An Integrated Picture of Conflict

Eric T. Dunford[*] David E. Cunningham[†,‡] Karsten Donnay[§]
David Backer[¶]

Spring 2020

Abstract

Growth in event datasets is fostering research about patterns, dynamics, causes, and consequences of conflict. Studies typically rely on a single dataset. Instead, we advocate integrating multiple datasets to improve measurement and analysis. We have generated an integrated dataset covering all violent events for Africa from 1997-2018 from three leading datasets (ACLED, UCDP-GED, and GTD). Our approach involves both pre-processing the data so that they are comparable and using an automated approach to produce an integrated dataset that is transparent and reproducible. Through examining these integrated data, we find substantial overlap across these three datasets. At the same time, each dataset includes events that conceptually should be captured in the other datasets, but are not. Thus, we view these integrated data as offering a better measure of violent conflict. A statistical analysis shows that geographic features frequently used in analyses of the location of conflict events — including the distance from the capital or a border, terrain, economic development, and population–have different effects on the incidence and frequency of conflict events when using integrated data as compared to individual datasets. These illustrations highlight the potential for integration to advance conflict research by yielding a more complete and accurate picture of activity, which has repercussions for both descriptive and theoretical findings. Integration is likely to be increasingly worthwhile as event datasets proliferate, expand in coverage, and exhibit wider applications.

[*]McCourt School of Public Policy, Georgetown University, Washington DC, USA
[†]Department of Government & Politics, University of Maryland, College Park, MD, USA
[‡]Peace Research Institute Oslo, Norway
[§]Department of Politics and Public Administration, University of Konstanz, Germany
[¶]Center for International Development & Conflict Management, University of Maryland, College Park, MD, USA

Introduction

Researchers interested in questions about the patterns, dynamics, causes, and consequences of conflict have ready access to detailed information on the location and timing of political violence. These data are recorded in an emergent array of event history datasets capturing violent activity at country, regional, and/or global scales.[1] These datasets facilitate empirical studies, including fresh lines of inquiry, especially leveraging the distinctive geographical and temporal granularity of the data. To date, however, researchers have not grappled sufficiently with the potential implications of the array of datasets for the measurement and analysis of conflict activity, let alone capitalized on the array for those purposes. Studies that employ multiple conflict event datasets simultaneously are rare and usually of narrow scope (e.g., Findley and Young (2012); Fortna (2015); Weidmann (2015); Polo and Gleditsch (2016)). Instead, the default is to rely on a single conflict event dataset, often without mentioning available options.

We argue that integrating information from multiple different conflict event datasets has the potential to greatly improve the measurement of conflict and to facilitate improved understanding of the patterns, dynamics, causes, and consequences of conflict. The basic rationale of integration is that it yields a more complete and accurate measurement of the concept being measured. At a minimum, integration offers a valuable means to evaluate the sensitivity of results as a function of how activity is conceptualized and measured. We are not advocating that integration ought to be employed universally. Integration may be

[1] For example, the xSub project (Zhukov et al., 2019) has assembled geo-referenced event data on violence and protests from 21 different sources.

unnecessary for pursuing a given topic, if one dataset perfectly and demonstrably captures a particular conceptualization of interest. Moreover, integration is inappropriate absent multiple datasets with relevant coverage of a concept of conflict of interest to analysis. In practice, however, conflict event datasets often exhibit coverage that is both overlapping and incomplete, which makes integration worthwhile to consider.

The conflict research community has shown an interest in integration, as shown by the articles cited above and additional efforts to manually integrate data from different event datasets. Yet the efforts at integration to date remain limited, which we believe reflects the stringent requirements and substantial challenges of integrating event datasets. Ideally, integration ought to (*i*) avoid double-counting, (*ii*) triangulate to confirm information or acknowledge uncertainties; and (*iii*) identify gaps in one dataset that another dataset can fill. Therefore, integration entails considering whether different datasets record cases that are the same, similar or unique—all vital to measurement. Systematic comparisons of datasets along these lines are hard to conduct manually at scale with the appropriate level of accuracy, transparency, and reproducibility.

We have assembled a novel dataset of violent conflict activity across Africa from 1997-2018, drawing on three leading data resources in the study of armed conflict: the Armed Conflict Location and Event Data (ACLED) (Raleigh et al., 2010), the Uppsala Conflict Data Project-Georeferenced Event Data (UCDP-GED) (Sundberg and Melander, 2013), and the Global Terrorism Database (GTD) (National Consortium for the Study of Terrorism and Responses to Terrorism (START), 2013). These three datasets are widely used in published research. Each dataset seeks to measure the location and timing of a specific set of conflict events, as each initiative defines them. Integrating these datasets

presents multiple challenges. First, the datasets have different conceptions of what an event is. Second, in some cases the datasets use different rules to report the geolocation where an event takes place. Third, the datasets have different descriptive labels for the types of events they code. In light of these challenges, we argue that proper integration requires deep knowledge of the data projects that allows for identifying where and how they overlap (or not) and a transparent, reproducible, and efficient process for integrating the data. The integration presented in this article addresses both of these needs. We pre-process the data from each dataset to account for differences in the conceptualization of events, geolocation rules, and labeling of event types. Then, we integrate the pre-processed data using an automated approach that has been shown to work in an accurate, transparent, reproducible, and efficient manner (Donnay et al., 2019)

This article first outlines the process of integrating violent events from ACLED, UCDP-GED and GTD. Next, we describe the integrated data that is produced. We compare our integrated data to the raw data from each dataset. This comparison reveals substantial overlap across the three datasets—in our preferred integration, 18.47% of total entries match to an entry in at least one of the other two datasets. We also find that each dataset captures unique events. These results are tangible evidence that integration provides a more complete and accurate picture of violent conflict activity. We look in more detail at the differences by comparing the number of events in each dataset and the integrated data by month and administrative unit. This analysis reveals variation across the integrated data and the three datasets in the degree of under and over-counting. Researchers frequently examine the occurrence or count of conflict events aggregated to administrative unit-months. The variation we find reinforces our conclusion that better measurement from integrated data will have

implications for such descriptive analysis of outcomes. In addition, we conduct a basic statistical analysis in which we examine the relationships between conflict events and several common features of geographical locations. Our analysis shows that the nature of these relationships varies across the integrated data and the different datasets. Taken together, these empirical illustrations highlight the utility of integration of conflict event data for improving measurement and analysis of the patterns, dynamics, causes, and consequences of conflict.

Integrating Conflict Data in Theory

We see three key benefits of integrating conflict event data. The first is to enhance measurement of conflict activity. Individual datasets may omit specific types or instances of events. Omissions can be intentional: events fall outside of the scope of the conceptualization of conflict for a given dataset. Alternatively, omissions can be unintentional by-products of a data collection process. Some events omitted from one dataset may appear in another dataset. Also, different datasets may contain complementary or contradictory details about the same event. All these aspects of measurement affect the ability to describe the prevalence, distributions, trajectories, and other patterns of conflict activity. Integration can address those needs by providing a more complete and accurate picture of activity.

The second benefit is to bolster evaluation of theoretical arguments about causes and consequences of conflict. Analyses about conflict increasingly employ measures of activity derived from event datasets. Integration that enables better measurement of this activity improves the accuracy, validity and utility of results.

The third benefit is to assist examinations of dynamics of relationships among types

of conflict. Many conflict event datasets focus on select types of conflict. Yet different types of contentious activity can occur in close proximity to and influence one another. Integrated data positions researchers to analyze more types together. Refining the measurement of these types is a vital foundational step.

Our empirical illustrations specifically demonstrate the first two benefits. We do not illustrate the third benefit within this article, but follow up on this point in the conclusion and our discussion of other work.

A necessary condition for integration is event datasets with substantial overlaps in coverage. When datasets should (as a matter of intentions and concrete conceptualizations) or could (allowing for coding differences) capture the same events occurring within the same area during the same time period, integration is both feasible and warranted.[2] In our situation, ACLED, UCDP-GED, and GTD currently cover all of Africa spanning a 21-year period. This overlap defines the scope of the integrated dataset we present and analyze here.[3] Each dataset overlaps with at least one of the other datasets in terms of various types of violent events they capture. The datasets are regularly used in peer-reviewed publications—signs the research community views them as being of high quality.

Integration requires making determinations about whether dataset entries concern the same event or different events, complicated by the fact that multiple datasets may record the same event, with similar but not identical information about attributes of the event. Making these determinations on a large scale is difficult. ACLED, UCDP-GED, and GTD

[2]Absent any overlaps, datasets can simply be pooled, splicing together discrete segments. Measurement and analysis is then an artifact of coverage of each constituent dataset for a given area, period and/or type.

[3]At the time of writing, only two of the datasets (GTD and UCDP-GED) that we use are global in scope. We leave to future research the task of integrating conflict event data globally, which may capitalize on additional datasets extending coverage.

collectively report over 155,000 entries for Africa from 1997-2018, recording many events in close geographical and temporal proximity to one another. Evaluating manually whether these events are unique or duplicates is cumbersome, time-consuming, and hard to replicate.

To address these challenges, Donnay et al. (2019) developed an automated protocol that can be used to integrate datasets containing information on the location and timing of events, based on inputs that researchers supply. The procedure, known as MELTT ("Merging Event Data by Location, Time and Type"), requires two main inputs that require researchers to be explicit about the assumptions they make when integrating data.

First, a researcher must define a spatiotemporal window within which entries compared across datasets are considered as candidate matches. The logic hinges on the reliability of measurements: only entries recorded in different datasets as being proximate in space and time are likely to reflect the same event. A researcher specifies this window based on knowledge of the datasets and expectations about the degree of uncertainty or error in recording locations and timings of events. Bounding the plausible scope of matches limits the number of comparisons of entries that are performed, reducing the processing time. Such efficiencies are especially vital when performing integration at scale.

Second, the researcher must define relevant taxonomies, which indicate how the codes for a variable in one dataset correspond to the codes for a variable in another dataset. Datasets differ in terms of the variables included and the codes possible for each variable. Defining rules of correspondence that generalize across coding schemes allows for systematic comparison of similar attributes of entries in multiple datasets, refining the identification of candidate matches. A researcher again relies on knowledge of datasets—referencing the codebooks and other information—to generate a taxonomy. Multiple taxonomies can be

defined for different sets of variables.

