In data we trust? A comparison of UCDP GED and ACLED conflict events datasets

Kristine Eck

Cooperation and Conflict 2012 47: 124
DOI: 10.1177/0010836711434463

Article

In data we trust? A comparison of UCDP GED and ACLED conflict events datasets

Cooperation and Conflict
47(1) 124–141

DOI: 10.1177/0010836711434463
cac.sagepub.com
SAGE

Kristine Eck

Abstract

In recent years, several large-scale data-collection projects have produced georeferenced, disaggregated events-level conflict data which can aid researchers in studying the microlevel dynamics of civil war. This article describes the differences between the two leading conflict events datasets, the Uppsala Conflict Data Program Georeferenced Events Dataset (UCDP GED) and the Armed Conflict Location Events Dataset (ACLED), including their relative strengths and weaknesses. The aim of the article is to provide readers with some guidelines as to when these datasets should be used and when they should be avoided; it finds that those interested in subnational analyses of conflict should be wary of ACLED's data because of uneven quality-control issues which can result in biased findings if left unchecked by the researcher. The article concludes that those interested in non-violent events such as troop movements have only ACLED to choose from, since UCDP has not coded such data, but again warns researchers to be wary of the quality of the data. Finally, while the creation of these datasets is a positive development, some caveats are raised in relation to both datasets about the reliance on media sources.

Keywords

armed conflict, civil war, conflict data, events data, geocoding, geographic study of war

Introduction

In recent years, researchers studying civil war have increasingly come to embrace a "microlevel" approach to the topic – an approach predicated on the idea that there is a fundamental mismatch between many civil war theories and their empirical applications. A reaction to the country–year level analyses common in the literature (Collier and Hoeffler, 2001; Fearon and Laitin, 2003), the microlevel approach posits that researchers should focus on subnational or individual levels of analysis, which are better

Corresponding author:
Kristine Eck, Department of Peace and Conflict Research, Uppsala University, Box 514, SE-751 20 Uppsala, Sweden
Email: Kristine.Eck@pcr.uu.se

suited to studying the causal claims found in the literature. Within the drive for empirical disaggregation, there are two trends. The first is for household or individual surveys within conflict-stricken countries which seek to address individual-level factors to explain the causes and outcomes of civil war (Blattman, 2009; Humphreys and Weinstein, 2008). The second—and the focus of this paper—is a trend towards geographical and temporal disaggregation of conflict events.[1] Researchers who embrace this approach employ refined data on the spatial elements of where violence occurs and use disaggregated temporal units in order to study the dynamics of warring parties' behaviour. For example, Balcells (2010) uses data on over 1,000 municipalities in Catalonia during the Spanish Civil War to show that the degree of violence against civilians was higher in areas where prewar electoral competition between rivals approached parity. Similarly, Kopstein and Wittenberg's (2011) data on 231 localities in Poland reveal that pogroms against Jews during the summer of 1941 were more likely to occur in areas where there were greater levels of pre-existing intercommunal polarization between Jews and the titular majority group. Using municipality-level fatality data for Bosnia's civil war, Weidmann (2011) finds that ethnicity affects wartime patterns of violence in two ways: macroterritorial explanations that focus on efforts by ethnic groups to create ethnically homogenous territories and microterritorial explanations that are the result of local ethnic resentment and fear activated once the war had started. Eck (2010) uses events-level data in Nepal to examine how battlefield outcomes affected rebel recruitment and found that when rebels suffered losses on the battlefield they were more likely to abduct civilians for indoctrination and recruitment efforts in the immediate aftermath. These sorts of studies would not be possible without disaggregated and systematic data: while qualitative research can often provide insights into the causal mechanisms driving these patterns, the patterns themselves are often difficult to observe without a large number of data points. Likewise, large-N country-level analyses lack the fine-grained data necessary to answer many questions about conflict dynamics. In adopting such a disaggregated approach to the study of civil war, researchers hope to study internal dynamics across time and space to be able to draw better inferences about the local conditions which affect the production of violence and the strategic behaviour between warring parties.

Microlevel approaches are hindered, however, by a lack of available data. Most microlevel research has been based on within-country studies, which often rely on historical and archival work and which leverage the researcher's knowledge of a particular area (see also Kalyvas, 2006). It remains unclear, however, the extent to which findings from these single-country studies can be generalized to other civil war settings. In recent years, several large-scale data-collection projects have been undertaken to rectify this problem. Based on the collection of events data, these projects include the Armed Conflict Location Events Dataset (ACLED), the Political Instability Task Force (PITF) Worldwide Atrocities Dataset and the Social Conflict in Africa Database (SCAD).[2] While several of these datasets have only just been released, those that have been available for several years have seen relatively little traction in the research community. The most famous of these datasets, ACLED, has of yet only been used in a handful of articles, most of which are authored by its lead researchers (Hegre et al., 2009; Raleigh and Hegre, 2009; Raleigh et al., 2010). Given the interest in the research community for such

data, one must consider why ACLED has not been employed to a greater extent in empirical analyses.

There are two reasons why this may be the case. First, ACLED's unit of analysis is the event (violent or non-violent), yet researchers rarely theorize about events *per se*, but rather the production and targets of violence. ACLED's approach means that an event like the massacre at Srebrenica is given the same weight in the data as a sniper attack in Sarajevo. It is conceptually problematic for many theories of civil war when no distinction is made in the intensity and nature of violence. Second, ACLED does not provide users with the information that is often needed to study theories of civil war: it provides no information distinguishing between whether the actor is connected to the state (i.e. military or police forces), nor does it provide Actor IDs which could be used to track the behaviour of a warring party. Similarly, it does not provide a Conflict ID and contains a wide array of violent events, including some criminal violence perpetrated by unknown individuals. This makes it difficult for researchers to weed out events which actually pertain to armed conflict in an area, and virtually impossible to link up with other data on civil war available from other outlets, such as the Uppsala Conflict Data Program (UCDP) or the Non-State Actor Dataset (Cunningham et al., 2009).

