focusing on three dynamics. First, that often media outlets are more interested in furthering political or ideological interests than objectively reporting on violent events, and as a result, some events may be methodically omitted while others are represented exclusively. Second, media outlets usually focus on large events, which forces them to leave out 'smaller' events due to editorial concerns. Third, most media corporations report on events that are relevant to their readership; for example, media companies in the US may be more interested in covering events

that affect American troops or assets nearby, while disregarding those that have no direct relevance.

The advent of social media was initially thought to be a positive development to bridge the above gaps. After all, social media not only contains reports and narratives from a broad range of news sources, but it also contains messages, videos, and images from the frontline combatants and civilians themselves. Such forms of data have no filter, no editorial oversight or immediate government censorship, and flow at a higher volume than regular media-based data. To that end, harvesting and logging social media event data has the potential to mitigate the data availability gap between various event datasets and conflicts. However, precisely because of the unmediated nature of social media event data, it has also fallen prey to disinformation, redundant reports, and reporting bias. For example, one of the most frequently recurring bias problems in social media data availability is the proximity to cell phone or 3G towers. Pierskalla and Hollenbach (2013) particularly demonstrated how cell phone coverage significantly affects the volume and quality of the event data we extract from ICT-based reports (see also Hollenbach and Pierskalla, 2017). Such distortions and biases between event datasets have a direct impact on conflict analysis as well as more practical applications such as aid distribution or peacekeeping (Duursma, 2018).

A more interesting point is that conflict events may often be non-violent events. The traditional method of measuring conflict intensity via the number of casualties is a practice that is being increasingly questioned in the literature for this very reason (Gleditsch, 2020). This is because literature definitions of 'peace' often stray into the 'absence of violence' territory, which is misleading because not all absences of death and violence constitute peace (Diehl, 2016). Sometimes frozen conflicts may record zero casualties, but the armed dispute (frontlines, mobilised fighters, clashes) may endure and cause forced migration. In other cases, sides may not be able to mobilise resources to fight a war, but may be in direct hostilities nonetheless. A very robust new sub-field on 'rebel governance' aims to address this problem by establishing new conflict datasets that include diplomacy, administration, and social work-related events (Arjona et al., 2015). While these new datasets and their research focus undoubtedly enrich conflict research, it also brings in new challenges with regard to measurement, generalisability, and reporting bias, as the micro-analysis of such event types are often less clear and harder to quantify in relation to casualties (Galtung et al., 2013).

## 'Off-the-shelf' conflict event datasets: A comparison

The origins of conflict event data go back to the 1960s, to Charles McClelland's 'The acute international crisis', and David Singer's 'Correlates of War project' (McClelland, 1961; Singer, 1988). Both of these earliest attempts to generate event data for the quantitative study of international diplomatic and violent events establish some of the most important benchmarks on the classification of conflict types,

periodisation, and the coding of event actors. Today, the publicly available conflict-related data ecosystem is quite robust, with generalised 'broad event datasets' such as the Uppsala Conflict Data Program – Peace Research Institute of Oslo (UCDP/PRIO) datasets, Armed Conflict Location and Events Dataset (ACLED) or Integrated Crisis Early Warning System (ICEWS), as well as more specialised datasets such as the Global Terrorism Database (START-GTD), TRAC (Terrorism Research and Analysis Consortium), International Crisis Behavior (ICB), or the Mass Mobilization (MM) Data Project. These datasets are used by governments, civil society, and NGO analysts, as much as serving as the basis of scientific research on the micro-dynamics of violent events.

Today, 'conflict event data' is widely defined as any observable information related to the interaction between violent parties. This may take the form of actual violent events such as deaths, bombings, airstrikes, drone strikes, terrorist attacks, or infrastructure targeting, or non-violent events researchers measure or observe because they have a direct impact on the course of a conflict. These may be threats, public declarations, alliance formations or break-ups, merging or splitting of organisations, and so on. ACLED's definition of an 'event' is any violent event that occurs between two designated actors that can be narrowed down into a specific time frame; these designated actors are coded as organised political groups, including militant organisations or rebel groups (Raleigh et al., 2010). UCDP's definition of an event, on the other hand, is more specific and has clear criteria for them to be coded as events:

> the incidence of the use of armed force by an organised actor against another organised actor, or against civilians, resulting in at least one direct death in either the best, low or high estimate categories at a specific location and for a specific temporal duration. (Sundberg and Melander, 2013)

These definitional variances (i.e., whether an 'event' constitutes fatality, injury, strategic action, or a threat) across datasets and over what constitutes a 'conflict event' are often so broad that they determine the selection rationale of one dataset over others, as it fits a research question or measurement method better. After all, a terrorist organisation blowing up a pipeline, which forces a nearby village to evacuate, a split within a militant group that leads to higher taxation of a particular network of towns under new leadership, and a verbal threat of a state military force to attack a rebel stronghold that is dug-in within civilian areas all constitute 'events' that directly cause forced migration and population movements (Wood, 1994).

For any newcomer to the field, choosing which dataset to use can be a daunting task. For researchers of forced migration, the most important criterion has to be disaggregation. For a very long time, aggregated datasets that were structured on a 'country/year level' coding format dominated the field. The 'disaggregation revolution' increased pace in the wake of the September 11 attacks in the US, and the growing need for more granular and detailed conflict and violence event datasets. Disaggregation in conflict event data implies separating 'event metadata' into constitutive actors, date, time, location, and type. In exploring causal mechanisms of

conflict that affect forced migration, disaggregation is critical in determining the effect of each variable (date, actor, location) on migration intensity, direction, and target (Shellman, 2008; Hegre et al., 2009). This provides a much better data architecture that enables researchers to explore deeper causes of conflict, via either surveys in conflict areas or geolocated spatiotemporal events.