The outputs of the MELTT procedure identify candidate matches and consolidate entries deemed to be duplicates. The integrated dataset retains all information from the input datasets. For each event duplicated across input datasets, a row in the integrated dataset includes the information about this event on every variable in the relevant input datasets.

Donnay et al. (2019) show that MELTT is reliable and consistent by conducting simulations on synthetic data and through qualitative manual validation of outputs of the integration of ACLED, UCDP-GED, GTD, and the Social Conflict Analysis Database (SCAD) (Salehyan et al., 2012) for three select country-years in Africa (Nigeria 2011, Libya 2014, and South Sudan 2015). We refer the reader to their work for a more in-depth explanation of the methodology. Here, we leverage MELTT to generate the first integrated dataset covering violent conflict events in all of Africa for a 21-year period, maximizing the overlap of ACLED, UCDP-GED, and GTD.

Integrating Conflict Data in Practice

Computationally integrating event datasets at scale is possible. Doing so, however, requires that the datasets are actually comparable, in terms of their coverage and other key attributes. In this section, we highlight three challenges to integrating conflict event data in particular: (*i*) spatial imprecision, (*ii*) temporal imprecision, and (*iii*) variation in the unit of analysis. We outline ways to resolve each issue. These decisions ultimately take the form of assumptions that can be adjusted given the relative ease with which an integration task

can be re-run using an automated protocol, such as MELTT. Finally, we pre-process and integrate ACLED, UCDP-GED, and GTD and explore the resulting data.

Spatial imprecision

One prominent concern when integrating geo-referenced data is the precision of the geo-coordinates specifying the event locations. The MELTT algorithm already adjusts for spatial imprecision due to minute differences in geo-referencing software (Donnay et al., 2019). An important consideration is that precise point locations cannot always be determined based on available information. Rather, information may reveal only that the location was within a geographic area, such as an administrative division or near a border. Across each dataset, the precision code captures this uncertainty in measurement. Alternatively, an event could have a large geographical scope (e.g., a battle spread over a wide area). Datasets typically assign to either type of event the coordinates of an associated location, such as a centroid or a population center (e.g., the capital of a country or administrative unit). Differences in assignment conventions across the datasets can therefore yield misleading reporting of locations.

To address this challenge, we use location coordinates as serviceable approximations for the level of aggregation of each event entry. We break entries up into four bins that correspond to an entry's level of geo-coded precision and integrate entries located in each bin separately. Entries coded as having the most precise locations, according to each dataset's coding scheme, are treated as precisely geo-coded events and the full extent of the MELTT algorithm is employed to evaluate matches. For entries with less precision, corresponding to higher levels of spatial aggregation, we assign the entry to a common centroid location (i.e.,

Table 1: Event precision codes and spatial unit bin assignment scheme

Dataset	Precision Code	Centroid Assignment
acled	1	Event Unit
ged	1 - 2	Event Unit
gtd	1	Event Unit
acled	2	Second Adminstration Unit
gtd	2 - 3	Second Adminstration Unit
ged	3	Second Adminstration Unit
acled	3	First Adminstration Unit
gtd	4	First Adminstration Unit
ged	4 - 5	First Adminstration Unit
gtd	5	Country Unit
ged	6	Country Unit
ged	7	Excluded (international waters)

the centroid for the respective administrative unit) and only block temporally when integrating with MELTT. This results in four separate integrated sets: (event-level) precise entries integrated with other precise entries, and less precise entries (adm2, adm1, and country) integrated with other entries on their respective unit of aggregation. Table 1 summarizes the assignment scheme, given the precision level as recorded in each dataset.

Temporal imprecision

Likewise, event data can also be measured with temporal imprecision. This form of imprecision is driven by conflict activity that cannot be pinpointed to a single day. Activity may transpire over a number of days that is clearly related. A common way to capture such uncertainty is to allow entries to cover spans of time by assigning them start and end dates (hereafter "episodes"). Episodes present a challenge for integration. Any day within an episode window could plausibly coincide with the single day of an event entry included in another dataset; in fact, multiple days within an episode window could plausibly coincide

with the timing of multiple single-day event entries included in another dataset.

For purposes of facilitating the comparison of entries across datasets, one strategy is to unpack an episode into separate event-day entries so that all entries have identical temporal units. That is, the information recorded about an episode could be expanded so that that each day contained within the time window contains a copy of the relevant data—effectively turning the episode into multiple single-day events. The challenge is that episodes vary substantially in duration: with some episodes accounting for large windows of time (i.e. a month or year). Unpacking these episodes can lead to a proliferation in event entries, creating its own distortions for count and severity metrics.

We view a researcher's tolerance of episodes as an essential parameter (assumption) for any integration in which one or more of the datasets permits temporal uncertainty in the form of episodal entries. Specifically, we recommend that a threshold be established that determines when an episode window is sufficiently short enough in duration to be unpacked without introducing additional distortions. For example, a researcher might only consider episodes that are no longer than a week as viable candidate episodes, discarding the rest. The logic would be that episodal entries extending in duration beyond this threshold are too ill-defined and lack the necessary characteristics to be considered event data.[4]

In addition to episodes, datasets also record temporal uncertainty in the form of a precision field, similar to geospatial imprecision. Temporal imprecision is a categorical variable that specifies the window (in days) when the event could have possibly occurred. The current version of MELTT deals with this form of imprecision by allowing the specification

[4]Episodes are the exception. Only 19.6% of the 152,616 entries contained within the UCDP-GED reflect episodes and just 4.5% of entries reflect episodes that extend beyond 7 days.

of temporal windows when considering entries as candidates to match. We explore the implications of varying this temporal window below.

Variation in the unit of analysis: aggregating entries to events

Event-based data denotes a data generating process where unique information is recorded regarding activity that occurred at a specific spatiotemporal location. For most event datasets, however, a single observation in the data does not consistently correspond to a single event occurrence. In some instances, events are disaggregated into individual entries capturing unique perspectives (or angles) of an event. Some datasets are explicit when these instances of disaggregation occur. For example, on December 18, 2014 at lon-34.117064, lat-31.212225, GTD records two explosive devices being discovered and defused in Sheikh Zuweid town in Egypt. One entry captures a bomb planted along the road, and the other captures a bomb planted inside a tunnel under the road. Both entries point to the same event but are disaggregated by the strategic placement of the explosive devices. GTD records when events like these are disaggregated into separate entries through the provision of a "related" field.

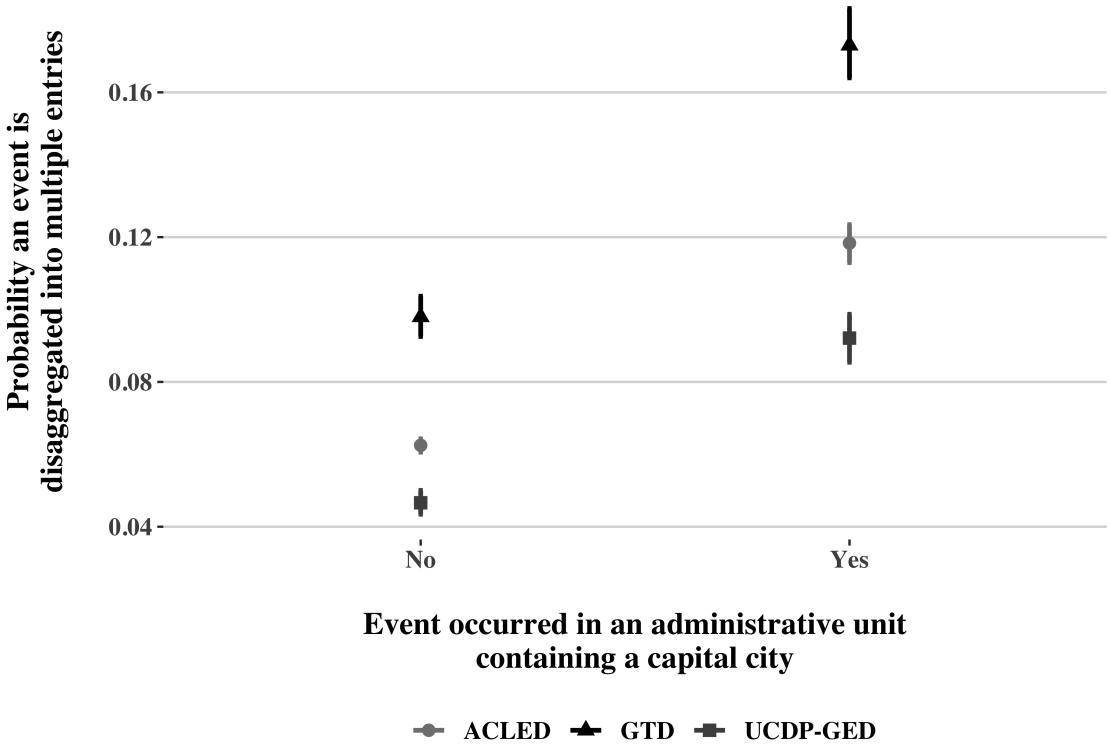
For other datasets, however, event disaggregation is less explicit. Some datasets categorize the same event into different event types with the aim of capturing different types of violence occurring simultaneously during the same attack. For example, on 2012-03-19 at lon-45.33, lat-2.055 ACLED reports a shootout between Al Shabaab militants and military officers while at the same spatiotemporal location reports Al Shabaab fighters firing mortar rounds into the town, killing four persons located within the town. The entries are recorded as "Battle - No change of territory" and "Remote Violence" event types, respectively. Both entries capture perspectives on different forms of the violence perpetrated by the same actors

taking place during the same event.[5] Similarly, other datasets disaggregate by actors or dyadic exchanges, capturing variety in the targets and engagements. For example, on 2016-04-22 at location lon-44.400876, lat-33.340582 UCDP-GED records Islamic State militants detonating a suicide bomb at a Shiite mosque, killing 9 civilians and 2 military personnel. As before, this single event occupies two entries: one capturing the rebel-state interaction and the other the violence against civilian component.

