Introducing the UCDP-Candidate Events Dataset and the ViEWS Outcomes dataset.

Monthly updated organized violence data in the form of events data as well as aggregated to the country-month and PRIO-GRID-month level*

Håvard Hegre[1], Mihai Croicu[1], Kristine Eck[1], and Stina Högbladh[1]

[1]Department of Peace and Conflict Research, Uppsala University

June 6, 2018

Abstract

This article discusses a set of issues that emerge when coding data on organized violence in near-real time according to the UCDP definitions. To ensure consistent application of the definitions without sacrificing the timeliness of the data, UCDP events data have two components. The 'UCDP Candidate Events Dataset' is released monthly, and contains the most recent observations in the form of candidate events that have not yet been subject to careful vetting. These candidate events are eventually included in the UCDP-GED as part of the UCDP annual update procedures. This paper also introduces the ViEWS outcome dataset that includes aggregations to the country and PRIO-GRID level at a monthly temporal resolution, based on both the UCDP-GED and UCDP-Candidate Events Dataset. The article describes the definitions, sources and procedures employed to code the candidate events as well as aggregating the data from the UCDP-GED.

*The authors would like to thank Frederick Hoyles and Naima Mouhleb for comments and assistance. The research was funded by the European Research Council, project H2020-ERC-2015-AdG 694640 (ViEWS). For more information on the project see `www.pcr.uu.se/research/views/` and `ucdp.uu.se`.

1 Outline and motivation

This paper presents two new datasets that are published at a monthly release cycle. The first is the UCDP-Candidate Events Dataset (UCDP-Candidate) that makes available monthly releases of UCDP candidate events data with not more than a month's lag for all of Africa. The second is the ViEWS outcomes dataset which combines the UCDP-Candidate with the last version of the UCDP-GED dataset into convenient aggregations to two standard units of analysis – countries and PRIO-GRID geographical locations. These aggregations facilitate a straightforward link between the UCDP-GED data and other available data at the different levels of analysis. The ViEWS outcomes dataset is also the one underlying the ViEWS forecasting project (`http://www.pcr.uu.se/research/views/`) (Hegre et al., 2018).

By moving from annual to monthly releases, these data facilitate research and policy endeavours which require up-to-date information. This paper introduces the concept of 'candidate events' so that users understand the difference between the candidate events data and the annual validated data released in UCDP-GED.

2 The UCDP-Candidate Events Dataset

UCDP-Candidate Events Dataset provides monthly updates of UCDP event data. UCDP-Candidate includes a number of candidate events recorded for all African countries based on the information that is available at the end of each month.

2.1 UCDP-GED events

The UCDP-Candidate includes the three types of conflict covered by the UCDP: state-based armed conflict (Gleditsch et al., 2002), one-sided violence (Eck and Hultman, 2007), and non-state conflict (Sundberg, Eck, and Kreutz, 2012). The dataset is based on the UCDP-GED dataset (Sundberg and Melander, 2013), where the unit of analysis is the 'event' – an individual incident of lethal violence. An event is defined as the 'incidence of the use of armed force by an organized actor against another organized actor, or against civilians, resulting in at least 1 direct death in either the best, low or high estimate categories at a specific location and for a specific temporal duration' (Sundberg and Melander, 2013, p. 524).

UCDP-Candidate departs from UCDP-GED by also including events that do not, 'in their aggregated form, constitute the UCDP's country-year datasets' (Sundberg and Melander, 2013, p. 525). The UCDP-Candidate include events carried out by dyads that do not satisfy the 25 battle-related deaths in a calender year criterion usually used as a threshold for inclusion by the UCDP. Since the UCDP identifies the stated incompatibilities by new groups on an annual basis, this criterion was also relaxed for the UCDP-Candidate events.

2.2 Scope

The combination of UCDP-GED and UCDP-Candidate together form a complementary pair with separate update schedules and different geographic coverages. The final component, which is fully consistent with UCDP's traditional annual output, has complete global coverage for the period covered by the most recent public release of the UCDP-GED. At the time of writing, the final data cover the 1989–2017 period. This component is updated along with the annual UCDP update.

The candidate component is currently restricted to Africa. As of late May 2018, for instance, data through April 2018 are available. The candidate data are preliminary in the sense described in Section 2, and are replaced with corrected and completed data at every annual release of the UCDP-GED. Hence, the candidate component only covers the most recent 4–18 months; candidate events for the preceding year become obsolete upon the release of the annual data.

