Online Appendix for "Integrating Conflict Event Data"

Journal of Conflict Resolution

Karsten Donnay, Eric T. Dunford, Erin C. McGrath,
David Backer, and David E. Cunningham

This document provides supplementary information and analysis referred to in the article, as well as further details for analyses that, in the interest of space, could only be discussed briefly. Section A of this document introduces the empirical data used in our study and elaborates on the replication of our findings. The full replication material can be downloaded from the journal website. The open-source *R* software tool `meltt` we developed for our analyses is available at `https://cran.r-project.org/package=meltt`. Section B provides a short user guide for the MELTT protocol, including additional details on its functionality referred to in the article. Section C explains the construction of the synthetic data and stylized scenarios used to test the performance of MELTT, as well as supplies additional details on the test results. Section D goes into greater depth on our empirical analyses. We document the process of the integration of conflict events for Nigeria 2011 and the various sensitivity checks and manual validation we performed. Finally, we describe the selection of the two further cases we analyzed (South Sudan 2015 and Libya 2014) and discuss the procedure we used to validate the performance of MELTT in these cases.

Contents

A Data and Replication

A.1 Empirical Data

The empirical analyses presented in the manuscript rely on four well-established conflict event datasets: 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), the Global Terrorism Database (GTD) (START 2013), and the Social Conflict Analysis Database (SCAD) (Salehyan et al. 2012). The analysis of Nigeria 2011 relies on ACLED Version 5 All Africa 1997-2014, UCDP-GED Version 3.0, the June 2015 release of GTD and SCAD Version 3.1. For the subsequent analysis of Libya and South Sudan we used more recent versions of each dataset that also include the year 2015; specifically, we relied on ACLED Version 7 All Africa 1997-2016, UCDP-GED Version 17.1, the 2017 release of GTD and SCAD Version 3.2.

An important consideration is that these datasets are continuously updated, including retrospective correction of entries. Using different versions of the datasets may therefore yield different integration results, given the changing empirical basis. For this reason, a basic premise is that MELTT does not produce a final, definitive integration. Instead, integrated data can be generated for any pertinent versions of datasets. In order to replicate the figures presented in this article, the exact data used for each analysis are provided as part of the replication material (see below).

A.2 Replication Material

The replication material for this article is available from the journal website `http://journals.sagepub.com/home/jcrb` alongside the main article and this online appendix. We provide two separate replication scripts in *R*. The first script (`replication-simulation-tests.R`) performs the full Monte Carlo analyses summarized in Table 1 of the article. The second script (`replication-conflict-event-data.R`) replicates the empirical data integration for Nigeria (Figure 3, Table 2 and 3), as well as for Libya and South Sudan. We also show how the `meltt` software tool we developed can be used to enable manual validation using human coders, in line with the process outlined in the article (Table 4).

Empirical data and the taxonomies used for each case are provided as three .Rdata files (`Nigeria_2011.Rdata`, `Libya_2014.Rdata`, `SSudan_2015.Rdata`). The simulated data used for the performance tests of MELTT reported in Table 1 is generated on-the-fly as part of the (`replication-simulation-tests.R`) script. For more details on the scenarios considered, please refer to Section C. Replication material for the supplementary analyses presented in this appendix is available upon request.

B MELTT Protocol

We first walk step-by-step through the MELTT data integration protocol introduced in the article. Next, we focus on three additional aspects of the procedure in greater detail: the blocking strategy, iteration and optional configurations. We also discuss the extension to multiple datasets and the generalization to episodal data.

B.1 User Guide

In this guide, we refer back to the example of integrating ACLED, UCDP-GED, GTD and SCAD data for Nigeria 2011 discussed in the article. Each of the four datasets we consider contains columns that code:

- `date`: when the event occurred;
- `enddate`: if the event transpired across more than one day, i.e., an "episode";
- `longitude` & `latitude`: geo-location information on where the event occurred;
- `event_tax`: coding scheme of the type of event;
- `actor_tax`: coding scheme of the type of actor;
- `precision_tax`: coding scheme of the geo-precision of event locations.

For details on the three taxonomy variables used, see Section D. Note that for running the data integration protocol, the variable names of these "shared" variables across dataset have to be standardized, such that MELTT correctly recognizes which columns to use for event comparisons.

Getting Started

The goal is to evaluate these four event datasets in order to identify which of the reported events are the same and which are unique. The protocol formalizes all input assumptions the user needs to make to conduct this evaluation.

First, the researcher has to specify a spatial and temporal window within which any potential match could plausibly fall. Put differently, how proximate in space and time does a recorded event need to be to qualify as potentially the same?