These examples illustrate a subtle reality when using conflict event datasets: the unit of observation can vary given the availability of information used to code a given event. When more information is available, these datasets often disaggregate beyond the event level, capturing features of an event as multiple entries. In theory, this practice provides researchers with more information on the forms violence takes. In practice, however, when and where such disaggregation is possible tends to track with locations where information is more available, such as urban centers (Croicu and Kreutz, 2017). Figure 1 plots the probability of a precisely coded event being disaggregated into multiple entries given that the capital is located in the administrative unit where the event occurred. The figure shows that, on average, there is a 5.7 percentage point increase in the likelihood of a precise event being disaggregated into multiple entries if the event occurred in an administration unit containing a capital city. This effect is consistent across each of the datasets we consider in this article. An implication is that analysis based on data in which events are disaggregated is prone to skew toward calculating higher ratios between the frequency of violence in urban areas as compared to rural areas.

[5]Both entries are recorded as exchanges between the same actors: Al Shabaab and the Military Forces of Somalia (2004-2012).

Figure 1: Information availability and event disaggregation



The figure plots the predicted probability that an event is disaggregated into multiple entries given the presence of a capital city in the first-order administrative division in which the event occurred. See Appendix A.1 for complete model details.

With the above in mind, standardization of the unit of observation is a prerequisite for mapping event datasets onto one another. Since the concept of an "event" varies both within and across datasets, we opt to aggregate all entries to the same spatiotemporal unit: the latitude-longitude-day (hereafter "latlon-day"). Specifically, we aggregate all entries coded with both the same geo-spatial coordinates and the same day into a single observation. Note that we only aggregate entries with precise geospatial coordinates. Entries with lower precision are forced to the same spatial location by definition. Thus, aggregation would force disparate events into the same entry.

To facilitate aggregation to the latlon-day, we standardize how an event is recorded so that information loss is minimized. Specifically, we generate indicator variables that activate when specific actor, event, or fatality information is recorded within any of the available entries being aggregated. These metadata are essential to the integration logic at the heart of MELTT; however, we use these metadata differently from the strategy outlined in Donnay et al. (2019) by making each dimension shallow (a single level taxonomy) rather than deep (a granular to broad taxonomy).

First, event information is reduced down into one of two categories: "violence" or "civilian violence". The former category encompasses all events involving interactions between armed combatants (state-on-rebel, rebel-on-state, and rebel-on-rebel), while the latter category encompasses all aggression by state or rebel actors directed toward civilians and other non-combatant populations. Substantial differences in event coding schemes across datasets precluded one-to-one correspondences of metadata for event types. Thus, we opted to reduce the consolidation of information on event types to the lowest common denominator. Our categorizations are reasonable since each of the three datasets codes entries in a manner that fundamentally captures whether violence is between combatants, or against civilians.

Next, we generate 5 categories capturing information on the actors involved in the entry: government, violent, ethnic, religious, and civilian. Raw comparisons of actors across datasets are often unreliable since they exhibit considerable differences in spelling and naming conventions. Instead, we codify actor types using a dictionary of common terms (see Appendix A.2). For a given entry, an actor category takes on the value of 1 if a relevant type is reported, or 0 otherwise. Both actors (i.e., the perpetrator and target) recorded in

each entry are coded, in separate fields. A shallow taxonomy triggers if any of the actor types is flagged in either actor field. For example, an interaction between an Islamic rebel organization and the government of a state would trigger the "violent actor", "religious actor", and "government actor" fields. When actor information is unknown, all actor fields are triggered. Since the actor composition is unknown, all combinations are possible.

Finally, we capture information on the severity of an event by generating three severity categories: none, low, and high. Specifically, "none" corresponds to fatalities = 0; "low" corresponds to $0 >$ fatalities ≤ 10; and "high" corresponds to fatalities > 10.[6] When aggregating multiple entries into a given dataset-latlon-day unit of observation, we first sum up all available fatality data for the unit, then categorize the unit by the sum. For each episode that is unpacked into single-day event entries, we distribute the fatality count uniformly across each lonlat-day within the window. While preventing over-counting deaths in a single episode window, this procedure also results in under-counting for a single latlon-day as severity is likely not equally distributed across days. This reality reflects the uncertainty in the episode itself and thus we opt for distributing activity uniformly as we have no additional information to reliably make a more specific assumption.

The result of the aggregation pre-processing step is a single observation that corresponds to a single event. When analyzing the spread, frequency, or severity of conflict, variability in the unit of analysis can be misleading and introduce bias. Aggregating to a common unit is both necessary and sufficient to generate consistency both across and within these data. More importantly, we contend that pre-processing is essential *whenever* employ-

[6]When fatality data is missing for an entry, the taxonomy codes the entry as "none". Our assumption is that entries without reported fatality data are more likely to be cases in which no fatalities occurred than cases resulting in either a low or high number of fatalities. Overall, fatality data is missing for only 0.002% (699) of entries.

ing the current generation of event data in any empirical analysis. That is to say, this is not just something that needs to happen pre-integration, but also pre-analysis when researchers are only using one dataset. Otherwise, if all entries are included as separate events, or if episodes are treated as one event, event count models may identify a greater number of events in the kind of areas where events are more likely to be disaggregated into multiple entries.

An Integrated Picture of Violent Conflict in Africa

We integrate ACLED, UCDP-GED, and GTD datasets for the subset of entries that refer to violent events occurring on the African continent and in Madagascar from 1997 to 2018.[7] The scope of the integration corresponds to overlapping spatiotemporal coverage across all three datasets. In this section, we outline our decisions in designing the integration, then explore the distribution of violent conflict activity in Africa when these datasets are integrated.

To integrate the datasets, we employ a spatiotemporal window of 25 kilometers and 2 days. The 25km window serves as a sufficient spatial buffer, in keeping with prior work by others who integrated conflict data manually (Weidmann, 2015). The 2-day temporal window mirrors the approach in Donnay et al. (2019). To reiterate, the spatial buffer only comes into play for precisely geo-coded events; otherwise, our implementation of algorithmic blocking relies strictly on the temporal buffer. As described in the previous section, the data

[7]Specifically, we discard all non-violent entries from ACLED, including the following event types: "Non-violent transfer of territory", "Headquarters or base established", "Protest" (note that "Protest/Riots" are split into protests and riots. We differentiate between the two types using the number of fatalities. When fatalities are greater than 0, we assume the event was a riot.), and "Non-violent activity by a conflict actor". All event types are retain from the UCDP-GED and GTD data.

is pre-processed so that (a) episodes are expanded[8], (b) spatial assignment is standardized given an imprecise geo-code, and (c) multiple entries are collapsed into a single entry (latlon-day).

Event type, actor, and severity metadata are incorporated as taxonomy information for running the MELTT procedure. These taxonomies are shallow: they are dichotomous and do not contain broader levels to be compared when matching. For the integration task to function with this approach, we allow for partial matches, meaning that not all criteria must align perfectly for to entries to be considered candidate matches. For example, if two candidate entries both report rebel actors and civilian actors and are coded as civilian violence, but differ on the level of severity, then these entries can still potentially match. Given not all taxonomy dimensions matter equally, we assign greater weights to the event type and some of the actor dimensions.[9]

Table 2 summarizes results of the integration, including the number of unique and matching entries between each dataset. We focus on two useful metrics when assessing the output of the integration: (*i*) the share of all entries in the pooled data that are flagged as matches and (*ii*) the percent reduction in the size of the pooled dataset after duplicate entries are consolidated. Of the 164,932 pooled entries, 30,463 entries are found to match (18.47%) for a spatial window of 25km and 2 days. Of those matching entries, 16,269 are found to be duplicates and therefore removed (9.86%).

[8]We employ a threshold of 7 days for episode, which means that episodes less than or equal to 7 days in duration are considered as part of the integration, whereas episodes of longer durations are discarded prior to integration.

[9]Specifically, we assign both event type taxonomy dimensions a weight of 0.2 apiece, while the government, violent, and civilian actor dimensions are assigned weights of 0.1 apiece. We emphasize these dimensions because they are most capable of discriminating among events. The remaining dimensions are equally weighted so that the total weight of all dimensions adds up to 1.

Table 2: Summary of the conflict data integration (25 km, 2 days)

ACLED	UCDP-GED	GTD	N Matches	N Entries
X			0	93,787
	X		0	27,996
		X	0	12,686
	X	X	866	1,732
X		X	3,907	7,814
X	X		7,346	14,692
X	X	X	2,075	6,225

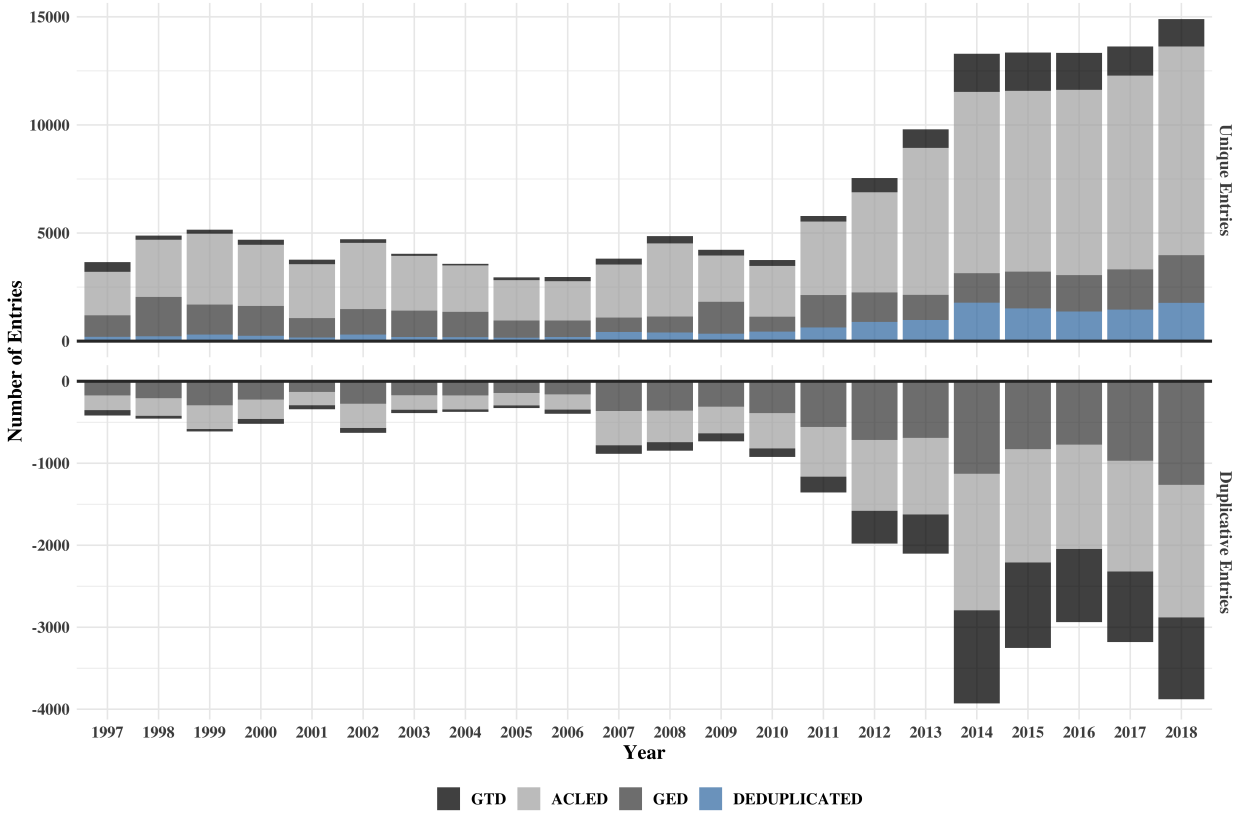
Total number of entries: 164,932
Total number of entries after de-duplication: 148,663
Number of duplicative entries removed: 16,269
Percent of entries that matched: 18.47%
Percent reduction in the size of the pooled dataset: 9.86%