Here are some of the most frequently used 'off-the-shelf' conflict event datasets popular in the scientific and policy community alike:

**UCDP/PRIO**: As one of the 'industry standard' conflict event datasets, both the joint UCDP/PRIO Armed Conflict Dataset and UCDP's other datasets such as the Georeferenced Disaggregated Event Dataset, one-sided, non-state, or the battle-related death datasets have been among the most frequently used in both policy and scientific analysis (Kreutz, 2010; Croicu and Kreutz, 2017; Sundberg et al., 2012; Sundberg and Melander, 2013; Gleditsch et al., 2002). Various levels of detail in these individual datasets relate to the granularity of spatial data (specific coordinates, village-, town-, and city-level), as well as temporality (hour, day, month, year). The UCDP/PRIO Armed Conflict Dataset and other UCDP datasets contain a unique conflict identifier (*conflict_id/integer*), location of event (*location/string*), parties to the conflict (*side_a, side_b, side_b_id, side_b_2nd/string*), source of disagreement (*incompatibility/integer*), conflict area (*territory_name/string*), along with a large selection of disaggregated variables related to the violent dispute that go back to either 1946 or 1989. Uppsala University has a related side project called ViEWS, a political Violence Early-Warning System, that forecasts violence through UCDP data (Hegre et al., 2019). This project is of particular importance to the researchers of forced migration.

**ACLED**: Another frequently used dataset, ACLED data is separated into regions: Africa, Middle East, Eastern and Southeastern Europe/Balkans, and so on (Raleigh et al., 2010). This is great for carrying out research on a specific region, but cross-regional historical analysis gets complicated as different regions start from different years. For example, while the Africa dataset starts from 1997, most Central Asia and Caucasus events start from 2016 or 2017. Likewise, Latin America data collection has begun very recently, in 2019. The temporal limitations of this dataset are formidable, but ACLED is also a better dataset for researchers of recent conflicts, as they update their dataset more frequently (weekly) than, say, UCDP/PRIO (annually). The events are coded by a unique API key, date range, event type, sub-event type (if the type of event can be categorised in two ways; for example, an act that can both be logged as 'violence against civilians' and 'excessive use of force against protestors' at the same time), type of actor, and location/region.

**START-GTD**: The University of Maryland's Study of Terrorism and National Responses to Terrorism Global Terrorism Database is a specialised dataset that only logs events that are defined as terrorist incidents. These incidents are defined as '[t]he threatened or actual use of illegal force and violence by a non-state

actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation' (LaFree and Dugan, 2007, p. 10). However, the dataset only contains actual use of force and does not log any threats. Additional problems originate from the fact that it becomes very difficult to distinguish most crisis events that include non-state violence against civilians as a terrorist tactic, or a threat of violence by states that targets populations indiscriminately as a battle tactic. Regardless, GTD has been a cornerstone of quantitative terrorism research since 1970, although newer datasets like ACLED and UCDP/PRIO have refined their event definitions and operationalisations to a greater degree. GTD variables include target country, type of attack (assassination, armed assault, bombing, hijacking, hostage-taking, infrastructure targeting, unarmed assault), target type (police, citizens/property, government, utilities-transportation, etc.), weapon type, perpetrator, and casualties (fatalities and injuries).

**GDELT**: The Global Database of Events, Language, and Tone project is among the earliest attempts to harvest large, multi-language, global media reports and log them into a unified event dataset. It monitors and scrapes web news media sources (both national and local) from around the world with a 15-minute regular update tempo (Leetaru and Schrodt, 2013). Recently, GDELT has begun feeding real-time, multi-language, machine-translated social media data into its monitoring systems to provide a broader coverage of events. GDELT has an event database of around 250 million data nodes collected since 1979. GDELT is not confined to conflict events, and includes a far broader repertoire of event types but, regardless of the type of protest, crisis, violence, economic, legal, events it logs are directly relevant to researchers of population movements and forced migration. The GDELT dataset does not attempt to provide the 'most accurate' form of event reporting, and rather focuses on collecting a trove of media narratives and perspectives. To that end, it is a dataset of 'reports about events', rather than events themselves, which is a crucial difference when it comes to using GDELT for migration research. More specifically, the GDELT dataset is more vulnerable to false reporting and misinformation in the form of false positives, both because of its very large data harvesting capacity, and the fact that the harvesting and coding are conducted by algorithms.

**ICEWS**: The Integrated Crisis Early Warning System is a more US-focused 'dataset of datasets', as its primary goal has been to assist American government analysts to predict and respond to crises that are relevant to its political objectives (Hammond and Weidmann, 2014). It is structured upon four components: iData, which is the underlying raw data of around 50 million stories from around the globe in four languages (English, Spanish, Portuguese, and Arabic) with 25 million geocoded events; iTrace, which converts iData events into index-based structured datasets that demonstrate interactions between actors, groups, and countries; iCast, which is a forecasting module that produces projects for up to six months ahead on political crisis, international crisis, religious violence, and rebellion event types; and finally, iSent, which provides a more real-time

analysis of social-media-heavy content to provide more recent situational awareness. The main problem with ICEWS is that it is not publicly available. As it is a product designed for US government use, it is accessed through specialised servers, although the scientists in charge of the project upload past versions of it on Dataverse.[1] This is still very useful for analyses of past events (for example, the relationship between conflict, crisis, and forced migration in previous years), but unlike GDELT, most up-to-date ICEWS data is not open for public use.