Second, the researcher needs to input event taxonomies. The idea is that each dataset can record information differently, and a taxonomy is a formalization of how coding of specific attributes corresponds, moving from as granular as possible to as general as possible. Here, we describe how the coding of three specific event attributes we considered (type, actor, geo-precision) corresponds across our four datasets.

Section D.2 explains how we arrive at these exact mappings and what these taxonomies look like for the case of Nigeria 2011. The event and precision taxonomies contain more levels than the actor taxonomy. For example, the `event_tax` goes from 'level1', which contains a schema with 10 unique entries using specific, narrow categories of conflict events to 'level3', which just differentiates between violent and nonviolent events.

Generally, specifications of taxonomy levels can be as granular or as broad as one chooses. The more fine-grained the levels one includes, the more specific the potential match. Yet the use of fine-grained categories can also fail to identify potential matches across datasets, since the categories may be distinct to individual datasets. As a rule, therefore, a trade-off exists between specific categories that can better discriminate among possible duplicate entries and broader categories that are more capable of identifying candidates for matches across datasets.

As a general rule, we recommend to include, whenever it is conceptually warranted, both specific fine-grained categories and a limited number of increasingly broader categories. In this case, MELTT will have more information to work with when identifying and discriminating potential matches. When establishing which entries are most likely to be the same, in case of more than two potential matches in one dataset, the protocol automatically favors the match that more precisely corresponds. *A good taxonomy is the key to matching data, and is the primary vehicle by which a researcher's assumptions—regarding how data fits together—are made transparent.*

Matching Data

Once the user has formalized any taxonomies, integrating several datasets is straightforward. In our main analysis in the article, we assume that any two entries in different datasets that indicate events occurring within 3 kilometers and 1 day of each other could plausibly capture the same event. This "fuzziness" sets the boundaries on how precise we believe the location and timing of events are coded. Varying these specifications systematically is usually best practice, to ensure that no specification drives the outcomes of the integration task. We perform this check in Section D.3.2. We then assume that event categories map onto each other according to the way that we formalized in the taxonomies discussed above.

At times, one might want to know which taxonomy level is doing the heavy lifting. By specifying certain taxonomy levels by which to compare events, or by weighting taxonomy levels differently, one is able to better assess which assumptions are driving the final integration results. This approach, which can help with fine-tuning the input assumptions for MELTT to gain the most valid match possible, is easy to implement within the `meltt` *R* package.

Inspecting the Data

Following integration, the MELTT protocol offers a summary of the output, including how many entries were matched (or not) and how those matches are structured across datasets. This output allows the user to engage in manual inspection and to adjust the input assumptions to seek to produce better integration results. This manual review is particularly important when matching events to episodes. Technically, episodes (events with different start and end dates) and events are specified at different units of analysis. Thus, user discretion is required to help sort out these types of matches. The `meltt` *R* package offers functionality that eases this process.

Visualization

For quick visualizations of the output, `meltt` includes three plotting functions. The first function generates a bar plot that graphically displays the unique and matching entries, as seen in Figure 3 (upper left panel) of the article.

The second function generates a time-series plot of the integrated output. Raw counts of the unique entries are plotted on top of the timeline, while raw counts of the matches are plotted below. This plot offers a quick snapshot of *when* matches are found. Temporal clustering of matches may indicate an issue with the input assumptions, or could be an artifact of data-generating processes of the datasets. Users can specify the temporal unit that the data should be binned (day, week, month, year) for purpose of the plot. In Figure 3 (lower left panel) of the article, we use a monthly resolution.

The third function presents a summary of the spatial distribution of the data, by plotting the integrating entries onto a map. Unique and matching entries are labeled separately using color coding. The utility is to gain a sense of the spatial distribution of the integrated output, which can be especially valuable to identify any clustering or other disproportionate coverage in where unique and matching entries are located. Figure 3 (right panel) in the article presents a generic map with just the country border outlined.

Validating Output

The `meltt` protocol provides a method and framework for validating the output from the data integration task, which operates in three steps. First, the total population of matches located by the protocol are expanded to dyadic pairings (e.g., if entry 1 from dataset A, entry 3 from dataset B, and entry 5 from dataset C are identified as a match, then this A1-B3-C5 pair is expanded into a dyadic pairing and permuted to account for each combination: A1-B3, A1-C5, B3-C5). The function randomly generates a sample from these located matches. The user specifies the size of this sample as a proportion of the total available dyadic-form of the matches via the `sample_prop=` argument. A "control group" is generated by scanning through the data and finding all entries that fall within the same spatio-temporal window as used in the integration step, but that were not identified as matches. Two non-matching but proximate entries are located for each matched dyad. Note that the function locates the closest entry that was not flagged as a match to the dyadic entry being referenced; however, sometimes the closest entry in the data is substantially distant in terms of space and/or time.