Among the entries flagged as matches, the greatest overlap is between ACLED and UCDP-GED (48% of the total found matches), followed by all three datasets (20%), ACLED and GTD (25%), and UCDP-GED and GTD (6%). Among these datasets, ACLED is designed to cover the most expansive variety of conflict event types, including those involving violence. Thus, the substantial extent of matches between ACLED and both UCDP-GED and GTD is not surprising. Meanwhile, the low extent of the overlap between the latter two datasets is noteworthy in light of previous studies that have analyzed the relationship between organized armed conflict and terrorism Findley and Young (2012); Fortna (2015); Polo and Gleditsch (2016). Identifying duplicates between GTD and other datasets is relevant to ensuring that violent events are not double-counted across the respective measurements of activity, though not the specific focus of this integration.

At the same time, the integration indicates that ACLED overlooks some conflict activity. Our analysis identifies unique events present in UCDP-GED and GTD, even though their conceptualizations of conflict activity should be subsumed by what ACLED aims to

capture. Within the integrated dataset, 69% of entries are unique to ACLED, but 20% are organized armed conflict events unique to UCDP-GED, and 9% are terrorist events unique to GTD. The integration results indicate that if we rely strictly on ACLED to measure violent conflict across Africa from 1997-2018, this source could capture as much as 65% of relevant events (combining the unique entries and de-duplicated entries associated with ACLED), while overlooking 35% of events.[10]

Figure 2: Unique and Duplicate Entries by Year



This figure captures the number of unique entries (top panel) and the number of duplicate entries (bottom panel) by dataset-year. The y-axis for the duplicate entries is negative to denote that these entries are removed from the total. In the unique entry panel, all de-duplicated entries are emphasized in blue as "de-duplicated" to highlight that these entries do not belong to any one dataset in particular.

[10]Our calculation of these shares only takes into account the events that the three datasets record. Donnay et al. (2019) report results of an analysis using multiple systems estimation that suggests these datasets may collectively fail to capture a majority of actual conflict events.

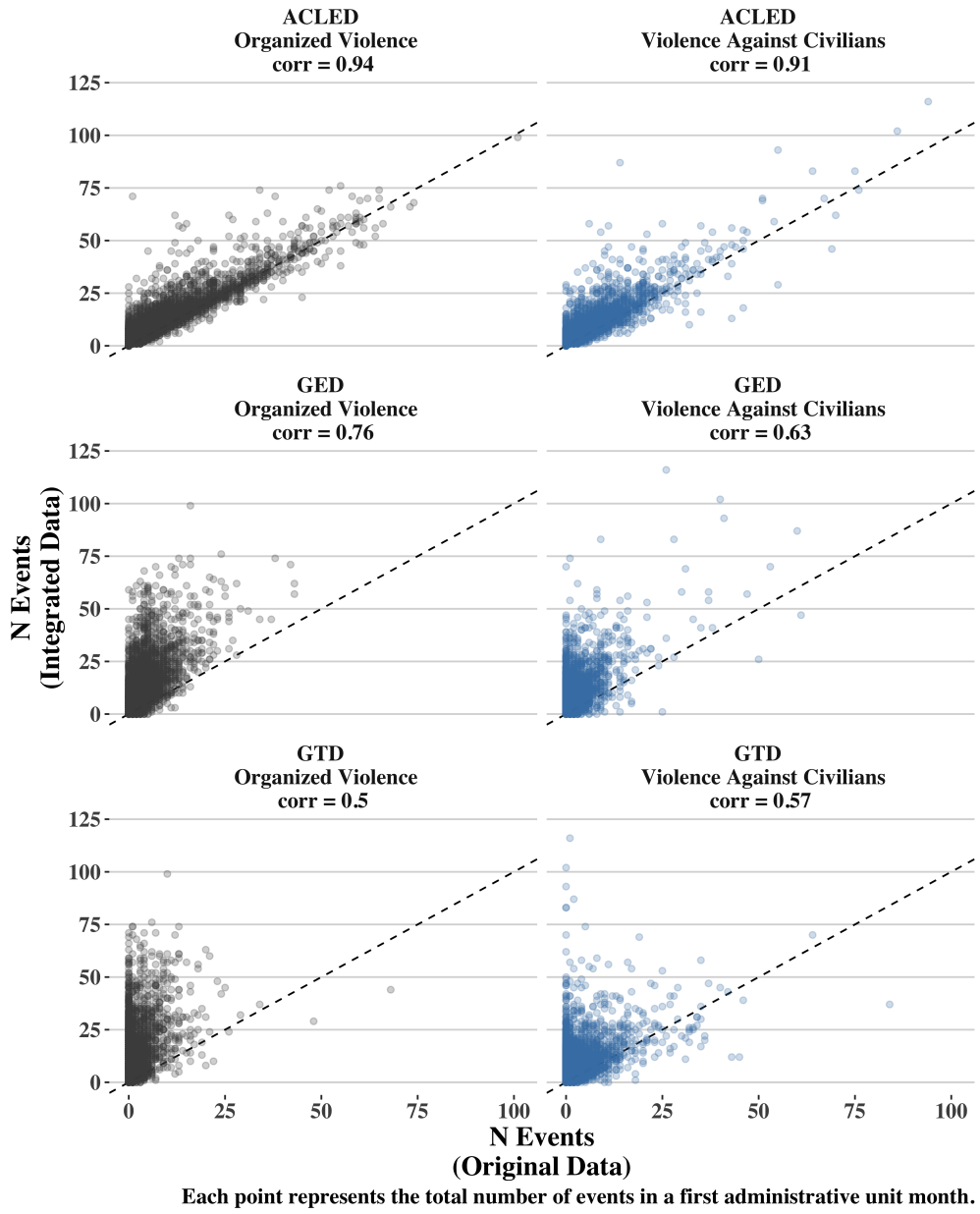
Figure 2 shows the numbers of unique and duplicate entries, by dataset, on an annual basis. We denote all matching events as "de-duplicated" to highlight that these entries do not belong to any one dataset, but are shared across multiple datasets. Unique entries increased substantially in all datasets from 2011-2014, before leveling off in 2015-2018. We attribute at least part of the increase to expanded media reporting, since the datasets rely to a large degree on coding news stories.[11]. Duplicate entries likewise grew substantially from 2011-2018. This trend potentially highlights a growing inter-reliance among the dataset initiatives. Each initiative may review the records of other sources with similar coverage to check for missingness, which can lead to post-hoc updates to the data.

We also explore the substantive differences between the integrated data and each conflict event dataset. Figure 3 tracks the number of reported events for each first-order administrative division-month (hereafter "admin-months") on the African continent and in Madagascar. The x-axis captures the event counts as reported in the original data; the y-axis reports the counts from the integrated data. Furthermore, we disaggregate the event counts with respect to two prominently utilized event types: organized violence and violence against civilians. The dashed line in the figure tracks with a perfect correlation—that is, a one-to-one mapping between the original and integrated data. Values that fall *above* the dashed line capture admin-months for which events are *under-reported* in the original data as compared to the integrated data. Values that fall *below* the dashed line capture admin-months for which events are *over-reported* in the original data.