2.3 Sources

The conflict data are in their entirety based on the UCDP-GED dataset (Sundberg and Melander, 2013; Croicu and Sundberg, 2013), which again are based on the UCDP suite of conflict data (Melander, Pettersson, and Themnér, 2016). Historical data covering 1989–2016 are extracted from the (published) UCDP-GED version 17.1 (Croicu and Sundberg, 2013).

The UCDP-GED raw data are publicly available through the UCDP-GED API (Croicu and Sundberg, 2013).[1]

2.4 How UCDP-Candidate and UCDP-GED differ

UCDP-Candidate has exactly the same structure as the annually updated UCDP-GED. It differs somewhat in content, however, since the update schedule precludes the same level of scrutiny as the UCDP applies in the annual update. This section outlines the sources of these differences.

The data collection in the candidate stage of the coding process involves retrieving relevant articles from a specified set of sources.[2] This process returns a set of possible events that is then basis for coding according to the UCDP's criteria for coding the different types of organized violence. The coding entails evaluating a number of criteria summarized in Table 1.

The set of candidate events that underlies UCDP-Candidate deviates from UCDP-GED in several ways. Many UCDP definitions (see Gleditsch et al., 2002) are applicable only on a calendar-year basis, and the final UCDP-GED dataset can only be compiled after the end of the year. The UCDP-Candidate consequently relaxes the UCDP requirement of a 25-battle related deaths threshold in order for a conflict to be recorded, as well as the requirement of a 'stated goal of incompatibility'.

[1] Usage of the API is described in `http://ucdp.uu.se/apidocs/`; the data are available as versions 5.0 (1989–2015), 5.9.99 (January 2016 to February 2017) and 5.9.201703 (March 2017).

[2] A Factiva search provides the major bulk of items. This body is complemented by country-specific sources such as local NGO reports, reports of the Secretary General on different UN missions, articles from local news agencies, Africa Confidential, Africa Research Bulletin, Amnesty International, Crisis Watch, Human Rights Watch, International Crisis Group reports.

	State-based	Non-state	One-sided
Who?	Government vs government or government vs formally organized groups (rebel groups)	Formally organized groups (rebel groups), informally organized groups (communal groups), same level of organisation	Governments or formally organized groups (rebel groups)
What?	Battle-related deaths (normal warfare)	Deaths in communal violence, deaths in clashes between rebel groups	Deaths in violence against civilians (genocides, massacres)
Why?	Stated incompatibility, government or territory or both	N/A	Targeting civilians

Table 1. UCDP coding criteria by conflict type

Since UCDP-Candidate is designed to serve as a candidate estimate of recent conflict intensity, we have developed a coding procedure aimed at making the monthly candidate event sample as close in content to the final dataset as possible.

Practical constraints, however, imply that the very strict requirements in terms of known and clear parties to the conflict are relaxed as long as there are sufficiently strong indications that such events have a high likelihood of inclusion in the final UCDP datasets at the end of the year. The decision whether to include a report in the UCDP-Candidate data is taken by the UCDP coder and project manager in charge of data collection. In the UCDP-Candidate releases, these relaxed-criteria events are marked with a code status. A slight majority of the candidate events that go into UCDP-Candidate are straightforward instances of UCDP-GED that will make it to the final version, and are immediately flagged as 'clear' events. Approximately half of them, however, are flagged for further investigation before being included in UCDP-GED proper. This two-stage process allows the UCDP to provide monthly updates without sacrificing the thorough scrutiny of ambiguous cases required for the annual update.

The careful scrutiny in the final stage of coding can lead to events being retained, discarded, or recoded. One example of recoding concerns the type of violence. In many conflict regions all types of UCDP organized violence are present at the same time. For example, the violence that broke out in the Kasaï region of the Democratic Republic of Congo in August 2016 was a mixture of the government fighting *Kamwina Nsapu*, a government-sponsored militia *Bana Mura* fighting *Kamwina Nsapu*, all three armed groups targeting the civilian population, as well as communal violence between the Lulua-Luba and Chowe-Pende ethnic groups. When reports from the first screening of sources only report fatalities from political violence in the Kasaï region, it is likely that they indicate events that end up in the annual UCDP-GED release. However, without more detailed information UCDP cannot attribute a dyad or specify the type of violence. Such events are included in the UCDP-Candidate Events Data, but with a code-status flag 'check type of violence'.

In other cases the initial reports indicate deaths from clashes between the government and insurgents, but in order to know the exact pair of actors involved, the UCDP often has to retrieve information matching the location of the reported incidents with UCDP's records of territorial location of groups. Such records would be retained but flagged as 'check dyad'.