Many of these problems are rectified by the newly released UCDP Georeferenced Events Dataset (UCDP GED; Melander and Sundberg, 2011), which is currently available for Africa for the period 1989–2010. UCDP GED provides fatality estimates for each event as well as actor and conflict IDs that allow users to merge the data with other existing datasets on armed conflict. Nonetheless, researchers interested in using these data need more information about their coverage and quality, and about the types of research questions for which they can be used. The purpose of this article is to describe the differences between UCDP GED and ACLED, including their relative strengths and weaknesses. The article also provides readers with some guidelines as to when these datasets should be used and when they should be avoided, and finds that those interested in subnational analyses of conflict should be wary of ACLED's data due to uneven quality-control issues which can result in biased results if left unchecked by the researcher. The article concludes that those interested in non-violent events such as troop movements have only ACLED to choose from, since UCDP has not coded such data, but again warns researchers to be wary of the quality of the data. Finally, while the creation of these datasets is a positive development, some caveats are raised in relation to both datasets about the reliance on media sources.

Conceptualizing events data

Events data break down armed conflict into the basic interactions between parties. Each event constitutes an observation, and so each armed conflict can produce thousands of individual events. Researchers—including the two datasets of interest here, ACLED and UCDP GED—operationalize "event" in different ways. ACLED does not provide a definition of event, but does specify that events occur between designated actors and are coded to occur at a specific point location on a specific day (Raleigh et al., 2009), and that conflict actors "include rebels, militias, and organized political groups who are involved in events over issues of political authority" (ACLED, n.d.: 2). UCDP GED

defines a conflict event as "the incidence of the use of armed force by an organized actor against another organized 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 et al., 2010).[3] UCDP GED specifies that conflict events must adhere to the general and established UCDP definitions that are the basis for the UCDP–PRIO Armed Conflict Dataset as well as the UCDP One-Sided and Non-State Datasets.[4]

There are a number of conceptual differences between these definitions. The most obvious difference is what constitutes a conflict event for the purposes of the respective datasets. UCDP restricts its domain to events which result in a fatality, while ACLED also includes non-fatal events (injuries, etc.) and non-violent events (arrests, troop movements, demonstrations, etc.). In doing so, however, ACLED does not specify what constitutes armed conflict, making it difficult to determine what behaviour is included and what excluded. For example, ACLED included the following incident in its category "violence against civilians": "A rebel group attacked a livestock farm along Maramvya's 15th Avenue, very near Bujumbura, the capital, stealing about 51 cows. They wounded four cows, including three calves. The other animals were not that lucky as the rebels shot dead 17 cows."[5] The question is whether attacks on livestock should be conceptualized as belonging to armed conflict because of the purported involvement of an (unnamed) rebel group, or whether this strays too far from common understandings of conflict behaviour. There is a trade-off at play here. ACLED is able to include far more conflict events because of its lack of restrictions on inclusion; while UCDP GED can include far less, but those incidents it does include are attributable lethal behaviour by established warring parties. UCDP GED has greater confidence that all events included in its dataset indeed are conflict events, but at the expense of excluding many of the diffuse and unidentifiable actions which occur in the context of civil war. ACLED is able to capture these events, but at the expense of the conceptual validity of the data. It is up to the end-user to determine which of these datasets is more appropriate for the research question at hand.

There are other implications of coding all conflict events (ACLED) versus only conflict events which result in fatalities (UCDP GED). On the one hand, ACLED is more inclusive and, generally speaking, inclusiveness is an attractive characteristic in a dataset. But ACLED makes no distinction between events in terms of their lethality (see Weidmann, 2011). This means that all events have the same weight: the massacre of over 8,000 people at Srebrenica constitutes a single event in ACLED, as does a sniper killing in Sarajevo; these are both categorized as "violence against civilians" and are thus indistinguishable in ACLED's dataset. Researchers must ask themselves whether it is reasonable that these two events carry the same weight in the dataset. Many theories of civil conflict would suggest that these acts are driven by different dynamics and would suggest that distinguishing between them provides analytical leverage; mass killings are arguably different in causes and effects from sniper fire. But others would argue that both events share a similar politico-strategic intent and objective and are only differentiated by tactical technique and scale. Ultimately, which viewpoint is correct is contingent upon the research question and theoretical interests of the end-user.

Table 1 gives the number of observations by country and by type of violence in ACLED and UCDP GED.[6] I restrict the UCDP GED data to the period 1997–2010 so that it overlaps with ACLED, but UCDP GED's data stretch back to 1989 in the full dataset. I also restrict the sample to Africa, although ACLED has several other countries available. We should expect ACLED to have many, many more events because ACLED makes no requirement of an identifiable actor, no requirement of a fatality and no requirement that the violence reaches the threshold of 25 annual deaths as is required for inclusion in the UCDP dataset.[7] This is compounded by the fact that a news report stating that "fighting took place over the past 3 weeks in Region X" will result in 21 events, one for each day, in ACLED.[8] In UCDP GED this report would result in one event, with an indication in the time precision variable that the event took place over 3 weeks. Looking at the table, we can see that UCDP has about one-third of the number of observations that ACLED has. If anything, this is higher than should be expected given the narrower score of the UCDP data.[9] In two cases (Algeria and Congo), UCDP actually records more events than ACLED, which is surprising given the broader scope of ACLED's data collection. The biggest difference between the projects lies in the one-sided violence against civilians category. While UCDP observations total approximately 50% of ACLED's total number for armed conflict, for one-sided violence this number is only 17%. This is almost surely due to the requirement of an identifiable actor and ACLED's generous inclusion of all forms of violence against the civilian population, which may include the types of events discussed above in the Bujumbura example.