Figure 3 demonstrates that our integrated dataset deviates from each of the indi-

[11]Studies have found that media reports often miss substantial conflict activity (Donnay and Filimonov, 2014; Weidmann, 2015)

Figure 3: Comparing the original and integrated versions



vidual datasets in meaningful ways. First, ACLED tracks most closely with the integrated data among the three datasets that we consider. This finding is not surprising since ACLED aims to capture violent activity known to be excluded from the other data projects. Specifically, ACLED captures "low-level" violence, especially within countries that have never experienced an organized armed conflict resulting in 25 deaths. This violence is perpetrated

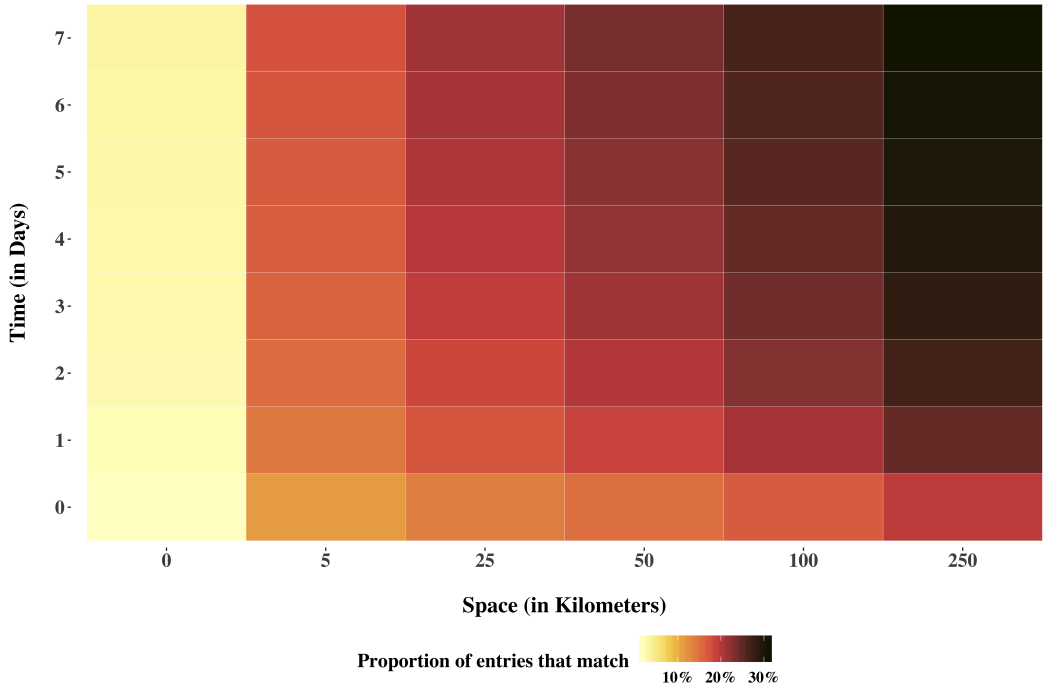
by a broad array of actors that may fall outside established lists of violent non-state actors (Cunningham et al., 2013). For projects interested in political violence and instability, broadly construed, tracking such activity is key. Second, even though ACLED correlates strongly with the integrated data, ACLED still deviates substantially from the dashed line that represents perfect correspondence. The implication is that ACLED either overcounts or undercounts substantially the number of violent events taking place within many admin-months. Third, the more focused data collection projects — UCDP-GED and GTD — exhibits a high degree of under-counting. These results are unsurprising, since these data projects do not aim to capture all violent activity. Figure 3 highlights the potential for selection bias if one aims to measure and analyze political violence using any one of these datasets.

The point of these analyses is not to single out evident incompleteness within any one dataset or to critique current data collection practices of these initiatives. Rather, we wish to emphasize the potential value of considering these datasets in concert when measuring and analyzing violent conflict. Our integration shows that each dataset exhibits the same phenomenon of capturing events that should be captured (given their coding criteria), but incompletely. The extent of completeness varies. If a researcher has a different aim than ours, the shares of relevant events captured by the datasets as illustrated in Figure 2 and Figure 3 could likewise differ. The important implication is that integration of multiple datasets can yield a more complete and accurate measure of violent conflict activity than relying on any one dataset.

The effect of varying the integration assumptions

Any integration is partially a function of the assumptions (i.e., pre-processing steps, spatiotemporal windows, taxonomy criteria, etc.). We reinforce this point in Figure 4, which examines the effects of variation in the spatiotemporal window, holding all other assumptions constant. We explore both larger and smaller windows.

Figure 4: Percent of matching entries given different spatiotemporal window configurations



Each bin reports the percent of events flagged as matches given specific spatial (x-axis) and temporal (y-axis) configurations for the integration window.

Figure 4 shows that as the window expands, the proportion of matches increases, which is an intuitive result. The relationship between the window and matches is primarily driven by the spatial dimension. When we require entries to match at the exact location, less than 10% of entries match, whereas the proportion of matching entries rises to about 30%

when we increase the spatial window to 250 km. Increasing the temporal window from 0 to 7 days has a much smaller effect on the proportion of matches. The fact that the proportion of matching entries does not exceed 32% even when setting a very wide spatiotemporal window (e.g., 250km, 7 days) means that a large number of events in each dataset provide new information not previously coded in any of the other data projects.

To be clear, there is no "correct" window, rather, to a large degree the appropriate window is a function of researchers' preference for minimizing false positives (unique entries that are incorrectly identified as matches) versus false negatives (duplicate entries that are incorrectly identified as unique entries). Larger windows likely increase the ratio of false positives, while smaller windows likely increase the ratio of false negatives. Donnay et al. (2019) suggest a manual validation procedure to assess the ratio of the false positive rate to the true positive rate for any integration output. An advantage of this procedure is a sense of the robustness of the integration output to different assumptions. The procedure has the drawback of being time intensive — and may be unrealistic to conduct if a researcher seeks to explore many varying integration assumptions as we do here.

At its core, however, integration is data pre-processing and should be thought of as a tuning parameter or source of sensitivity for any downstream analysis rather than a definitive (perfectly integrated) dataset. Thus, to best understand the implications of any one assumption (such as the decision on a spatiotemporal window), one should assess the effect any given assumption has on their final results and findings. One of the advantages of using an automated procedure is that re-running the integration under a different set of assumptions and then re-running a subsequent analysis is easy to do. We explore this point in the next section as we further explore insights gleaned from an integrated conflict event

dataset.

One interesting result from Figure 4 is that, even when setting a very large window (e.g. 7 days, 250km), the proportion of matching entries does not exceed 32%. This result means there are still a large number of events in each data set that are providing new information not previously coded in any of the other data projects. It is important to emphasize that integrated data is not always more appropriate than a specific data project if researchers' conceptualization of a conflict matches that of the specific data project. However, within published research, each of these datasets are frequently used to measure concepts such as "violence," "political violence," or "political instability" and each of these data projects seek to measure some or all instances of these concepts. We see the results in Figure 4 as further illustrating that integration, when done properly, can produce a better measurement of broad concepts such as these.

The Repercussions of Integration

We here examine the repercussions of integrating conflict event data for testing theoretical claims. A basic consideration is whether using integrated data affects the precision of statistical estimates. Specifically, do we arrive at different conclusions about observed relationships using integrated data as compared to using any one dataset independently? To explore this question, we design a statistical example using controls common to disaggregated studies of conflict. Specifically, we explore variables that aim to measure rugged terrain (mountain and forest coverage), distance (from border and capital), wealth (proxied by night lights) and population. As each of these measures are used as controls, our intent is not to

make a theoretical claim but rather to explore if the estimates are sensitive to differences in measurement.

We aggregate our event counts to a common disaggregated spatial unit: the PRIO-Grid (Tollefsen et al., 2012). The PRIO-Grid is an increasingly common unit of analysis in conflict studies as the grid construct contains valuable metadata regarding environmental and socioeconomic factors as relatively granular units. For each grid location, we calculate the raw event counts for the individual conflict datasets, as well as for our integrated data. Our decision to aggregate to the grid cell, without examining temporal variation, is reasonable for two reasons. First, most of the indicators we analyze do not vary across time (e.g., mountainous terrain). Thus, the temporal dimension carries no information. Second, those of the indicators that do vary over time tend to exhibit a large number of missing values. Our concern is that these missing values may be correlated with the likelihood of any one of the conflict datasets capturing (or failing to capture) a violent event.

When generating the event counts in each PRIO-grid unit, we only consider entries that are *precisely* geo-coded. Event entries whose locations are imprecisely coded to administrative centroids could artificially inflate the event count in grid units that contain those centroids, potentially biasing measurement. Thus, we only explore events that are sufficiently disaggregated. Note that by dropping all entries that are imprecisely geo-coded, we risk introducing selection bias into the analysis. We see this as the lesser of two evils but note that a more robust strategy for dealing with imprecision in these data should be considered. We would again point out that this is not a problem specific to integrated data—any researcher using event data in an analysis at a subnational level has to decide how to deal with events that are artificially placed in a location (such as a capital or centroid) due to

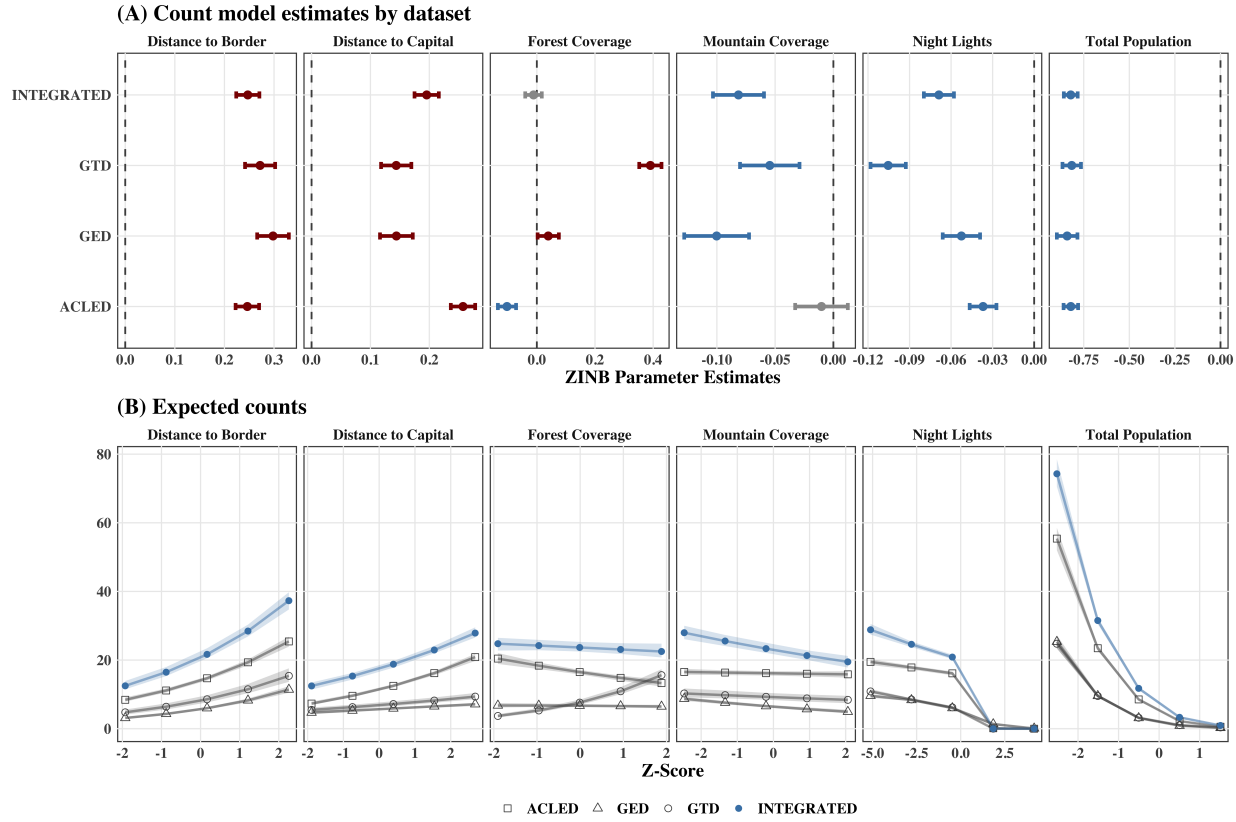
spatial imprecision.