As shown in Table 1, all rebel groups need to state their incompatibility against the government

before meeting the state-based armed conflict definition. Such statements are frequently tricky to find, and for new groups fighting the government UCDP will often have to flag events as 'check incompatibility'.

Another feature of events data that often will change from candidate version to final version are the number of deaths. This is particularly true for events where an article suggests that there have been deaths without specifying the number and for 'summary events' giving a total death count for a longer time period; where the second round of coding involves trying to confirm death estimates or derive deaths in single-day events from the given summary. UCDP flags such events as 'check deaths'.

Recoding also occur for events flagged as 'check vague or biased source'. An example of such reports is an Al-Shabab website claiming government deaths (in such a case, translated to English by BBC Monitoring). Here, the UCDP routinely looks for corroborating claims of deaths or at least of fighting from other sources before deciding whether to include the event in the final dataset.

Moreover, the exact geographic location sometimes changes when it was impossible to find the exact location of an event in the candidate coding stage. In such cases, UCDP codes the location as the relevant administrative region and flags the event as 'check geography'.

Finally, a generic 'check' flag is used for some additional uncertainties, such as investigating whether two events on different dates with very similar descriptions really are two distinct events.

2.5 Citations

Users of the ViEWS-Candidate dataset should cite this article as well as the article introducing UCDP-GED (Sundberg and Melander, 2013).

3 The ViEWS Outcomes dataset

The ViEWS Outcomes dataset provides monthly aggregations of the UCDP-GED and the UCDP-Candidate datasets; aggregating all recorded events up to the month they occur within. ViEWS automatically retrieves these data from the UCDP API each month and aggregates to the units of analysis described below.

In the current version, the ViEWS Outcomes data are aggregated to two levels of analysis: PRIO-GRID version 2.0 (Tollefsen, Strand, and Buhaug, 2012), and countries (Gleditsch and Ward, 1999). We retrieved the grid-level structure directly from the PRIO-GRID API to ensure full compatibility with the variables provided by PRIO-GRID. Country level information is taken from the latest version of CShapes (Weidmann, Kuse, and Gleditsch, 2010). In future releases, ViEWS Outcomes data will also include aggregations to the actor-month level.

3.1 Aggregation procedures

Note that the country and PRIO-GRID levels are not fully compatible with each other, and aggregation of the PRIO-GRID-monthly dataset to country level will not result in a perfect copy of

the country-month dataset. There are two reasons for the imperfect compatibility between the two datasets:

- PRIO-GRID cells may span across the border of two or more countries. In that case, all events spatially contained in that PRIO-GRID cell are aggregated to it, ignoring which country they actually took place in. On the other hand, such events are assigned to the appropriate country where the event took place in the country-month dataset.[3]

- PRIO-GRID cells are purely geographical entities, and exist for the entire duration of the dataset. Countries, on the other hand, occasionally enter and exit the international system. Only those months in which a country has existed in the Gleditsch and Ward (1999) country list are included in the country-month datasets.

The UCDP-GED and UCDP-Candidate both aspire to a fine-grained coding resolution spatially (village/town) and temporally (day). In practice, the recorded locations are heterogeneous in nature. Even if UCDP is certain that an event actually took place, spatial or temporal information may be missing or insufficient to locate an event precisely. UCDP does not perform list-wise deletion for such events, but rather assigns them to a place-holder location (the centroid of the administrative region or the country) or to a imprecise date (UCDP provides two dates for each event; the earliest possible date an event took place and the last possible date an event took place). Approximately 16% of all events are coarser than a grid cell, and just under 2% as coarser than a calendar month. This is indicated by a set of precision scores in both UCDP GED and the UCDP-Candidate datasets.

Since the place-holder locations can misrepresent the actual location of the events (in some cases the assigned centroid coordinates are over 1000 km from the nearest plausible location of fighting), the ViEWS Outcomes data exist in two editions with different modes of aggregations at the PRIO-GRID-month level.

The first is the **baseline** edition.[4] This will only include those events that are sufficiently precise to be clearly attributable to a single PRIO-GRID cell. This aggregates all events that the UCDP codes as having precision scores 1 (exact location), 2 (assigned location within a 25-kilometer radius of the most likely actual event location), 3 (location assigned to the centroid of the second-order administrative unit the event occured within)[5] and 4 (linear or fuzzy area known).[6] We aggregate all these events using their end date, as this is the date at which UCDP is fully certain the event has taken place.

[3]PRIO-GRID provides a 1:1 cell-to-country correspondence using a majority rule, i.e. assigns the grid cell to the country taking up the largest area in the cell. We do not make use of this rule to aggregate events to the country-month; we instead use UCDP's own coding of the country (Tollefsen, 2012).