**xSUB**: As another attempt to create an integrated database, xSub merges more than 25,000 event datasets and organises them into easily navigable variables related to spatial (coordinates, grid, town/city name), temporal (dd.mm.yy format), and actor-specific conditionals that go back to 1969 (Zhukov et al., 2019). xSub is a very important project for researchers that seek to understand the comparative advantages of various datasets and test hypotheses using datasets with different measurement, periodisation and actor focus types. The need for xSub originated from the observation that 'grand' datasets like ACLED and UCDP/PRIO are often not detailed enough for micro-level analyses of very clearly defined crises, especially at the spatial level. In order to remedy this gap, most studies end up creating their own specialised, dedicated datasets, leading to a proliferation of a large number of very detailed but disconnected subnational event dataset ecosystem. In addition, multiple datasets that cover the same (period, actor) dyad or events use different definitions and operationalisations, leading to important incompatibilities that impair replicability. xSub's data sources include ACLED, Empirical Studies of Conflict's Worldwide Incidents Tracking System, the US government's Iraq Significant Acts (SIGACTS) database, the Social Conflict Analysis Database (SCAD), and other major 'off-the-shelf', as well as donated, research data. These data sources are integrated with a dedicated R package called MELTT (Matching Event Data by Location, Time, and Type), which dictates the fundamental aggregation principle of the xSub project.

The conflict event data ecosystem goes beyond these usual suspects. Although these datasets are some of the most frequently used in both more established and more recent scholarly works, there are important additional datasets that would enable researchers to cross-validate data sources. Some of those important supplementary datasets are:

**Social Conflict Analysis Database (SCAD)**: covers protests, intra-state conflicts, strikes/riots, and one-sided violence across the African continent, Latin America, and the Caribbean through 1990–2017 (Salehyan et al., 2012);

**Mass Mobilization Data Project**:[2] logs anti-government protests and demonstrations across 162 countries through 1990–2018, and includes protester demands, government reactions, location, and mobilisation identity variables;

[1] https://dataverse.harvard.edu/dataverse/icews

**Integrated Network for Societal Conflict Research (INSCR)**:[3] supported by the US government to create a policy-relevant 'master database' of forcibly displaced populations, major episodes of political violence, state failure, high-casualty terrorist bombings, and state fragility markers in 167 countries from 1946 to date;

**Sexual Violence in Armed Conflict (SVAC)**: an ambitious attempt to identify and log one of the most difficult-to-measure indicators in civil wars. It systematically lists all conflict-related sexual violence committed by government militaries, pro- and anti-government militias as well as terrorist/insurgent groups, and contains data about perpetrators, victims, time, and location of sexual violence events (Cohen and Nordås, 2014).

## Extracting social media as conflict event data: A sample workflow

Depending on the research question or the scope of the study, 'off-the-shelf' data solutions may not be sufficient. This may be due to the recency of the event that is being studied: prominent datasets may have been late to log them into their data pool. Or the event may be taking place in a region that is not covered by any of the datasets. Equally likely is the possibility that a study may be focusing on a different level of analysis compared to what other datasets are offering. In such cases, researchers may have to create their own event datasets, and social media offers an excellent data source to establish those tailored event data pools.

This section demonstrates a sample workflow (see Figure 18.1) that is intended to serve as an inspiration for conflict researchers and a way for newcomers to the field to gauge how to build their own workflows. Although most of the specific

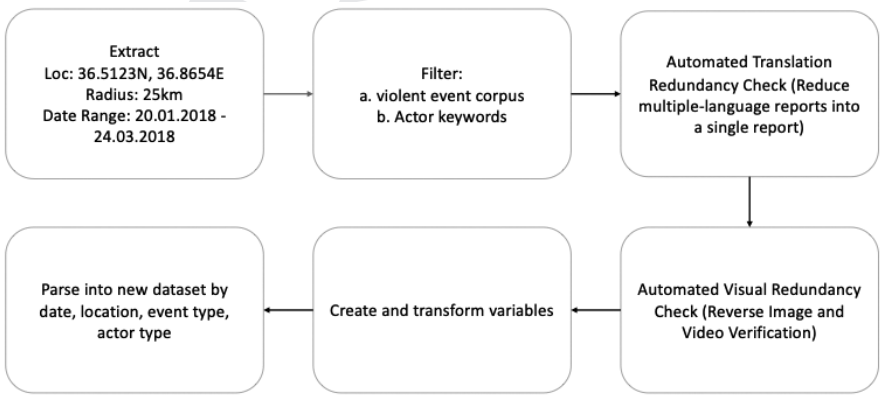


**Figure 18.1** A sample workflow for automated violent event data extraction and processing.

techniques mentioned here will inevitably become obsolete with time, the underlying principle of the workflow should have some methodological shelf-life. Most importantly, social media sites continually update their scraping and data extraction policies, so the method outlined here may not be fully available in this form at a future date.

There are two prominent ways of extracting violent event data from social media: manually, and automatically. Manual extraction, such as those based on the Textual Analysis By Augmented Replacement Instructions (TABARI) coding system (Best et al., 2013), involves employing coders (read: armies of assistants) that scroll through social media feeds on Twitter, Facebook, or any other platform to identify and log events one by one, by hand. This method is feasible for the study of short-lived isolated incidents with a moderate amount of data, such as small-scale clashes or incidents. Manual extraction has a lower likelihood of producing redundant or faulty data, due to the fact that the coder extracts such information one by one, checking for such errors in real time (King and Lowe, 2003). Automatic extraction, on the other hand, uses one extraction algorithm (or a combination of several) to set extraction criteria (such as named-entity recognition) to fish for data within a larger pool of social media feeds. Box 18.1 provides an overview of reliability in content analysis and annotation quality assessment, which is particularly important for large-scale data annotation.

The first decision a researcher will have to make will inevitably depend on the research question: is the study focusing on a particular event (i.e., armed clash, threats against energy infrastructure, suicide bombing), a group (a militant organisation, or members of a particular tribe, ethnic/sectarian populace), or a region (town/village, pipeline network, migration route)? This is crucial because a study that tests whether violence against a particular ethnic or a religious group, or violence in a region close to border areas, creates a greater magnitude of forced migration will have to follow different entity extraction protocols. Similarly, a study that focuses on whether a certain threshold of violence against civilians in areas under peacekeeper control triggers greater mistrust towards peacekeeping, and hence, greater likelihood of migration, will require different named entities.