Even when collapsing the temporal dimension of the data, the event counts indicate that conflict does not occur in a substantial number of grid units. Consequently, we conduct the analysis using the zero-inflated negative binomial model (hereafter "ZINB"), a standard version of a count model that deals with an over-proliferation of zeros alongside an over-dispersion in the count. As applied, the model aims to estimate the data-generating processes of both the onset and severity of violent conflict. In presenting our results, we focus on the count output; the full results of the estimations are reported in Appendix A.4.

Figure 5 reports the parameter estimates from the count portion of the ZINB (panel A), as well as the expected counts when the respective variable is varied two standard deviations above and below the mean value (panel B). We standardize all six independent variables (also, we log the distance and population measures due to a large right skew). Thus, the estimates can be read as a unit increase in the dependent variable given a standard deviation increase in the independent variable.

Figure 5 shows meaningful differences in the parameter estimates given which conflict data are employed. The forest coverage variable exhibits the biggest differences in results: analysis conducted on ACLED shows a negative and significant effect, whereas analyses with UCDP-GED and GTD show a positive and significant effect. In contrast, the estimated relationship with the integrated data is not statistically significant—yielding a notable middle ground between the divergent results from the analyses using the other datasets. The differences in results for the mountain coverage variable are less profound; this variable has a negative sign in the analysis with each dataset and the integrated data, but the relationship is insignificant with ACLED. For the other four independent variables, the same sign and

Figure 5: Estimates of the count of violent conflict events as a function of standard independent variables



Panel A reports the parameter estimates (along with 95% confidence intervals) for the count portion of a zero-inflated negative binomial model (ZINB). The dashed line denotes the location of 0 and the colors highlight a positive (red), negative (blue), or statistically insignificant (grey) effect. The x-axis captures the parameter estimates, while the y-axis reports the violence data used as the dependent variable to generate the estimate. **Panel B** shows the expected count from the count portion of the ZINB model as each variable is varied by two standard deviations above and below the variable mean. All variables are standardized to have a mean of 0 and a variance of 1. Integrated data is differentiated in blue with a solid point. All other variables are held at their observed values.

significance is observed across the individual and integrated datasets. The nighttime lights and distance to capital variables do exhibit significant differences in the estimated effect of a one-unit change depending on which conflict data are used.[12]

[12]In Appendix A.4.1, we explore the sensitivity of the results to the spatiotemporal window used for integration. The parameter estimates remain largely stable even as the window is increased to 250 kilometers and 7 days. This implies that the result presented here are not sensitive to true/false positives resulting from the integration.

The expected counts in Panel B of Figure 5 show the differences in the predicted number of events as the values of the variables increase. Again, we observe divergent effects of the Forest Coverage and Mountain Coverage variables. Panel B also shows that, for each of the variables, the integrated data yields a larger expected count, which highlights a higher base rate in the integrated data. The implication is that when predicting conflict counts, integrated data will generally yield higher expected counts while tracking with the trends captured in the other datasets. This increase in the base rate is not surprising given that integrated data fundamentally leads to the inclusion of more event activity.

As stated above, we are not seeking to test theoretical claims in the analyses in Figure 5. We see these analyses as informative, however, in that they reveal that the choice of dataset can affect both the sign and magnitude of the effect of variables on the number of events observed. In some ways, this should not be surprising, given that the datasets have somewhat different goals—UCDP-GED is trying to capture events related to its conceptualization of armed conflict, GTD is focused specifically on terrorism, and ACLED much broader. The differences between ACLED and the other two datasets in the Forest Coverage variable, for example, could suggest that terrorism and organized armed violence are more likely in areas of significant forest cover, but that the kind of low-level, unorganized violence picked up in ACLED is more likely in less forested territory.

The divergence between the integrated data and the ACLED data for both terrain variables, however, suggests that the patterns are not just driven by different conceptualizations of conflict. Our empirical explorations above suggest that there is a significant degree of missingness across all of the data projects. If integration produces a more accurate measure of conflict activity—with less events missing—then the analyses in Figure 5 suggest that

using a more accurate measure can lead to different interpretations of the effect of variables affecting the occurrence and location of violent conflict.

Conclusion

The proliferation of conflict event datasets is a boon to research on the disaggregated dynamics of conflict. Yet these resources have not been used to their full potential, since each study typically draws on one of the datasets, even when another available source is relevant. Scholars show an interest in integrating multiple datasets (e.g., Findley and Young (2012); Fortna (2015); Weidmann (2015); Polo and Gleditsch (2016)). Those efforts are limited in scope, which we attribute to challenges inherent to integration. Integration requires both detailed knowledge of the datasets being used and an efficient process to generate an integrated dataset that is transparent, replicable, and reliable.

The integrated dataset we present here meets these requirements. We pre-process the data contained in ACLED, UCDP-GED, and GTD so that the observations are comparable with respect to the type of events that are included and the translation of locations into geospatial coordinates. We then use the MELTT protocol to integrate the data (Donnay et al., 2019). This automated protocol allows us to conduct a large number of integrations (Figure 4 presents results reflecting 192 separate integrations) efficiently and to explore systematically how the extent of overlap across the datasets varies across these integrations (as well as the impact of different integration assumptions on the results of subsequent analyses).

Each of these integrations shows overlap across these data projects in the events that

are captured, as well as events that are unique to individual datasets. Our results confirm well-known patterns among the datasets, as well as important nuances that depart from those patterns, which are important to measurement and analysis. In particular, ACLED reports more events than either UCDP-GED or GTD, and greater proportions of events match between ACLED and both of the other datasets than between UCDP-GED and GTD. Yet the degree of matching across the datasets varies among administrative units. Thus, each dataset appears to capture a greater proportion of violent conflict activity in certain locations then in others. For this reason and others, our results make the case that the integrated data provide a more complete and accurate measure of violent conflict activity than any one of the datasets alone. The illustrative statistical analysis we conduct shows that using this measure can lead to different conclusions about relationships between commonly used independent variables and the occurrence of violent conflict activity in specific locations.

The integrated dataset that we present here, which integrates all entries related to violent conflict from ACLED, UCDP-GED, and GTD for the African continent and Madagascar from 1997-2018, represents the largest integrated dataset to date, by orders of magnitude. We make the dataset itself and all associated inputs, including the pipeline to pre-process the data, publicly available. Researchers can use the new dataset directly to conduct empirical analysis of their own. They can also make changes to the inputs, to reflect pertinent information or to conduct robustness checks, then re-run the integration.

We see several paths forward for future research. First, the scope of integration can be expanded. We restrict the scope to Africa from 1997-2018 as a reflection of the current overlap among the three datasets. Both UCDP-GED and GTD already have global coverage, while ACLED is periodically adding coverage of additional countries and regions, opening

up opportunities of data integration covering an even richer array of conflict activity beyond Africa. Complementing those prospects are newer data initiatives, which offer valuable coverage of specific types of conflict.

Second, integration is revealing in ways that can help to upgrade data collection processes. Datasets should ideally capture what they intend, as completely and accurately as possible. Integration sheds fresh light on conceptualizations of conflict and the efficacy of operationalizations. Firmer boundaries can be established where warranted. Greater alignment of variable schema across datasets would foster comparison. Integration also documents deficiencies of datasets in capturing events, which may isolate biases to correct. In addition, the partial overlap established via integration is a strong signal that the data initiatives collectively understate conflict activity. More efforts ought to be devoted to bolstering assessments of the cumulative burden of conflict, bringing together integration with ground-truth validation.

Third, generation of integrated data can spur re-examination of a range of topics in conflict research. Studies regularly rely on a single source of event data that may be incomplete and not even suitably tailored to the topic. Using integrated data, researchers can investigate the results obtained with measures that do a better job of capturing the conflict activity of interest.

Fourth, data integration can be a catalyst for innovations at the frontiers of conflict research. One line of inquiry where integration should be integral is studying relationships among types of conflict. Traditionally, different types are studied on separate tracks. These divisions have been dissolving. Notable topics of growing interest in the field include why actors choose certain conflict tactics from a menu of options, how various manifestations of

conflict intersect, interact and influence each other, and when conflict can be expected to morph in type. Addressing the questions is likely to require working with multiple sources of data affording coverage of distinct types of conflict. Integrating these datasets can ensure appropriate refinement in the measurement of conflict activity.

References

Becker, R. A., A. R. Wilks, R. Brownrigg, T. P. Minka, and A. Deckmyn (2016). maps: Draw geographical maps. r package version 3.1. 0.

Birnir, J. K., D. D. Laitin, J. Wilkenfeld, D. M. Waguespack, A. S. Hultquist, and T. R. Gurr (2018). Introducing the amar (all minorities at risk) data. *Journal of Conflict Resolution 62*(1), 203–226.

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

Cunningham, D. E., K. S. Gleditsch, and I. Salehyan (2013). Non-state actors in civil wars: A new dataset. *Conflict Management and Peace Science 30*(5), 516–531.

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

Donnay, K. and V. Filimonov (2014). Views to a war: systematic differences in media and military reporting of the war in iraq. *EPJ Data Science 3*(1), 25.

Findley, M. G. and J. K. Young (2012). Terrorism and civil war: A spatial and temporal approach to a conceptual problem. *Perspectives on Politics 10*(2), 285–305.

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

National Consortium for the Study of Terrorism and Responses to Terrorism (START) (2013). Global terrorism database. Available from http://www.start.umd.edu/gtd.