Table 1. Number of observations in ACLED and UCDP GED, 1997–2010

	ACLED				UCDP GED			
Country	Total	State-based	Non-state	One-sided	Total	State-based	Non-state	One-sided
Algeria	798	396	13	389	2371	2107	8	256
Angola	2399	1926	0	473	717	489	0	228
Benin	8	0	0	8	.	.	.	.
Botswana	11	5	1	5	.	.	.	.
Burkina Faso	33	5	13	15	.	.	.	.
Burundi	2754	1363	105	1286	1142	727	24	391
Cameroon	105	42	22	41	2	0	2	0
CAR	522	236	21	265	169	61	0	108
Chad	456	259	17	180	172	96	13	63
Comoros	0	0	0	0	4	1	3	0
Congo	21	4	1	16	177	97	0	80
DRC	3782	2028	655	1099	1010	210	151	649
Djibouti	57	50	0	7	5	5	0	0
Egypt	264	56	22	186	49	33	0	16
Equatorial Guinea	16	6	0	10	.	.	.	.
Eritrea	175	94	0	81	33	33	0	0
Ethiopia	1285	877	111	297	989	761	127	101

(Continued)

Table 1. (Continued)

Country	ACLED				UCDP GED			
	Total	State-based	Non-state	One-sided	Total	State-based	Non-state	One-sided
Gabon	9	1	0	8	.	.	.	.
Gambia	41	3	2	36	.	.	.	.
Ghana	73	5	29	39	13	0	13	0
Guinea	286	133	9	144	49	24	1	24
Guinea Bissau	107	88	2	17	21	21	0	0
Ivory Coast	598	236	78	284	122	47	26	49
Kenya	1831	474	374	983	238	4	136	98
Lesotho	75	47	2	26	5	5	0	0
Liberia	762	515	53	194	126	73	0	53
Libya	11	6	1	4	.	.	.	.
Madagascar	86	9	2	75	39	0	38	1
Malawi	41	0	2	39	.	.	.	.
Mali	84	38	21	25	34	28	5	1
Mauritania	31	19	1	11	3	2	0	1
Morocco	19	4	4	11	2	0	0	2
Mozambique	107	10	2	95	.	.	.	.
Namibia	129	47	0	82	20	10	0	10
Niger	184	120	10	54	35	32	1	2
Nigeria	1945	537	461	947	303	27	227	49
Rwanda	330	136	0	194	147	78	0	69
Senegal	353	159	42	152	100	67	2	31
Sierra Leone	1012	329	301	382	5	4	0	1
Somalia	3638	1476	572	1590	1499	991	452	56
South Africa	859	237	73	549	13[a]	.	.	.
Sudan	2559	1077	340	1142	1350	638	152	560
Swaziland	54	30	0	24	.	.	.	.
Tanzania	168	16	13	139	10	0	2	8
Togo	24	4	2	18	93	0	0	93
Tunisia	12	4	0	8	.	.	.	.
Uganda	3096	1572	138	1386	1333	871	51	411
Zambia	133	17	0	116	1	0	0	1
Zimbabwe	3399	14	22	3363	40	0	0	40
Total	***34742***	*14710*	*3537*	*16495*	***11418***	*7332*	*1283*	*2803*

ACLED includes additional events which cannot be categorized in any of the three categories (e.g. events with both actors missing; events involving peacekeepers as main actors, etc.). A dot indicates that the data are not included because no conflict/violent actor reached the threshold of 25 annual fatalities anytime during the period.

[a]The difference between ACLED and UCDP GED for South Africa is due in large part to ACLED's inclusion of unidentified parties. Once these are removed, ACLED only records 92 events, most of which are non-fatal or involve protesters killing police, and other forms of violence which UCDP GED does not consider to be related to armed conflict or which fall under its 25 fatality per year threshold.

The question of whether violence should play such a central role in the study of war is again dependent on the research interests of the researcher. UCDP GED, more than ACLED, is narrowly focused on fatalities. This is in part because a major impetus behind the project is to determine global trends in armed conflict, driven by questions such as whether there are more or fewer conflicts in the world today compared to before, whether these conflicts are more violent, and which areas are becoming more violent (and therefore in more critical need of intervention). But from a theoretical perspective, researchers studying civil war are usually interested in many non-violent facets of warfare; for example, planning, troop movements, weapons sales, destruction of property, threats, alliances, recruitment and training of forces, and so on. It is reasonable to study some of these things as events, while others may be better conceptualized as continuous processes (such as support from a state/group to a warring party). Some of these variables can be found in such a format in the UCDP Conflict Encyclopedia, which provides data on negotiations, third-party mediation, secondary support to warring parties, troop size, peace agreements, etc.[10] But the vast majority of the day-to-day non-violent events that occur in the context of civil war (e.g. recruitment, threats and troop movements) are not found in UCDP data, whether the UCDP GED or UCDP Conflict Encyclopedia. Thus, the dataset is best suited to research projects focused on the production or effects of conflict violence.

ACLED, on the other hand, includes about 2,700 non-violent events and 6,500 events of riots or protests. Because UCDP GED does not include these categories, researchers wishing to study these phenomena have only ACLED available. The category of non-violent event includes behaviours such as troop movements, the establishment of bases, the establishment of alliances, etc. Sometimes the dates given for these events are the dates on which the event occurred, sometimes they are the dates the event is reported. It is worth noting that users considering ACLED data on troop movements should consider why the number of events is only a small fraction of all of the violent events. Presumably, most battles require the movement of troops, or result in the movement of troops afterwards. Troops also move around to gain access to resources and survival necessities. Given this, we should expect non-violent events in conflict zones to far outnumber violent events. Yet non-violent events are only 8% of the total violent events in ACLED data. Thus, it is relevant to consider how data are generated on non-violent events. While news of fighting tends to be well reported in the media, news of troop movements is often clandestine and rarely reported in the news media. We should therefore expect that it is troop movements that occur in connection with (major) acts of violence that will be noted in media reports; as such, there is likely to be a strong bias in the sort of non-violent events that are reported. For this reason, users should exercise caution and consider the data-generating process before using ACLED's non-violent events data in analysis. This is a shame because the inclusion of non-violent events is ACLED's strongest relative advantage vis-à-vis UCDP GED; should ACLED eventually find a way to overcome the current biases in the data-generation process, then these data could be quite useful to the research community.