For the sake of demonstration, this section will focus on Operation Olive Branch, the cross-border incursion by the Turkish Armed Forces into northern Syria to push the People's Protection Units (YPG) out of the town of Afrin (20 January to 24 March 2018) using Twitter data only. As mentioned earlier, the extraction criteria for social media data largely depend on the research question. Here we will apply a two-tier methodology: first, we'll extract all data from a 25 km radius of Afrin (36.5123°N, 36.8654°E), then omit irrelevant data based on a corpus of keywords. These keywords can either be fed into a corpus via pre-extraction from newspaper sources or user-generated word combinations, or simply can be extracted through machine-learning algorithms that are trained on datasets like UCDP/PRIO, ACLED, and others. A sample violent event type list could be listed as follows. Given space constraints it is impractical to create a sub-list of keywords that belong to these event lists, but further information on this could be found in (Atkinson et al., 2017;

**Box 18.1** Reliability in content analysis

Scraping or collecting large datasets from the internet is possible, but annotation of these datasets can be tedious and expensive. Manual annotation can be subjective and occasionally erroneous. Subsequently, using multiple annotators for the same data item and looking at their agreement is a typical solution to ensure annotation quality.

Amazon's *Mechanical Turk* service made large-scale, distributed, and cheap human annotation possible, and other initiatives followed suit. Approaches to improve the quality of annotations include specifying qualification levels (i.e., filtering the annotators), inserting items with known labels (i.e., *implicit screening questions*) to perform quality assessment, analysis of correlations between annotations, IP address analysis, attention checks, outlier analysis, anomalous pattern detection, response time checks, and such. Providing good instructions to annotators, randomising items, and constraining task completion times are other recommendations.

For two annotators annotating nominal data, *Cohen's kappa* is used as an inter-annotator agreement measure, where the probability of chance agreement is also taken into account (Cohen, 1960). In this and other agreement coefficients, the general form is:

$$Agreement = 1 - \frac{Observed\ disagreement}{Expected\ disagreement}.$$

Accordingly, Cohen's kappa is written as:

$$\kappa = 1 - \frac{1 - p_o}{1 - p_e},$$

where $p_o$ is the relative observed agreement among raters, and $p_e$ is the hypothetical probability of chance agreement. In practice, $p_e$ is also estimated from the data of the annotators, and not assumed via problem structure. It is taken as a squared geometric mean of proportions. In a similar measure, called *Scott's pi* (Scott, 1955), the main formula is the same, but $p_e$ is calculated by the squared arithmetic means of the marginal proportions. For more annotators, *Fleiss' kappa* can be used, which generalises Scott's pi (Fleiss et al., 2013).

Krippendorff (2004) discussed reliability in content analysis, and noted that 'agreement is what we measure; reliability is what we wish to infer from it'. According to him, an agreement coefficient can become an index of reliability only if (1) it is applied to proper reliability data, resulting from duplicating the process of coding, categorising, or measuring a sample; (2) it treats units of analysis as separately describable or categorisable, whereby a coding procedure is used and coders are treated as interchangeable, and observable coder idiosyncrasies are counted as disagreement; (3) its values indicate, and correlate with, the conditions under which one is willing to rely on imperfect data, which means it should produce meaningful and interpretable values.

Alhelbawy et al., 2016). Depending on the research question, researchers could use automated translation API services or human translators to create Arabic, Farsi, or Kurdish versions of these lists, as required (see Table 18.1). Then, optionally, a

**Table 18.1** A sample meta-keyword list containing macro-level violent event types

| |
|---|
| aerial_attack, armed_clash, border_incident, chemical_attack, drone_strike, espionage, explosion, military_exercise, military_operation, sabotage, shelling, shooting, curfew, protest, riots, road_blockade, assassination, bomb_defusal, kidnapping, security_incident, security_operation, smuggling, pipeline_damage, pipeline_shutdown |

**Table 18.2** Actor-specific entity recognition list for Kurdish groups in, or relevant to, northern Syria

| |
|---|
| Partiya Karkerên Kurdistanê, PKK (Kurdistan Workers' Party) – Hêzên Parastina Gel, HPG (People's Defence Forces) – Yekîneyên Jinên Azad ên Star, YJA STAR (Free Women's Units) – Koma Civakên Kurdistan, KCK (Group of Communities in Kurdistan) – Teyrênbazê Azadiya Kurdistan, TAK (Kurdistan Freedom Falcons) – Partiya Yekîtiya Demokrat, PYD (Democratic Union Party, Syria) – Tevgera Ciwanen Welatparêz Yên Şoreşger, YDG-H (Patriotic Revolutionary Youth Movement) – Yekîneyên Parastina Gel, YPG (People's Protection Units) – Yekîneyên Parastina Jin, YPJ (Women's Protection Units) – Yekîtiya Nîştimanî ya Kurdistanê, PUK (The Patriotic Union of Kurdistan) – Partiya Demokrat a Kurdistanê, KDP (Kurdistan Democratic Party) – Encûmena Niştimanî ya Kurdî li Sûriyê, KNC (Kurdish National Council) – Peshmerga (Pêşmerge) – Partiya Jiyana Azad a Kurdistanê, PJAK or HRK (Party of Free Life of Kurdistan) – Hêzên Parastina Jinê, HPJ (Women's Defence Forces) – Yekîneyên Parastina Rojhilatê Kurdistan, YRK (East Kurdistan Defense Units) – Yekîneyên Berxwedana Şengalê, YPŞ (Sinjar Resistance Units) – Parastin u Zanyari (Protection and Information, KRG official intelligence) – Quwwāt Sūriyā al-Dīmuqrāṭīya, SDF (Syrian Democratic Forces) |

second layer of keyword filter can be introduced to sort violent event data according to actors (groups, individuals). A sample collection of Kurdish-specific group keywords (Kurdish and English-only) relevant during Operation Olive Branch are given in Table 18.2.