Polo, S. M. and K. S. Gleditsch (2016). Twisting arms and sending messages: Terrorist tactics in civil war. *Journal of Peace Research 53*(6), 815–829.

Raleigh, C., A. Linke, H. Hegre, and J. Karlsen (2010). Introducing ACLED–Armed Conflict Location and Event Data. *Journal of Peace Research 47*(5), 651–660.

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

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

Tollefsen, A. F., H. Strand, and H. Buhaug (2012). Prio-grid: A unified spatial data structure. *Journal of Peace Research 49*(2), 363–374.

Weidmann, N. B. (2015). On the accuracy of media-based conflict event data. *Journal of Conflict Resolution 59*(6), 1129–1149.

Zhukov, Y. M., C. Davenport, and N. Kostyuk (2019). Introducing xsub: a new portal for cross-national data on subnational violence. *Journal of peace research 56*(4), 604–614.

A Supplementary Information

A.1 Information availability and event disaggregation model

In Figure 1 in the main paper, we show that event disaggregation is a function of information availability. The presence of a capital city in an administrative unit increases the probability that events recorded in that unit will be disaggregated into multiple entries. This insight motivates event aggregation to the latlon-day. Table 3 contains the entire model used to estimate the predicted probabilities in Figure 1.

We define an event as containing multiple entries if more than one entry exists for the same latlon-day-dataset. The data is filtered to only consider precisely geo-coded event entries. By definition, multiple entries exist for entries aggregated to a centroid location: multiple events at the same latlon-day in these instances is a function of data aggregation rather than event disaggregation. Data on city and capital locations is draw from local data stored in the R package `maps` (Becker et al., 2016), which contains data on the coordinate locations of all world cities (circa 2006).

Table 3: Presence of a captial cities on the likelihood of event disaggregation

	Event Disaggregation
Capital City	0.356^{***}
	(0.016)
ACLED	0.155^{***}
	(0.019)
GTD	0.398^{***}
	(0.023)
Number of Cities	0.005^{***}
	(0.001)
Observations	106,367
Log Likelihood	−26,770.790
Akaike Inf. Crit.	53,715.590
Note:	*p<0.1; **p<0.05; ***p<0.01
	Unit of analysis: latlon-day.
	Baseline: UCDP-GED for dataset fixed effects.
	Country, year, and month fixed effects included.

Table 3 present the results of a probit model with country and temporal (year and month) fixed effects. Finally, fixed effects are added for each of the three datasets (ACLED, UCDP-GED, and GTD), where UCDP-GED is the baseline dataset. The results show that

both ACLED and GTD are more likely to disaggregate events in comparison to UCDP-GED. If an event was located in a first administrative unit containing a capital city, we find a substantively and statistically significant effect in the likelihood of observing a disaggregated entry. Finally, as the number of cities in the first administrative unit where the event was recorded increases, so too does the likelihood of event disaggregation increase.

A.2 Event Taxonomy Scheme

In the article, we outline pre-processing steps necessary for integrating event datasets with different spatial and temporal imprecision rules. In addition, we outline an aggregation strategy for events that are disaggregated into multiple entries. Given that event disaggregation is attributed to the existence of more information about an event, we advance a way of aggregating event information into a single observation that minimizes loss of the added entry perspectives.

Specifically, we adopt a shallow taxonomy scheme that encodes an event profile allowing for additional information to be encoded into an observation. A single observation commonly holds a field for event type, severity, and actors (perpetrator/target). As encoded, only one state can occupy said field. To allow for aggregation, we generate a series of dummy variables that encode all the information about the information for a specific latlon-day. Table 4 reports the encoded indicator fields.

Table 4: Standardized event information as shallow taxonomies

Indicator	Taxonomy Type	Coding Process
civilian_actor	actor	actor dictionary
ethnic_actor	actor	actor dictionary + AMAR (Birnir et al., 2018)
religious_actor	actor	actor dictionary
government_actor	actor	actor dictionary
violent_actor	actor	actor dictionary + GTD actor descriptors
violence	event type	relevant event types
civilian_violence	event type	relevant event types + civilian_actor target
fatal_none	severity	fatalities $= 0$
fatal_low	severity	$0 <$ fatalities ≤ 10
fatal_high	severity	fatalities ≥ 10

Table 4 summarizes the different indicator fields we generate for the aggregation and integration task, along with the coding process used to generate the field. We outline how actors were classified in greater detail below. For the event taxonomy, we leverage the event type field, which is consistent in each of the datasets. The only exception being when we code for civilian violence in the GTD dataset. GTD does not have an explicit violence against civilians event type, even though the majority of the activity it reports (terrorism) targets civilians. We code an event as civilian violence when the civilian actor field is triggered for a GTD entry. Finally, the fatality categories are generated using the fatalities field, which exists in each of the input datasets.

A.2.1 Classifying actors

To facilitate a way of generalizing across different actors contained within the input datasets, we generate general categories (or bins) that actors plausibly fall into. We generalize across different actor types by focusing primarily on government, rebel, and civilian actors. In addition, we flag instances when religious or ethnic information is contained within the actor name. Given that both side A and side B actor—which can be loosely understood as the perpetrator and target, though that convention does not hold across datasets— fields are used as metadata in the integration process, we must leverage actor lists across a number of data sources and build our own dictionary of keywords (textual flags that are unique to specific actor type) to classify actors.

Of the three datasets under consideration, only one (UCDP-GED) has an explicit actor list. The other three datasets all capture a larger range of activity, and for this reason, these data contain a more diverse array of actors. To identify and categorize these actors, we developed a general dictionary to help facilitate this process.

However, any actor taxonomy is not robust enough to deal with ambiguity in how the different datasets arrange dyadic actors in an exchange. For example, in any situation where there is a dyadic pair side_A to side_B, actor A in dataset X may correspond with actors coded as side_A or side_B in dataset Z or not. This capacity for the initiator to be flipped with the target and vice versa is plausible when integrating data that do not hold persistent rules regarding how actors are designated in dyadic arrangements.

We build off the actor taxonomy advanced by Donnay et al. (2019) by reconceptualizing the way actor entries are compared. Specifically, we build out indicator variables as a series of "shallow" taxonomy dimensions, one for each potential actor type. When entries are compared they either agree on these dimensions or not. We have found that this is the most general way to build flexibility with regard to ambiguity in the placement of the dyadic pairings without generating a penalty. As such, actor arrangements can be compared even if initiator-target orderings are ambiguous.

For rebel actors, we hone in on some common keywords, such as "rebel", "gunman", "armed", to flag armed actors. In addition, given that side_A is always reliably an armed perpetrator in GTD by design, we leverage unique adjectives and nouns in that list to bolster our list or keywords. For civilian actors, we manually reviewed all actors in the data and recorded common nouns and adjectives that are associated with civilian actors, e.g. "workers", "students", "refugees", etc. For ethnic actors (or actors that are identified using an ethnic moniker, e.g. the "Kurds"), we leverage actors identified in the AMAR dataset, which tracks all socially relevant ethnic groups (Birnir et al., 2018). In addition to this, we incorporate key words alluding to ethnicity (such as "kinsmen","clan","tribal"). Finally, for religious actors, we target keywords such as "sect", "cult", "muslim", "christian", etc.

A complete list of keywords in the actor taxonomy can be found with the replication files accompanying this project.

A.3 Integration Assumptions

In this section, we outline in greater detail specific steps taken in the integration process that differ from Donnay et al. (2019). The main aim of the authors' initial paper was to

increase transparency in how data was combined. Sticking with that tradition, we outline all the steps that we took in integrating the data. Most of these steps are outlined in detail in either the article (pre-processing, spatiotemporal windows) or in previous Appendix sections (see Appendix A.2); we merely itemize those instances here. However, we explain in greater detail any technical decisions that differ fundamentally from the integration of Nigeria 2011 data in Donnay et al. (2019).

The following integration steps all deviate from Donnay et al. (2019):

- Event data pre-processing
  - Episode expansion (only episodes with durations of 7 days or less are retained).
  - Convert entry metadata to shallow taxonomy scheme
  - Aggregate to the latlon-day
- Integrate using the MELTT algorithm
  - Allow for partial matches
  - Impose weight scheme on the taxonomy dimensions
    - upweight event type information
    - upweight consistent types of actor information (government, rebel, and civilian actor dimensions)
  - Allow for "across the board" (3-way) matches - See Appendix A.3.1.

A.3.1 Across the Board (3-way) matches

In Donnay et al. (2019), datasets are matched iteratively. If more than two datasets are being matched, MELTT first matches dataset 1 and 2, then uses the information from dataset 1 when matching any subsequent dataset. This scheme preferences the "leading dataset" by retaining its coordinate information when locating future matches.

For the integration in the paper, we deviate from this logic slightly. Rather than relying on a single dataset as the primary source of coordinate information, we allow each dataset the opportunity to pair with every other dataset, and then scan for what we refer to here as "across the board" (ATB) matches. We outline four potential pathways for an ATB match. Note that below the numbers denotes datasets, and the letters denote specific entries in those datasets.

- **Pathway 1**: match(1a, 2b), match(1a, 3c), match(2b,3c) → 1a-2b-3c
- **Pathway 2**: match(1a, 2b), match(1a, 3c) → 1a-2b-3c (Donnay et al. (2019)'s original logic)
- **Pathway 3**: match(1a, 2b), match(2b, 3c) → 1a-2b-3c (2b bridges)
- **Pathway 4**: match(1a, 3c), match(2b ,3c) → 1a-2b-3c (3c bridges)

All four match pathways are possible when integrating data. As a result, more than one ATB match can be identified for a specific set of events when pairing more than two datasets. For example, 1a can match to 2b, 3c via pathway 2, but then 1a can match to 2b and 2b to 3d via pathway 4.

For pathways 3 and 4, dataset 3 and dataset 2 operate as "bridging events" tying 1 to 2 and 1 to 3, respectively. These are what Donnay et al. (2019) call chaining events. Donnay et al. (2019) worried that such a matching scheme could lead to chain matching where largely disparate events were found to match via some intermediate observation from another dataset. Chaining presented a real challenge to the logic of MELTT, especially if that intermediate event linked events distant in time and/or space. However, we find that the original logic of meltt is overly-conservative in practice and often yields an integrated dataset that resembles the leading dataset: as only information in the leading dataset was used to find matches in any of the other datasets.