[4]The `ViEWS_Outcomes_baseline` dataset is available at `http://pcr.uu.se/research/views/data/downloads`.

[5]We include these data in the baseline, as second-order administrative units, on average, have a smaller area than a PRIO-GRID cell.

[6]This category was included after the investigation of the underlying events for a sample of 25% of all events included in this category. A vast majority describe relatively small, but non-point/non-formal areas such as borders, roads, rivers, informal and relatively small regions.

The second is the **multiply imputed** dataset, and contains all the events produced by UCDP.[7] The spatially imprecise ones are multiply imputed (5 times) using the procedure described in (Croicu and Hegre, 2018) while the temporally imprecise ones are also multiply imputed (5 times) by sampling from a discrete uniform distribution consisting of all the months between the start and the end month. Downloads of this edition include all five imputed datasets. Since no event is more spatially imprecise than the area of a country, no imputed edition is produced for the country-month dataset.

The aggregation process is mostly straightforward, as UCDP will always assign one single pair of coordinates, one country and a set of dates to all events. The only technical complication occurs if an event's coordinates are situated exactly on the boundary between two PRIO-GRID cells.[8] By convention, we assign it to the grid situated to the North, respectively West of the boundary. This also applies to events precisely placed on the corner of four such cells – these are attached to the cell situated to the North-West of the corner. This ensures consistency with the other UCDP datasets that all use the same rule where applicable.

3.2 Variables included in the ViEWS Outcomes dataset

The ViEWS Outcomes dataset makes available the following information for each cell month and country month:

- `ged_dummy_sb, ged_dummy_ns, ged_dummy_os` – one dummy variable for each outcome indicating whether there is any state-based conflict, non-state conflict or one-sided violence in the cell month/country month. One dummy is produced for each category.
- `ged_count_sb, ged_count_ns, ged_count_os` – event counts for each outcome for each cell month/country month in each category of conflict (state-based, one-sided and non-state violence)
- `ged_best_sb, ged_best_ns, ged_best_os` – sums of all fatalities in the 'best' category for each cell month/country month in each category of conflict (state-based, one-sided and non-state violence).
- `ged_best_sb_lag1, ged_best_ns_lag1, ged_best_os_lag1, ged_count_sb_lag1, ged_count_ns_lag1, ged_count_os_lag1` – first-order spatial lags for the above event counts and fatality sums (i.e. for the neighbors of the cell/country and for the neighbors of neighbors of the cell/countries). Queen contiguity was used for both countries and cells - i.e. sharing a single point makes one a neighbor. For countries, the boundaries presented in Weidmann, Kuse, and Gleditsch (2010) have been used.

Code for aggregating data as well as computing all the measures is made available with the dataset.

[7] The `ViEWS_Outcomes_imputed` dataset is available at `http://pcr.uu.se/research/views/data/downloads`.
[8] 9% of events have this property.

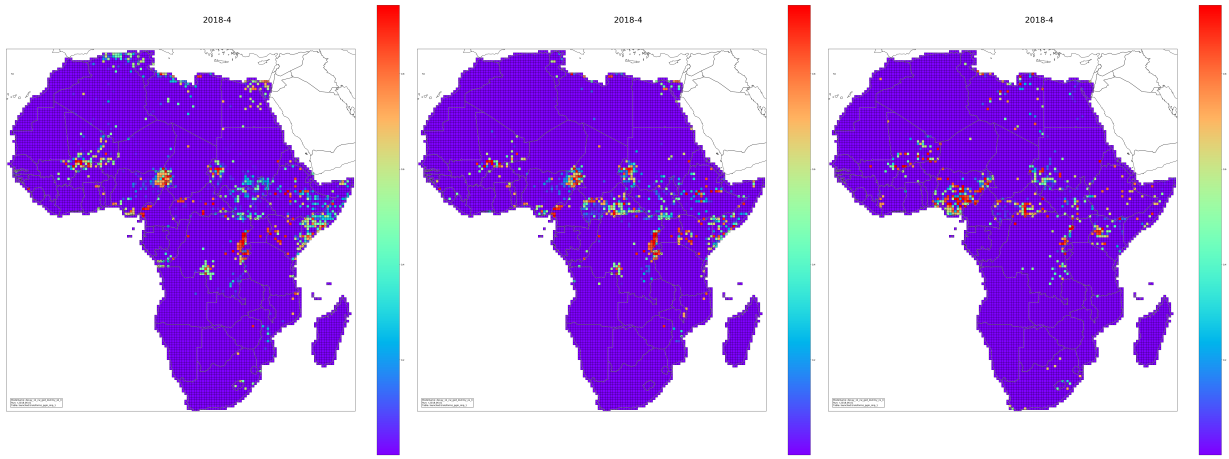

Figure 1. UCDP-Candidate events aggregated to the PRIO-GRID level as of April 2018. State-based conflicts (left), one-sided violence (middle), and non-state conflict (right). Locations with events in April have red color, and locations without events for several years have purple color. The color scale to the right refers to a decay function of the number of months since event in the PRIO-GRID cell, with halflife of 12 months ($y = e^{-msc/12}$ where msc is the number of months since event)