It is important to expect data redundancy in social media data extracted from war zones. Often a violent event may have multiple reports in the same language or across different languages, so it is imperative to build precautions that will reduce multiple reports of the same events into a single event at the level required by the research question. Most automated translation API services like Google Translate or Microsoft Translate work fine, at least in recognising simple violent events that do not contain sarcasm. While they are imperfect at this time, they omit quite a significant amount of redundancy, leaving a more manageable set of data cleaning tasks for human coders.

Similarly, the heavy presence of disinformation and information manipulation in war zones continue to be one of the most problematic hurdles against automated conflict event data generation from social media feeds. Often, videos and images taken earlier in a different location are reshared as if they are happening recently, either as a propaganda technique or as a form of deterring the other side. One way of dealing with this is to conduct reverse image or video checks of extracted content

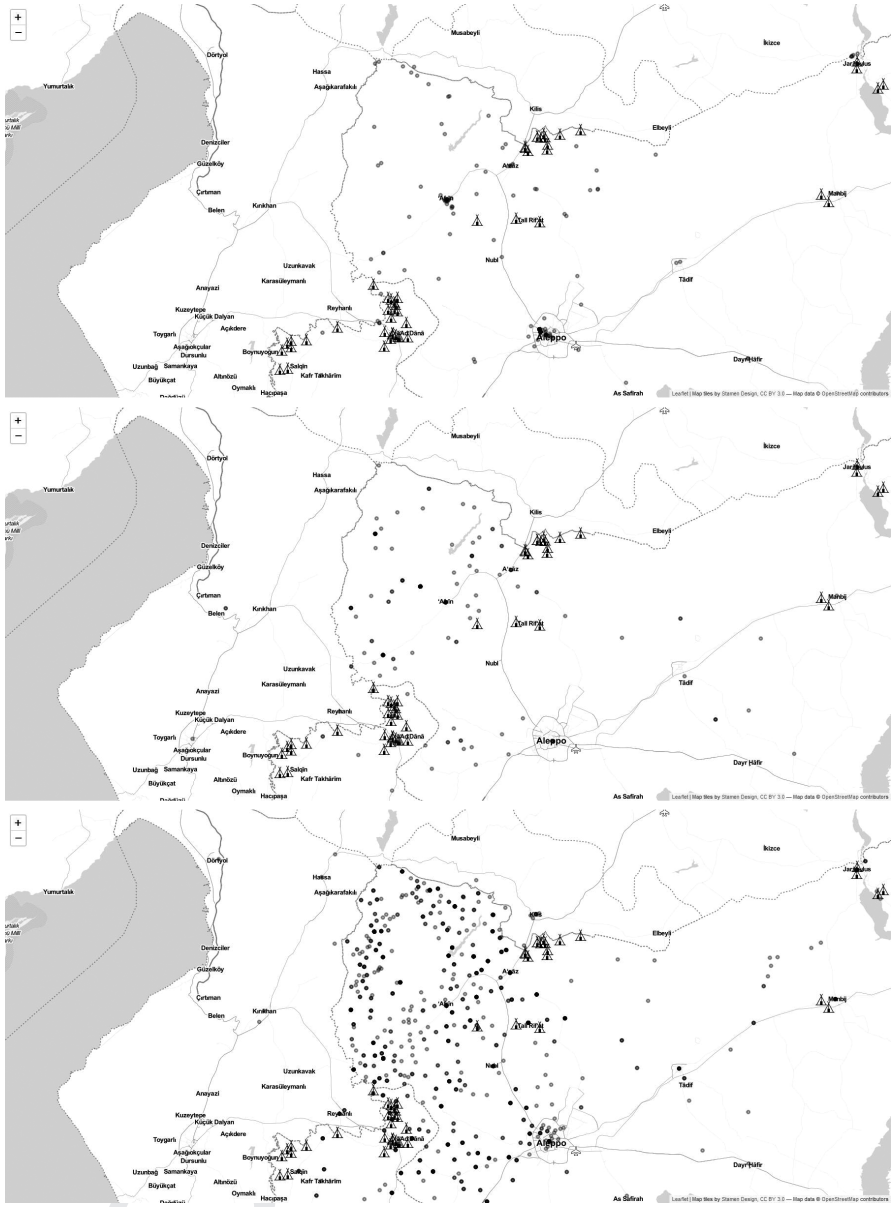


**Figure 18.2** Region-level violent event data comparison in geographic relationship to the nearest refugee camps. Top: Twitter extraction output (authors' work); middle: UCDP/PRIO; bottom: ACLED. Each dot represents a single logged violent event. Tent icons indicate The Assistance Coordination Unit (ACU) IDP camps (data: The Syrian IDP Camps Monitoring Study, Northern Syria Camps: https://data.humdata.org/dataset/idp-camps-monitoring-november-of-2018). While Twitter data yields more events compared to UCDP/PRIO, ACLED offers the largest set of events.


**Figure 18.3** Town-level (Afrin) violent event data comparison. Top: Twitter extraction output (authors' work); middle: UCDP/PRIO; bottom: ACLED. Each dot represents a single logged violent event. Twitter data yields more events at this level compared to both UCDP/PRIO and ACLED.