When multiple pathways emerge, a choice needs to be made on which pathway to accept. By only choosing Pathway 2 (as Donnay et al. (2019) do), the assumption is that the leading dataset information is the determinate factor when determining future matches. When strictly held, the assumption prevents matches along pathway 3 & 4 (and overlooks the additional evidence for an ATB match flagged by pathway 1).

We outline a preference among specific pathways when reconciling multiple pathway matches. We contend that pathway 1 is the most robust of potential pathways. In this arrangement, each dataset entry independently aligns to every other dataset entry. This is ideal as there exists maximal agreement, but again when strictly held, prevents the occurrence of bridging events (pathways 3 & 4).

We implement a procedure that scans all ATB pairings. When multiple pathways are detected for any ATB pairing, the procedure scans for a pathway 1 arrangement. If none exists, it preferences a pathway 2 arrangement. If none exists, it takes the remaining pathway. Note that only a pathway 3 or 4 can exist at this point as the existence of both would imply a pathway 1 arrangement exists.

Note that all ATB pairings that are dropped are still considered down the line as non-ATB pairings — i.e. matches that only occur between two datasets, which constitutes the majority of matches located by the integration procedure (see Table 2).

A.4 Empirical Results

Table 5 reports the full zero-inflated negative binomial model as presented — in part — in Figure 5.

A.4.1 Sensitivity Analysis

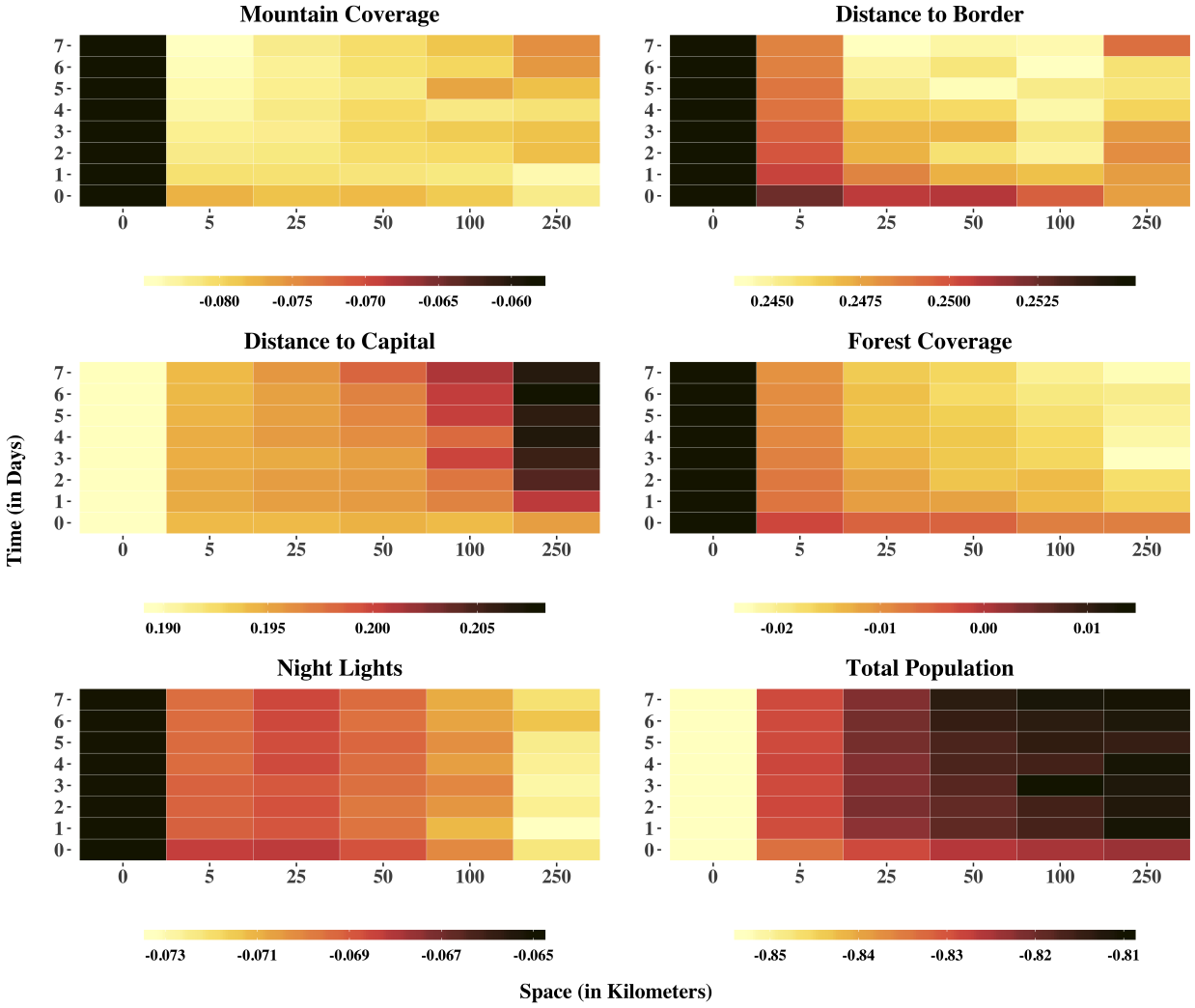
We explore the sensitivity of the parameter estimates presented in Table 5 to specific integration assumptions. Specifically, we examine the differences in parameter estimates as the spatiotemporal window is varied along the bins established in Figure 4. To do this, we re-estimate the same model as the one reported in Table 5 using different integrations that were generated with different window sizes.

Table 5: Full model output ZINB model

Dataset	Model Component	Variable	Estimate (Std. Error)
integrated	Count	Constant	1.94 (0.02)
integrated	Count	Mountain Coverage	-0.08 (0.01)
integrated	Count	Distance to Border	0.25 (0.01)
integrated	Count	Distance to Capital	0.2 (0.01)
integrated	Count	Forest Coverage	-0.01 (0.01)
integrated	Count	Night Lights	-0.07 (0.01)
integrated	Count	Total Population	-0.82 (0.02)
integrated	Count	Log Theta	-1.03 (0.01)
integrated	Inflation	Constant	-1.2 (0.11)
integrated	Inflation	Mountain Coverage	-0.04 (0.03)
integrated	Inflation	Distance to Border	-0.23 (0.03)
integrated	Inflation	Distance to Capital	0.33 (0.03)
integrated	Inflation	Night Lights	6.5 (0.58)
integrated	Inflation	Total Population	1.96 (0.05)
integrated	Inflation	Forest Coverage	0.25 (0.03)
acled	Count	Constant	1.67 (0.02)
acled	Count	Mountain Coverage	-0.01 (0.01)
acled	Count	Distance to Border	0.25 (0.01)
acled	Count	Distance to Capital	0.26 (0.01)
acled	Count	Forest Coverage	-0.1 (0.02)
acled	Count	Night Lights	-0.04 (0)
acled	Count	Total Population	-0.82 (0.02)
acled	Count	Log Theta	-1.08 (0.01)
acled	Inflation	Constant	-0.8 (0.09)
acled	Inflation	Mountain Coverage	0.02 (0.03)
acled	Inflation	Distance to Border	-0.22 (0.03)
acled	Inflation	Distance to Capital	0.33 (0.03)
acled	Inflation	Night Lights	4.78 (0.48)
acled	Inflation	Total Population	1.99 (0.05)
acled	Inflation	Forest Coverage	0.18 (0.03)
ged	Count	Constant	0.9 (0.03)
ged	Count	Mountain Coverage	-0.1 (0.01)
ged	Count	Distance to Border	0.3 (0.02)
ged	Count	Distance to Capital	0.14 (0.01)
ged	Count	Forest Coverage	0.04 (0.02)
ged	Count	Night Lights	-0.05 (0.01)
ged	Count	Total Population	-0.84 (0.03)
ged	Count	Log Theta	-1.23 (0.02)
ged	Inflation	Constant	0.27 (0.05)
ged	Inflation	Mountain Coverage	0.15 (0.03)
ged	Inflation	Distance to Border	-0.06 (0.02)
ged	Inflation	Distance to Capital	0.34 (0.03)
ged	Inflation	Night Lights	1.34 (0.13)
ged	Inflation	Total Population	1.17 (0.04)
ged	Inflation	Forest Coverage	0.35 (0.03)
gtd	Count	Constant	0.77 (0.03)
gtd	Count	Mountain Coverage	-0.05 (0.01)
gtd	Count	Distance to Border	0.27 (0.02)
gtd	Count	Distance to Capital	0.14 (0.01)
gtd	Count	Forest Coverage	0.39 (0.02)
gtd	Count	Night Lights	-0.11 (0.01)
gtd	Count	Total Population	-0.82 (0.03)
gtd	Count	Log Theta	-1.01 (0.02)
gtd	Inflation	Constant	0.41 (0.06)
gtd	Inflation	Mountain Coverage	-0.14 (0.02)
gtd	Inflation	Distance to Border	-0.1 (0.02)
gtd	Inflation	Distance to Capital	0.31 (0.03)
gtd	Inflation	Night Lights	4.16 (0.29)
gtd	Inflation	Total Population	1.39 (0.04)
gtd	Inflation	Forest Coverage	0.15 (0.03)

Model results are reported for each respective conflict event dataset.

Figure 6: Parameter sensitivity varying the spatiotemporal integration window



The figure explores the parameter estimates for each variable explored in Figure 5 given different spatiotemporal integration windows. The bars below each tile plot report the variability among the estimates for each variable. All variables remain statistically significant across all windows, except Forest Coverage, which never becomes statistically significant for any window size.

Figure 6 reports the parameter estimate for each variable as the spatial and temporal bin is varied. The figure shows that parameter estimates are somewhat sensitive to the integration assumption imposed: most notably when a spatial window is set at 0 then increased to 5 kilometers. The large increase in the proportion of matches reported in Figure 4 appears to correspond to the shift in estimates. It is important to note that no one estimate varies substantially. The Forest Coverage estimate shifts from negative to positive but never becomes statistically significant. However, the figure demonstrates how — like all pre-processing methods — the sensitivity to model output given specific integration assumptions should be explored.