Data quality

So far, the focus has been on data coverage, but the question of data quality is also important. To evaluate this, the author randomly selected one year to evaluate for Algeria and

Table 2. Comparison of coding quality of selected cases

	Algeria 1997		Burundi 2000	
	ACLED	UCDP GED	ACLED	UCDP GED
Total events	116	128	496	188
Events with problems	60 (52%)	6 (5%)	126 (25%)	3 (2%)
Incorrect region/admin1	23 (20%)	0 (0%)	34 (7%)	0 (0%)
Incorrect location/admin2	12 (10%)	0 (0%)	9 (2%)	0 (0%)
Incorrect geo. precision code	34 (29%)	2 (2%)	87 (18%)	3 (2%)
Events double coded	7 (6%)	1 (1%)	12 (2%)	0 (0%)
Missing events	2 (2%)	3 (2%)	...	0 (0%)

UCDP GED distinguishes between Admin1 (the first-order administrative division, i.e. province, etc.) and Admin2 (second-order administrative division, i.e. district, etc.). ACLED distinguishes between "Regions" and "Locations" though it does not make clear what the distinction is between the two; in the dataset, a region in ACLED can include an Admin1 or Admin2 location, or even an exact town.

Burundi. These countries were chosen because they are two of the very few in ACLED's data which provide enough information to evaluate the quality of the coding. For every observation, ACLED has a field called "Notes". Coders are discouraged from writing long notes (Raleigh et al., 2009), but in Algeria and Burundi the coders apparently disregarded this instruction and instead provided some text from the article used to code the event. UCDP GED also records this information in a field called "What" for every observation in the dataset.[11]

Table 2 gives the results from the quality-control analysis of all violent events. One caveat is that these determinations were made on the basis of the data provided by the respective projects, and so for some observations coders may have had access to additional data which clarified the issue.[12] In evaluating problems, the author erred on the side of generosity, and when in question gave the datasets the benefit of the doubt. Using the data available in the "Notes/What" sections, it was possible to examine whether the coder had recorded the correct region/admin1, the correct location/admin2 and the correct geoprecision code, all of which are explained in greater detail below. Sometimes it was quite clear that an event had been coded twice or that it had not been coded at all,[13] though this was much more difficult to determine without access to the original reports that the coders used; as such it is possible that these areas are more problematic than the results from Table 2 would otherwise indicate. The total percentage of events with problems indicates the percentage of events that suffered from one or more of these problems; indeed, many events were miscoded along multiple dimensions and that is why the percentage for each type of error will add up to more than the total error.

The differences between the two datasets are quite dramatic for Algeria 1997: over 50% of ACLED's observations were coded incorrectly on at least one of the dimensions listed in Table 2, while only 5% of UCDP GED's data suffered from such problems. For Burundi 2000, the extent of the quality problems is less: 25% of ACLED observations were miscoded, compared to 2% of UCDP GED observations. There were two recurring

problems with ACLED's geocoding. The first is miscoded location information. It appears that coders are not always distinguishing between villages/towns with the same name. The Algeria and Burundi data suggest that they often select the coordinates for a village without referring back to the "Notes" to ensure that they have identified the same village as designated in the news report; this is usually identifiable by the province or district in which it is located. For example, an ACLED event for Burundi on June 13, 2000 states that "Rebels tried to return to Tanzania through Musumba in Kinyinya Commune, but were repelled by police operating in Moso region". The incident is geocoded to Musumba in Ngozi province, which does not even border Tanzania. It should have been coded to Musumba in Ruyigi province, which is where Kinyinya commune can be found. The location is thus some 150 kilometers off, putting the location in northern Burundi instead of southeast Burundi. Other times, it is unclear how coders manage to get the incorrect coordinates; even major cities like Khartoum and Juba are sometimes coded with erroneous latitude/longitude coordinates.[14] UCDP GED avoids a great number of these problems through a triple-checking process. The first manual check is done by the coder, and the second by the UCDP project leader, who manually checks the data and uses Spatial Key, a visualization software for geographic data, to map the data and locate possible miscoded coordinates. In the third stage, automated scripts in Python and PHP are run to check for internal consistency in dates, actors, dyads, conflicts and fatality counts. The automated scripts pick up problems like the same city being given different coordinates.[15] The scripts normally pick up dozens of errors per country, suggesting that they are invaluable in the data-cleaning process.

The second recurring geocoding problem in the ACLED data is the misuse of the geoprecision codes. In ACLED and UCDP GED, a geoprecision code of 1 indicates that the coordinates mark the exact location that the event took place, usually an inhabited area. When a specific location is not provided, i.e. "Helmand province", ACLED and UCDP GED employ different strategies for managing this issue. ACLED selects the provincial capital, while UCDP GED selects the centroid point when available and the provincial capital when a centroid point is not available. One can debate which of these is the best practice, but what is crucial is that the data provider convey uncertainty about the location to the user. This is done through geoprecision codes; higher numbers on the geoprecision code indicate broader geographic spans and thus greater uncertainty about where the event occurred (the range for ACLED is 1–3, for UCDP GED it is 1–7).[16] As Table 2 indicates, the geoprecision code was incorrect for 29% of the observations in Algeria 1997 and 18% in Burundi 2000; ACLED often identifies coordinates as representing exact locations when in fact the original source states that the event took place "near X town" or "in Y region".[17]

It may not be immediately evident to end-users why this is so important: it is crucial because most consumers of geocoded events data are interested in examining the associations between various factors. These data are used to ask whether violence occurs in densely populated areas, in areas with low gdp, in areas with certain types of terrain or natural resources, in areas with certain types of infrastructure, and so on. If ACLED attributes violent incidents to towns when in fact they took place in rural areas, they are introducing a systematic bias in the data that can lead to invalid inferences. Using ACLED data, results will be biased towards attributes associated with urban areas due to the imprecision in ACLED's geoprecision coding.