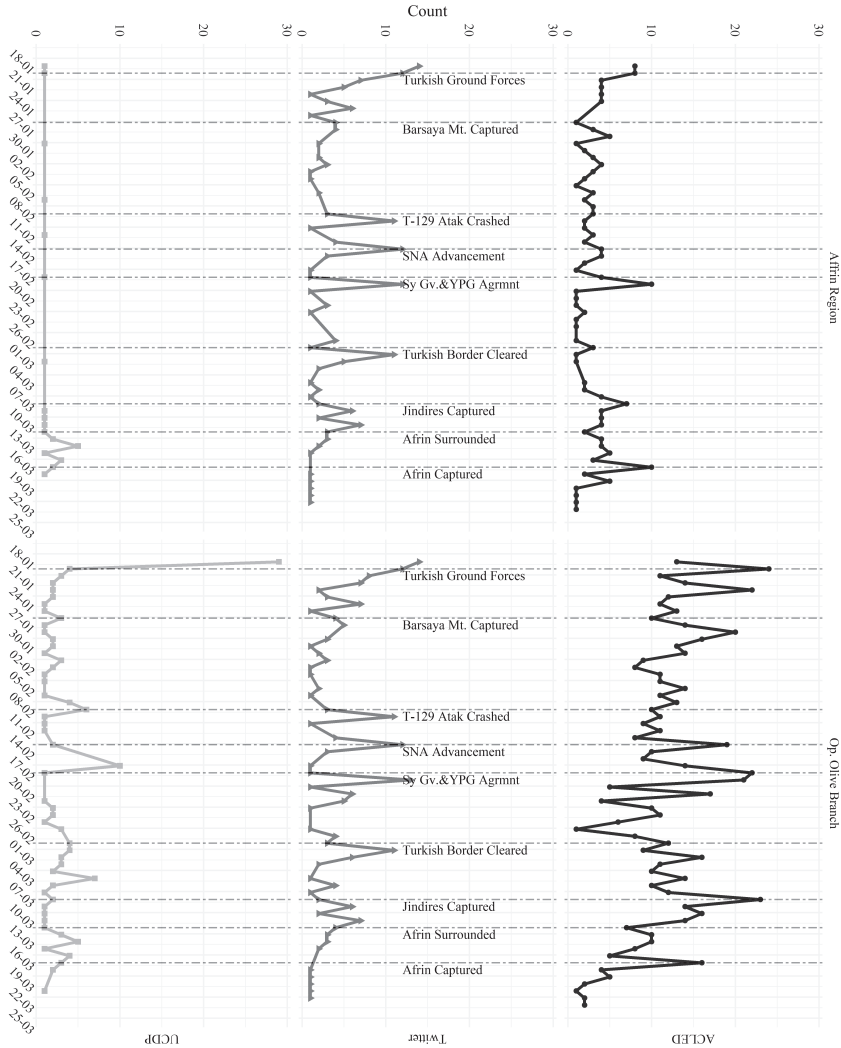


**Figure 18.4** Time-series graph showing ACLED, Twitter, and UCDP data frequencies corresponding to major violent events (18 January to 25 March 2018). Top: Afrin town and immediate vicinity ([latitude ≥ 36.49 & latitude ≤ 36.52 & longitude < 37 & longitude > 36.8, geo_code := 'Afrin Region']); bottom: wider Operation Olive Branch area ([latitude ≥ 36.49 & latitude ≤ 39 & longitude < 37 & longitude > 35, geo_code := 'Op. Olive Branch']).

via TinyEye[4] or Berify,[5] which can be integrated with the researcher's event data extraction algorithm to verify media data in real time. Once event data are extracted, they can be parsed into variables required by the researchers. For the purposes of

[4] https://tineye.com/
[5] https://berify.com/

this study we have parsed it into date, location, event type, and actor type variables. We have tested whether our Twitter-focused data extraction can offer us any advantages against two giants: the UCDP/PRIO georeferenced event dataset and ACLED. In this test, we have assessed both the spatial validity and granularity of social media data against UCDP/PRIO and ACLED, as well as their size (volume) advantages. The spatial test was conducted at both region (see Figure 18.2) and town level (see Figure 18.3).

At the broader Operation Olive Branch area, we see that ACLED data has better coverage compared to both Twitter data and UCDP/PRO data. This superiority is apparent both at the spatial level and time-series interpretation (see Figure 18.4). However, zooming into the more specific Afrin town area, ACLED and Twitter data become almost comparable in terms of granularity and volume. In some cases, such as the crash of the Turkish T-129 attack helicopter, the territorial gains of the Syrian National Army (SNA) and Turkish forces' control of the entire border area are better represented in the Twitter data. This means that at the micro level, Twitter data extraction does catch quite important and relevant violent events that are not covered by ACLED. In contrast, ACLED does better in terms of logging rural violent events. Surprisingly, Twitter data is less impressive in logging the conflict termination phase (the capture of Afrin by the Turkish forces), which is better represented by ACLED and UCDP/PRIO.

Overall, in this test the advantage of automated social media violent event data was its better performance in monitoring inner-settlement dynamics (control of streets, capture of buildings) compared to ACLED, which did better in logging violence in areas outside cities. ACLED also had an overall higher volume of violent event data, although it did miss a number of very relevant events that had a direct impact on the outcome and course of the 2.5-month operation.

## Conclusion

Scholars of forced migration have begun using conflict event datasets more frequently over the last decade. Part of this rising popularity is due to the datasets' rapidly increasing quality and granularity, enabling observers to predict, monitor, and explain forced migration events with greater accuracy and causal validity. Since the 1970s, various such datasets emerged testing different data extraction and recording techniques and various levels of analysis. However, with the advent of social media and computational research tools that allow us to harness such data, a new methodological thrust emerged that tries to build more 'bespoke' and 'research-specific' datasets from social media streams.