3.3 Availability

Subsequent monthly and annual updates as well as the imputed datasets are available from `http://www.pcr.uu.se/research/views/data/downloads`.

The ViEWS project combines the UCDP-Candidate and ViEWS data with a set of predictors used in their forecasting system. These variables are made publicly available as separate, monthly updates at `pcr.uu.se/research/views/data/replication-data`. The UCDP Candidate Events data are released monthly through the UCDP API.

3.4 Citations

Users of the ViEWS Outcomes non-imputed data should cite this article as well as the article introducing UCDP-GED (Sundberg and Melander, 2013). Users of the imputed versions of the data should also cite Croicu and Hegre (2018).

4 Descriptives

Figure 1 shows the distribution of recent events across Africa up to and including April 2018. Locations with orange or red color have had conflict events over the past six months. Locations with light blue or green colors have had events during the 2015–2017 period. Locations with purple color has had no events over the past 5–10 years.

Funding

The research has been funded by the European Research Council project H2020-ERC-2015-AdG 694640 (ViEWS).

References

Croicu, Mihai and Håvard Hegre (2018). *A Fast Spatial Multiple Imputation Procedure for Imprecise Armed Conflict Events*. Typescript, Uppsala University/ViEWS. http://www.pcr.uu.se/research/views/publications.

Croicu, Mihai and Ralph Sundberg (2013). *UCDP Georeferenced Event Dataset Codebook Version 4.0*. Typescript, Uppsala Conflict Data Program. URL: http://www.pcr.uu.se/research/ucdp/datasets/ucdp_ged/.

Eck, Kristine and Lisa Hultman (2007). "One-Sided Violence against Civilians in War: Insights from New Fatality Data". In: *Journal of Peace Research* 44.2, pp. 233–246.

Gleditsch, Kristian S. and Michael D. Ward (1999). "A Revised List of Independent States since the Congress of Vienna". In: *International Interactions* 25.4, pp. 393–413.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand (2002). "Armed Conflict 1946–2001: A New Dataset". In: *Journal of Peace Research* 39.5, pp. 615–637.

Hegre, Håvard, Marie Allansson, Mike Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Naima Mouhleb, Sayeed Auwn Muhammad, Desirée Nilsson, Håvard Mokleiv Nygård, Gudlaug Olafsdottir, Kristina Petrova, David Randahl, Espen Geelmuyden Rød, Nina von Uexkull, and Jonas Vestby (2018). *ViEWS: A political Violence Early Warning System*. Typescript, Uppsala University. www.uu.se/pcr/research/views/publications.

Melander, Erik, Therése Pettersson, and Lotta Themnér (2016). "Organized violence, 1989–2015". In: *Journal of Peace Research* 53.5, pp. 727–742.

Sundberg, Ralph, Kristine Eck, and Joakim Kreutz (2012). "Introducing the UCDP Non-State Conflict Dataset". In: *Journal of Peace Research* 49, pp. 351–362.

Sundberg, Ralph and Erik Melander (2013). "Introducing the UCDP Georeferenced Event Dataset". In: *Journal of Peace Research* 50.4, pp. 523–532. DOI: 10.1177/0022343313484347.

Tollefsen, Andreas Forø (2012). *PRIO-GRID Codebook*. Typescript, PRIO. URL: http://file.prio.no/ReplicationData/PRIO-GRID/PRIO-GRID_codebook_v1_01.pdf.

Tollefsen, Andreas Forø, Håvard Strand, and Halvard Buhaug (2012). "PRIO-GRID: A unified spatial data structure". In: *Journal of Peace Research* 49.2, pp. 363–374. DOI: 10.1177/0022343311431287. eprint: http://jpr.sagepub.com/content/49/2/363.full.pdf+html.

Weidmann, Nils B, Doreen Kuse, and Kristian Skrede Gleditsch (2010). "The geography of the international system: The CShapes dataset". In: *International Interactions* 36.1, pp. 86–106.