Because there is insufficient information in ACLED's "Notes" field, it is impossible to determine whether the high levels of incorrect geoprecision codes found in Algeria and Burundi are representative of the other countries in the dataset. For violent events in ACLED, 77% of the observations have a precision code of 1 while the corresponding estimate for UCDP GED is only 29%; Table 3 provides a breakdown by country. The data show that in ACLED quite a few cases have extremely high levels of geoprecision code 1, while the numbers are much lower for UCDP GED. The question is whether it is reasonable to assume that ACLED would be able to place such a large percentage of violent events in exact locales.

Table 3. Comparison of geoprecision codes

Country	ACLED % violent events with geoprecision 1	UCDP GED % violent events with geoprecision 1
Algeria	99.8	40
Angola	81.3[a]	29.8
Benin	50	.
Botswana	100	.
Burkina Faso	84.4	.
Burundi	96.2	35.2
Cameroon	87.5	0
Central African Republic	84.5	46.7
Chad	83.9	47.1
Congo	0[a]	79.1
DRC	91.2[a]	57
Djibouti	69.0[a]	100
Egypt	100	61.2
Equatorial Guinea	100	.
Eritrea	39.9[a]	42.4
Ethiopia	58.3	23.2
Gabon	100	.
Gambia	100	.
Ghana	97.3	92.3
Guinea	96.3	71.4
Guinea Bissau	100	81
Ivory Coast	90.5	80.3
Kenya	69	42
Lesotho	100	60
Liberia	92.8[a]	65.1
Libya	54.5	.
Madagascar	100	84.6
Malawi	100	.
Mali	84.5	35.3
Mauritania	93.5	33.3

Table 3. (Continued)

Country	ACLED % violent events with geoprecision 1	UCDP GED % violent events with geoprecision 1
Morocco	80	100
Mozambique	100	.
Namibia	83.1	25
Niger	99.5	42.9
Nigeria	81[a]	61.4
Rwanda	94.6	23.1
Senegal	89.8	39
Sierra Leone	91.9	40
Somalia	49.7[a]	83.8
South Africa	97.1[a]	61.5
Sudan	51.3	36.9
Swaziland	96.3	.
Tanzania	99.4	70
Togo	100	94.6
Tunisia	100	.
Uganda	38.3	17.6
Zambia	68.4[a]	0
Zimbabwe	98.7	62.5

[a]Indicates that the country contains geoprecision values which have no meaning in the codebook. A dot indicates that the data are not included because no conflict/violent actor reached the threshold of 25 annual fatalities anytime during the period.

Fighting often takes place in rural areas, away from human habitation, and it is usually impossible to get precise geocoordinates for a location out in the bush unless there happens to be a geographical landmark in that place (fighting between Eritrea and Djibouti in 2008, for example, took place at a particular hill which could be identified).[18] Violence that takes place outside inhabited areas is extraordinarily common in some places, such as Sudan, and occurs to some extent in virtually all major armed conflicts.[19] In these areas, coders will not be able to record precision level 1, and for this reason it is reasonable to advise users to be sceptical of cases in which geoprecision 1 exceeds 85–90% of the observations. In some cases, these may be accurate and the violence truly demonstrates a pattern of occurring in urban (or otherwise identifiable) areas; in other cases, this will be immediately obvious as untenable. UCDP GED has four countries in which the percentage of observations coded to geoprecision 1 exceeds 85–90%: Djibouti, Ghana, Morocco and Togo. All four of these cases contain relatively few events which did indeed take place in urban areas.[20] In other cases, it is less plausible that a large number of violent conflict events took place in inhabited areas, such as in Rwanda, Senegal and Sierra Leone, for example. Researchers interested in using the data for these countries (and others listed in Table 3) are strongly encouraged to validate the data before employing ACLED's geocoding information in analysis. The extent to which fighting

takes place in inhabited areas is an interesting empirical question to examine, but to do so one needs data that accurately capture the location of fighting.

Actor inclusion in events data

UCDP GED is strict in its definitions of which actors can be included in its various categories of organized violence. For interstate conflict, intrastate conflict and one-sided violence against civilians, actors must be the government of a state or an organized group with a name (e.g. the National Patriotic Forces of Liberia). For the category of non-state conflict, this requirement is relaxed to include communal groups that engage in conflict with each other (e.g. Nanumbas and Konkombas in Ghana). All actors are identified with a unique Actor ID so that researchers can easily identify and follow the same group (even when it changes its name). The UCDP requires each event to be attributable to a specific warring party for inclusion. Usually this means that the rebel group is named in the original data source (newspaper article, UN report, etc.), but attribution can also be coded through inferences based on the expertise of the coder; for example, if an event takes place in a region in which the coder knows that only a single group is active, the coder may attribute that event to the actor in question. This occurs rather infrequently, and UCDP GED errs on the side of caution when attributing incidents to a group.