This chapter aimed to introduce current debates, datasets, and measurements that are involved in conflict research, and to provide an insight into how researchers can leverage a social media stream to produce tailor-made event datasets for their own work. In our study, Twitter-based conflict event data was more advantageous to ACLED and UCDP/PRIO at the local level, but fell short against ACLED in a

wider operational area. Yet despite its numerical advantages, ACLED did miss a number of significant events in our case study, Operation Olive Branch, that had a direct impact on the outcome of the war. Overall, extracting violent event data from social media is still a labour-intensive, often frustrating, but nonetheless rewarding endeavour that will certainly become more relevant to scientific research in the coming years. Especially as social media membership proliferates and smartphone ownership expands well into war zones, we can predict that more and higher-quality conflict data will be available within the next few years.

That said, social-media-based event data suffers from two major biases that affect data quality, and thus research findings. The first is the access to physical infrastructure, such as smartphones, cell phone towers, and data coverage. In areas that lack any of these prerequisites, extracting event data is still difficult, which leads to biased data that omits information from low-access regions. Although in the last few years this issue has been remedied by makeshift routers and satellite uplink facilities, it is still a major problem in war zones. The second major problem is the proliferation of disinformation and information manipulation. Conflict areas are rife with misleading content, either for deterrence purposes, or with the aim of pursuing propaganda efforts. These misleading claims can be cleaned by human coders if the data volume is low, but in a major crisis that produces a very large volume of social media data, cleaning it manually becomes impossible. Automated disinformation-recognition tools are still in their infancy, as most machine learning fact-checking systems can be easily misled. As a result, there is currently no working automated model to accurately identify disinformation in war zones, causing significant data validity problems.

We expect these problems to be less relevant in the future, given the quality of research dedicated to remedying both problems. Along with the rapidly increasing usage of smartphones and social media among refugee and combatant groups alike, the volume and quality of the data that we extract from them will be significantly improved in the near future.

## Acknowledgements

This work was supported by The Scientific and Technological Research Institution of Turkey (TUBITAK), ARDEB 1001 Program, Project Number: 120K986, Title: ‘Silahlı Örgütlerin Dijital Kamu Diplomasisi – Suriye Ve Irak Örnekleri [Digital Public Diplomacy of Armed Organizations – Syria and Iraq Cases]’.

## References

Alhelbawy, A., Massimo, P., and Kruschwitz, U. (2016), ‘Towards a corpus of violence acts in Arabic social media’, *in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, 1627–31.

Arjona, A., Kasfir, N., and Mampilly, Z. (2015), *Rebel Governance in Civil War*, Cambridge University Press, Cambridge.

Atkinson, M., Piskorski, J., Tanev, H., and Zavarella, V. (2017), 'On the creation of a security-related event corpus', *in Proceedings of the Events and Stories in the News Workshop*, 59–65.

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.

Best, R. H., Carpino, C., and Crescenzi, M. J. (2013), 'An analysis of the TABARI coding system', *Conflict Management and Peace Science* 30(4), 335–48.

Bowsher, G., Bogue, P., Patel, P., Boyle, P., and Sullivan, R. (2018), 'Small and light arms violence reduction as a public health measure: The case of Libya', *Conflict and Health* 12(1), 1–9.

Chojnacki, S., Ickler, C., Spies, M., and Wiesel, J. (2012), 'Event data on armed conflict and security: New perspectives, old challenges, and some solutions', *International Interactions* 38(4), 382–401.

Cohen, D. K. and Nordås, R. (2014), 'Sexual violence in armed conflict: Introducing the SVAC dataset, 1989–2009', *Journal of Peace Research* 51(3), 418–28.

Cohen, J. (1960), 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement* 20(1), 37–46.

Croicu, M. and Kreutz, J. (2017), 'Communication technology and reports on political violence: Cross-national evidence using African events data', *Political Research Quarterly* 70(1), 19–31.

Czaika, M. and Kis-Katos, K. (2009), 'Civil conflict and displacement: Village-level determinants of forced migration in Aceh', *Journal of Peace Research* 46(3), 399–418.

De Mesquita, B. B., Smith, A., Siverson, R. M., and Morrow, J. D. (2005), *The Logic of Political Survival*, MIT Press, Cambridge, MA.

Diehl, P. F. (2016), 'Exploring peace: Looking beyond war and negative peace', *International Studies Quarterly* 60(1), 1–10.

Dustmann, C. and Kirchkamp, O. (2002), 'The optimal migration duration and activity choice after re-migration', *Journal of Development Economics* 67(2), 351–72.

Duursma, A. (2018), 'Information processing challenges in peacekeeping operations: A case study on peacekeeping information collection efforts in Mali', *International Peacekeeping* 25(3), 446–68.

Fleiss, J. L., Levin, B., and Paik, M. C. (2013), *Statistical Methods for Rates and Proportions*, John Wiley & Sons, Chichester.

Galtung, J., Fischer, D., and Fischer, D. (2013), *Johan Galtung: Pioneer of Peace Research*, Springer, New York.

Gleditsch, K. S. (2020), 'Advances in data on conflict and dissent', *in* E. Deutschmann, J. Lorenz, L. G. Nardin, D. Natalini, and A. F. X. Wilhelm, eds, *Computational Conflict Research*, Springer, Cham, 23–41.

Gleditsch, N. P., Wallensteen, P., Eriksson, M., Sollenberg, M., and Strand, H. (2002), 'Armed conflict 1946–2001: A new dataset', *Journal of Peace Research* 39(5), 615–37.

Hammond, J. and Weidmann, N. B. (2014), 'Using machine-coded event data for the micro-level study of political violence', *Research & Politics* 1(2), 1–8.

Harrington, C. (2005), 'The politics of rescue: Peacekeeping and anti-trafficking programmes in Bosnia-Herzegovina and Kosovo', *International Feminist Journal of Politics* 7(2), 175–206.

Hegre, H., et al. (2019), 'ViEWS: A political violence early-warning system', *Journal of Peace Research* 56(2), 155–74.

Hegre, H., Østby, G., and Raleigh, C. (2009), 'Poverty and civil war events: A disaggregated study of Liberia', *Journal of Conflict Resolution* 53(4), 598–623.