ACLED's requirements for actor inclusion are far looser and actors who cannot be identified (e.g. "unidentified armed men") are included. The following is an example of the kind of incident that can be found in ACLED but not UCDP GED: "[A] spate of killings by angry mob (sic) accusing people of being penis-shrinking sorcerers" (ACLED All Africa dataset; Ivory Coast, 1997–03–06). Approximately 20% of ACLED's violent events data are missing an actor name entirely (e.g. "unidentified actor") and a good portion of the remainder are stated only in general terms (e.g. "Hutu rebels"). ACLED does not provide Actor or Conflict IDs with which users could group events of interest or merge ACLED data with other data sources on civil war. A further consequence of actor identification for conflict events is that it affects what is recorded as the start and end dates of political violence and may affect analyses which seek to distinguish between armed conflict and post-conflict violence by ex-belligerents. In terms of actor identification, UCDP GED is thus far more transparent, rigorous and amenable to analysis. At the same time, one might argue that uncertainty is a key characteristic of conflict and the quest for certainty in UCDP GED might lead to the exclusion of many uncertain, but potentially, useful events. Again, researchers must weigh the trade-off between inclusion and validity.

ACLED's relative advantage in regard to actor data is that it specifies which force of the state (e.g. army, police) was engaged in the event. It usually does so in rather broad strokes, i.e. "army of Country X" rather than the specific military unit (i.e. "5th battalion of Country X") mainly due to the lack of precision in the original news sources. Journalists and even INGO reports only occasionally specify the actual unit involved in fighting, either because they do not have access to that information or because they deem it uninteresting to their audiences. Yet the organization of armed conflict is an important and fruitful avenue for research into the dynamics of war. For example, Arreguín-Toft (2007) notes that the composition of the Russian military forces deployed in Afghanistan

affected its strategic calculations and Butler et al. (2007) suggest that the type and composition of forces may help to explain various forms of violence against the civilian population. While giving greater attention to those who carry out the violence and addressing issues of control and discipline of forces is potentially important to civil war research, neither of these datasets is sufficiently detailed to allow researchers to study these questions using their data. ACLED certainly has the advantage in that it attempts to make some distinction, rough though it is, and thus for interested researchers ACLED is the only option available. Both datasets (and indeed, other researchers) should be encouraged to explore whether more and better data can be gathered on this topic.

The value of events data

Until now, the working assumption has been that events data of the type generated by ACLED and UCDP GED are useful for studying civil conflict and that the relevant dimensions to consider prior to use are the quality of the data and which dataset produces data best suited to the needs of a particular research project. But it is worth raising some general concerns about events data which researchers should consider before moving ahead with analysis. In many respects, events data are dangerous because they can convey a false sense of accuracy of the precision of the data if not considered carefully.

Events data sources like UCDP GED and ACLED are sometimes criticized for their heavy reliance on media sources.[21] There are two points to be made here before discussing the validity of this critique. The first is that UCDP GED employs news reports to generate a baseline because news reports are the only source on armed conflicts which are global in scope. UCDP GED then uses a wide array of case-specific sources such as UN and local and international NGO reports, Truth and Reconciliation Commissions, Wikileaks documents and case-oriented research to supplement the data. To the extent that others generate better data than can be found in news reports, they are incorporated into UCDP GED's data, although such sources will not exist for all cases or all time periods. ACLED takes a similar approach, though it is unclear how systematic this effort is across coders. The second point to make is that advocates of multiple systems estimation[22] are critical of ACLED and UCDP's heavy reliance on news reports and argue that analysis should not be run on data that is not based on multiple different data sources. What these researchers are missing is that the scope of both ACLED and UCDP GED is global in nature (and in UCDP GED's case, annually updated) and therefore not amenable to such approaches due to a lack of raw data.

That said, researchers have found that there are problems with media-reliant conflict events data. In his analysis of Bosnia, Dulic (2010) finds that news reports provide insufficiently detailed information to georeference data to the sub-municipal level which he argues is most theoretically relevant in the Bosnian context. He also finds that journalists frequently fail to distinguish between military and civilian victims of war, which has implications for both the study of armed conflict and for one-sided violence. Similarly, in his analysis of the Black Panther Party in the US, Davenport (2010) finds evidence of media bias in the field of contentious politics; in particular, he finds that coverage of groups and state actors varies depending on the political orientation of the source and the spatial distance between the source and the events in question. Both of

these studies suggest that source variation should become an area of inquiry in its own right and ultimately any bias should be modelled and incorporated into analyses. Scholars still do not know the extent to which media bias may occur in this setting and how it would affect causal inferences regarding the study of civil war. To this end, users should be encouraged to consider at the very least the issue of possible bias in regard to their own research questions.

Conclusion

For the end-users of events datasets, there are two main considerations. The first is what sort of data are needed. For those interested in non-violent and non-fatal events, ACLED is the only option to consider, since UCDP GED does not include such data. For those interested in fatality counts and in linking up the data to other datasets on armed conflict, UCDP GED is superior to ACLED. The second consideration for end-users is the quality of the data. In some cases, one has no choice: only one dataset provides the data needed. In such cases, the user should stop to consider whether the available data are of sufficiently high quality on the dimensions of interest, or whether it is better to explore other empirical strategies, such as collecting the data themselves (even if for a more limited temporal and spatial domain). To the extent that one's research interests allow them to choose between UCDP GED and ACLED, the analysis here shows that the quality of UCDP GED's geocoding and precision information is far superior to ACLED's. This is particularly important for anyone examining geographic dimensions of civil war. By this I mean any independent variable that is connected to the question of "where" things happen: terrain type (mountains, forest), natural resources, population sizes, gdp, infrastructure and so on. In other words, the vast majority of the types of questions examined in the civil war literature have a geographical dimension to them because of the level of measurement of the independent variables. The urban bias in ACLED's data can lead to incorrect causal inferences.

The creation of georeferenced, disaggregated conflict events datasets provides an empirical boost to the research programme on the microlevel study of war. These data allow researchers to study a myriad of questions related to the spatial and temporal dynamics of violence found within civil wars. But researchers should be alert to data quality issues and potential biases in these datasets. This article attempts to provide some guidance to end-users on how to evaluate these datasets.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Acknowledgements

Many thanks to Ralph Sundberg for his tireless assistance, as well as to Mihai Croicu, Lotta Harbom and Joakim Kreutz for useful comments. Andrew Linke was kind enough to answer my questions about ACLED. All errors are my own.