Hollenbach, F. M. and Pierskalla, J. H. (2017), 'A re-assessment of reporting bias in event-based violence data with respect to cell phone coverage', *Research & Politics* 4(3), 1–5.

Ibáñez, A. M. and Vélez, C. E. (2008), 'Civil conflict and forced migration: The micro determinants and welfare losses of displacement in Colombia', *World Development* 36(4), 659–76.

Kaiser, J. and Hagan, J. (2015), 'Gendered genocide: The socially destructive process of genocidal rape, killing, and displacement in Darfur', *Law & Society Review* 49(1), 69–107.

Kalyvas, S. N. and Kocher, M. A. (2009), 'The dynamics of violence in Vietnam: An analysis of the Hamlet Evaluation System (HES)', *Journal of Peace Research* 46(3), 335–55.

Katz, E. and Stark, O. (1986), 'Labor migration and risk aversion in less developed countries', *Journal of Labor Economics* 4(1), 134–49.

King, G. and Lowe, W. (2003), 'An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design', *International Organization* 57(3), 617–42.

Klabunde, A. and Willekens, F. (2016), 'Decision-making in agent-based models of migration: State of the art and challenges', *European Journal of Population* 32(1), 73–97.

Kreutz, J. (2010), 'How and when armed conflicts end: Introducing the UCDP Conflict Termination dataset', *Journal of Peace Research* 47(2), 243–50.

Krippendorff, K. (2004), 'Reliability in content analysis: Some common misconceptions and recommendations', *Human Communication Research* 30(3), 411–33.

LaFree, G. and Dugan, L. (2007), 'Introducing the global terrorism database', *Terrorism and Political Violence* 19(2), 181–204.

Leetaru, K. and Schrodt, P. A. (2013), 'GDELT: Global data on events, location, and tone, 1979–2012', *in Proceedings of the ISA Annual Convention*, 1–49.

Lyall, J. (2010), 'Do democracies make inferior counterinsurgents? Reassessing democracy's impact on war outcomes and duration', *International Organization* 64(1), 167–192.

McClelland, C. A. (1961), 'The acute international crisis', *World Politics* 14(1), 182–204.

Moore, W. H. and Shellman, S. M. (2004), 'Fear of persecution: Forced migration, 1952–1995', *Journal of Conflict Resolution* 48(5), 723–45.

Nettelfield, L. J. (2010), 'From the battlefield to the barracks: The ICTY and the armed forces of Bosnia and Herzegovina', *International Journal of Transitional Justice* 4(1), 87–109.

Pellegrini, P. A. and Fotheringham, A. S. (2002), 'Modelling spatial choice: A review and synthesis in a migration context', *Progress in Human Geography* 26(4), 487–510.

Pierskalla, J. H. and Hollenbach, F. M. (2013), 'Technology and collective action: The effect of cell phone coverage on political violence in Africa', *American Political Science Review* 107(2), 207–24.

Raleigh, C., Linke, A., Hegre, H., and Karlsen, J. (2010), 'Introducing ACLED: An armed conflict location and event dataset: Special data feature', *Journal of Peace Research* 47(5), 651–60.

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), 503–11.

Salt, J. and Stein, J. (1997), 'Migration as a business: The case of trafficking', *International Migration* 35(4), 467–94.

Schmeidl, S. (2001), 'Conflict and forced migration: A quantitative review, 1964–1995', *in* A. R. Zolberg and P. Benda, eds, *Global Migrants, Global Refugees: Problems and Solutions*, Berghahn, New York, 62–94.

Schon, J. (2019), 'Motivation and opportunity for conflict-induced migration: An analysis of Syrian migration timing', *Journal of Peace Research* 56(1), 12–27.

Scott, W. A. (1955), 'Reliability of content analysis: The case of nominal scale coding', *Public Opinion Quarterly* 19(3), 321–5.

Shellman, S. M. (2008), 'Coding disaggregated intrastate conflict: Machine processing the behavior of substate actors over time and space', *Political Analysis* 16(4), 464–77.

Singer, J. D. (1988), 'Reconstructing the Correlates of War dataset on material capabilities of states, 1816–1985', *International Interactions* 14(2), 115–32.

Stark, O. and Levhari, D. (1982), 'On migration and risk in LDCs', *Economic Development and Cultural Change* 31(1), 191–6.

Sundberg, R., Eck, K., and Kreutz, J. (2012), 'Introducing the UCDP non-state conflict dataset', *Journal of Peace Research* 49(2), 351–62.

Sundberg, R. and Melander, E. (2013), 'Introducing the UCDP georeferenced event dataset', *Journal of Peace Research* 50(4), 523–32.

Thobane, B., Neethling, T., and Vrey, F. (2007), 'Migration from the OAU to the AU: Exploring the quest for a more effective African peacekeeping capability', *Scientia Militaria: South African Journal of Military Studies*.

Todaro, M. P. (1969), 'A model of labor migration and urban unemployment in less developed countries', *The American Economic Review* 59(1), 138–48.

Weidmann, N. B. (2013), 'The higher the better? The limits of analytical resolution in conflict event datasets', *Cooperation and Conflict* 48(4), 567–576.

Weidmann, N. B. (2015), 'On the accuracy of media-based conflict event data', *Journal of Conflict Resolution* 59(6), 1129–49.

Weidmann, N. B. (2016), 'A closer look at reporting bias in conflict event data', *American Journal of Political Science* 60(1), 206–18.

Wood, W. B. (1994), 'Forced migration: Local conflicts and international dilemmas', *Annals of the Association of American Geographers* 84(4), 607–34.

Zhukov, Y. M., Davenport, C., and Kostyuk, N. (2019), 'Introducing xSub: A new portal for cross-national data on subnational violence', *Journal of Peace Research* 56(4), 604–14.