Notes

1. There are also researchers who combined the two approaches, cf. Østby (2008) and Østby et al. (2009), who use demographic and health surveys in research designs based on subnational units of analysis.
2. SCAD includes data on demonstrations, violent riots, strikes, pro-government violence (repression), anti-government violence (rebellion by actors not listed in the UCDP dataset), extra-government violence (non-state conflict) and intra-government violence (including coups). While they are sometimes described as events data, SCAD's conceptualization of an event is unconventional; for example, SCAD considers an entire war to be a single event. Because of their aggregate nature, these data are not geocoded.
3. Operationalization of each of the concepts contained in this definition can be found in Sundberg et al. (2010).
4. See UCDP (2011) for these definitions.
5. ACLED All Africa dataset: Burundi, 2000–10–22. This was not an isolated event; there are a number of events in ACLED which concern attacks against livestock.
6. *State-based* indicates armed conflict in which at least one of the parties is the government of the state. *Non-state* indicates armed conflict in which neither of the parties is the government of the state (i.e. militias, communal conflict, inter-rebel fighting, etc.). *One-sided* indicates violence against civilians. UCDP distinguishes between these categories and has strict coding rules for inclusion in each (see UCDP, 2011). ACLED makes only a distinction between violence against civilians and other forms of violent conflict. Because ACLED provides no variable which identifies when state actors are involved, the name of the actor is used in this article to make this distinction in the ACLED data. For how one-sided violence against civilians is determined in UCDP, see Eck and Hultman (2007). The distinction between battle-related violence and the deliberate targeting of civilians is the basis for an important and growing body of research; see Kalyvas (2006), Downes (2008) and Valentino et al. (2004), among others.
7. For the purposes of this article, events are reported based on the country where they occur, not on the basis of which conflict they belong to. For example, the Lord's Resistance Army (LRA) has been active almost exclusively in DRC, Sudan and the Central African Republic since 2008 (both fighting the government and attacking the civilian population). In UCDP GED, the data are recorded in two different ways: the country location data indicate the country where the violence occurred (in this case DRC or Sudan), while the Conflict ID indicates which conflict the violence belongs to (Uganda).
8. ACLED is not consistent about how such summary news reports are treated. Often they are treated as described: one event is recorded for each day and location, but sometimes they are treated as single events with time precision codes which indicate extended periods of time. It appears that different coders took different approaches to solving the problem of coding summary events.
9. The UCDP GED dataset is fully compatible with the other UCDP datasets, which means that in its current version it includes only events for conflict (or actor)-years which reach the threshold of 25 deaths per year. That said, UCDP collects data on all fatalities in collective violence but does not make public those which do not conform to its definitional specifications. If those observations were to be included, UCDP GED would have approximately 22,000 observations for the 1997–2010 period.
10. The UCDP Conflict Encyclopedia can be found at: http://www.ucdp.uu.se/gpdatabase/search.php

11. Because the "What" usually contains text taken directly from the original news articles/reports, UCDP GED is prevented by copyright law from releasing the "What" field to the general public. Users interested in this information are strongly encouraged to contact UCDP.
12. While I am affiliated to Uppsala University, I did not have access to any additional data for UCDP GED nor did I contact the project staff for clarifications when evaluating the quality of the data; ACLED and UCDP GED were thus treated equally. After the analysis, both programmes were contacted regarding the errors that were found. UCDP GED either corrected the error or was able to explain why I was mistaken in my estimation. ACLED has noted the errors and refers to the beta nature of the data.
13. Articles sometimes reference previous events, i.e. "Rebels X killed 10 civilians in Village A today. With the 5 killed in Village B last week, this brings this month's total to 15 killed". Sometimes the coder included the current event but neglected to check and ensure that the previous event had been coded as well.
14. This is not an isolated incident. In Sudan, for example, inconsistent names and coordinates are used for the same place across different events in the Jebel Marrah; no fewer than six different location coordinates are given and are sometimes given geoprecision codes of 1 (exact town) despite Jebel Marrah being a large mountain range.
15. The scripts also check for consistency between the UCDP GED and other UCDP datasets to ensure consistency across the datasets.
16. In ACLED, 849 total observations have geoprecision codes of 0 (502 if we restrict it to the subsample contained in Table 1), the meaning of which is not given. A handful of observations (18) have geoprecision codes of 4–9, which also have no meaning according to ACLED (n.d.). I was not able to get an explanation from ACLED as to why this is the case.
17. ACLED suggests that in some cases "near a town" might constitute something more like "edge of town" or "1 km junction road from town" (email correspondence, 2011–10–26). In some cases this indeed may be true, but based on the raw source material it is not likely to explain the large divergence in precision code 1 between the two programmes.
18. The quality of geocoding depends in large part on the quality of the gazetteers and maps used to identify locales. With excellent maps, coders can achieve greater levels of precision for inhabited areas; in DRC, for example, the International Peace Information Service (http://www.ipisresearch.be/mapping.php) has provided detailed maps which allow researchers to find small villages. But fighting outside of these inhabited areas cannot be located at a precision level of 1.
19. Fighting happens outside of inhabited areas for a number of reasons. In guerrilla warfare, ambushes are often staged by rebels along roads or occasionally at army/police bases or temporary camps, and the government must hunt rebels based in rough terrain like mountains and forests. Even in semi-positional and conventional warfare, battlefronts need not fall along urban areas; indeed, fighting often takes place outside towns as armies attempt to gain control over the territory. Both strategic and tactical considerations drive groups' choices about where to engage in violence, and whether this is urban/rural can vary considerably across conflicts and periods. This is not to say that conflict violence does not take place in inhabited areas; it most certainly does. But researchers familiar with the dynamics of conflict in many of these regions will recognize that in many major conflicts in sub-Saharan Africa a great deal of the fighting occurs outside of towns and cities.
20. For example, in Morocco, there were only two events recorded (a bombing and an assassination), both in Casablanca; and all of the events in Ghana related to the same communal conflict in Bawku.
21. While UCDP GED relies on international wire services, many of these reports come directly from local sources. The BBC Monitoring Service, for example, provides text reported by

local radio, print and television sources. UCDP GED complements its media reports with other sources, including those that are non-English language; for more information see Kreutz (in press) and Eck and Hultman (2007). For countries such as Angola, ACLED has also supplemented with non-English language sources.

22. Multiple systems estimation (or capture–recapture methods) use two or more separately collected but incomplete lists of a population to estimate the total population size. In terms of violence, these may be lists from governments, NGOs, truth commission interviews, etc., which are then jointly analysed to evaluate the amount of overlap between them and thereafter estimate the magnitude; see Lum et al. (2010).

References

ACLED (n.d.) User document. Available at: http://www.acleddata.com/templates/modernview-blue/documents/ACLED_user_guide.pdf (accessed September 20, 2011).

Arreguín-Toft I (2007) How to lose a war on terror: A comparative analysis of a counterinsurgency success and failure. In: Ångström J and Duyvesteyn I (eds) *Understanding Victory and Defeat in Contemporary War*. London: Routledge, pp.142–167.

Balcells L (2010) Rivalry and revenge: Violence against civilians in conventional civil wars. *International Studies Quarterly* 54: 291–313.

Blattman C (2009) From violence to voting: War and political participation in Uganda. *American Political Science Review* 103: 231–247.

Butler CK, Gluch T and Mitchell NJ (2007) Security forces and sexual violence: A0 cross-national analysis of a principal-agent argument. *Journal of Peace Research* 44: 669–687.

Collier P and Hoeffler A (2001) *Greed and Grievance in Civil War*. Mimeo. Washington, DC: World Bank.

Cunningham, DE, Gleditsch KS and Salehyan I (2009) It takes two: A dyadic analysis of civil war duration and outcome. *Journal of Conflict Resolution* 53: 570–597.

Davenport C (2010) *Media Bias, Perspective, and State Repression: The Black Panther Party*. Cambridge: Cambridge University Press.

Downes A (2008) *Targeting Civilians in War*. Ithaca, NY: Cornell University Press.

Dulic T (2010) Geocoding Bosnian violence. In: Paper presented at the International Studies Association Annual Convention, New Orleans, LA, 16 February.

Eck K (2010) Coercion in rebel recruitment. Unpublished manuscript, Uppsala University, Sweden.

Eck K and Hultman L (2007) Violence against civilians in war: Insights from new fatality data. *Journal of Peace Research* 44: 233–246.

Fearon JD and Laitin DD (2003) Ethnicity, insurgency, and civil war. *American Political Science Review* 97: 75–90.

Hegre H, Gudrun Ø and Clionadh R (2009) Poverty and civil war events. *Journal of Conflict Resolution* 53: 598–623.

Humphreys M and Weinstein JM (2008) Who fights? The determinants of participation in civil war. *American Journal of Political Science* 52: 426–455.

Kalyvas SN (2006) *The Logic of Violence in Civil War*. Cambridge: Cambridge University Press.

Kopstein JS and Wittenberg J (2011) Deadly communities: Local political milieus and the persecution of Jews in occupied Poland. *Comparative Political Studies* 44: 259–283.

Kreutz J (2012) From tremors to talks: Do natural disasters produce ripe moments for resolving separatist conflicts? Unpublished manuscript, Uppsala University, Sweden.

Lum K, Price M, Guberek T, et al. (2010) Measuring elusive populations with Bayesian model averaging for multiple systems estimation: A case study on lethal violations in Casanare, 1998–2007. *Statistics, Politics, and Policy* 1.

Melander E and Sundberg R (2011) Violence in space and time: Introducing the UCDP Georeferenced Event Dataset. Unpublished manuscript, Uppsala University, Sweden.
Østby G (2008) Polarization, horizontal inequalities and violent civil conflict. *Journal of Peace Research* 4: 143–162.
Østby G, Nordås R and Rød JK (2009) Regional inequalities and civil conflict in subSaharan Africa. *International Studies Quarterly* 53: 301–324.
Raleigh C and Hegre H (2009) Population size, concentration, and civil war: A geographically disaggregated analysis. *Political Geography* 28: 224–238.
Raleigh C, Linke A and Hegre H (2009) Armed Conflict Location and Event Dataset (ACLED) codebook. Available at: http://www.acleddata.com/templates/modernviewblue/documents/Codebook_CR_Fall2009.pdf (accessed October 4, 2011).
Raleigh C, Linke A, Hegre H, et al. (2010) Introducing ACLED: An armed conflict location and event dataset. *Journal of Peace Research* 47: 1–10.
Sundberg R, Lindgren M and Padskocimaite A (2010) *UCDP Geo-referenced Event Dataset (GED) codebook*, V 1.0. Uppsala: Department of Peace and Conflict Research, Uppsala University.
UCDP (2011) Uppsala conflict data program online database. Available at: http://www.pcr.uu.se/research/ucdp/datasets (accessed 25 October 2011).
Valentino B, Huth P and Balch-Lindsay D (2004) Draining the sea: Mass killing and guerrilla warfare. *International Organization* 58: 375–407.
Weidmann N (2011) Violence 'from above' or 'from below'? The role of ethnicity in Bosnia's civil war. *Journal of Politics* 43: 1178–1190.

Author Biography

Kristine Eck is Assistant Professor in the Department of Peace and Conflict Research, Uppsala University, Sweden. In addition to conflict data collection, her current research interests are rebel recruitment, female participation in contentious politics and state repression. Her work has appeared in *International Studies Quarterly* and *Journal of Peace Research*.