Second, the function initializes an *R* `Shiny` interface that presents a "main entry" along

with three "candidate entries." The candidate entries consist of the two proximate control entries and one of the matching entries from the dyad, meaning one of the three entries is a match identified by the protocol. The user is prompted to compare the three candidate entries to the main entry, with an instruction to identify the "most likely" match of the three. A button appears next to each candidate entry; the user presses one of the buttons to record his/her choice. When a choice is made, a small text field appears on the screen noting the choice made. The scope of the descriptive information provided for each entry is specified at the outset by the user; the default option is the variables for which taxonomies were defined and used in the integration. The user proceeds through each entry one at a time until all sampled entries have been reviewed. Options to go forward and backward along the list are available as arrow buttons in the application.

Third, once all entries have been reviewed, the function reports performance diagnostics: the number of true/false positives/negatives; the true positive rate (TPR), which is the sum of the true positives/negatives; and the false positive rate (FPR), which is the sum of the false positives/negatives. In addition, information about the sample is recorded. The performance statistics are generated by comparing the matches identified by the user to those identified by the protocol.

B.2 Additional Details of the Protocol

B.2.1 Blocking

We presented a blocking strategy that relies on the spatial and temporal attributes recorded in dataset entries. A possible alternative could entail comparing only entries that are reported for a particular country, in a particular year, and/or of a particular type. Such an approach would also certainly reduce the volume of comparisons, potentially by substantial amounts. The downside, however, is that country, year, and type may be much too inflexible as boundaries of differentiation.

Consider that an entry recorded in one dataset as being in Country A and an entry recorded in another dataset as being in Country B may be coded as occurring in close geographic proximity along a border shared by the two countries. An entry recorded in one dataset as being in Year T and an entry recorded in another dataset as being in year

T+1 may be coded as occurring in close temporal proximity, perhaps even on consecutive days. An entry recorded in one dataset as being of Type X and an entry recorded in another dataset as being of Type Y may simply be labeled differently, based on the conventions of the respective datasets.

In all these instances, the pairs of entries could actually relate to the same event. Potential matches between entries would be overlooked when applying blocking strategies that require within-country, within-year, and within-type matches, respectively. Instead, we believe a blocking strategy that limits the consideration to the recorded location and timing of events is the most plausible, straightforward and effective option.

B.2.2 Iteration

For computational efficiency, the protocol relies on a very efficient time iteration procedure. Prior to the comparisons, all entries in the first and second dataset are pooled together, then their temporal order is established. The ordering of the entries for each dataset is recorded in an index. Comparison of entries is then performed by iterating through the index of the first dataset, finding the closest entry in the index of the second dataset and then iterating through the index of the second dataset—first backward, then forward—within Δt. This approach ensures that the procedure automatically skips time intervals with no events in either dataset, minimizing the number of iterations required for exhaustive comparisons. The blocking approach ensures that the protocol only compares other event attributes for those entries that co-occur spatially and temporally.

If two or more entries record events as occurring at the same time, they will be compared in random order. Such randomization, however, has no impact on the calculated matches, since comparisons are ultimately performed by considering absolute time differences between entries. Entries with the same timestamp will be considered, regardless of their ordering in the index. The only thing that changes is the order in which they are identified as potential matches (or not). This order has no bearing on the matching score S and on the likelihood with which a potential match is ultimately selected as the most likely match.

We rely on the matching score S, using a "deferred-acceptance algorithm" (Gale and Shapley 1962), to disambiguate among potential corresponding entries and establish optimal matches. The algorithm is robust to incomplete assignment, i.e., not all entries must be

included in the ultimate set of matching entries we identify. Furthermore, the algorithm was recently shown to be an unbiased approach to identifying those entries that agree most closely, as long as the datasets being compared do not contain exactly the same number of entries (Ashlagi, Kanoria and Leshno 2017), which is rarely the case in empirical applications.

B.2.3 MELTT Options

Static Matching

The MELTT procedure, as a default, relies on a dynamic matching logic: it iterates through each level of the taxonomy and finds the most fine-grained level on which two entries match. As an option, `meltt` also allows for static matching across taxonomies. For this option, the user has to specify a specific level for each taxonomy at which to compare events. In that case, the matching score S is irrelevant because all matches identified by construction will have exactly the same score. Through the acceptance-deference algorithm, the protocol then—effectively randomly—selects a final set of matches, since all candidate matches are equivalent. The static matching option is meant for sensitivity analyses of integration results. Specifically, the researcher can generate statistics on how many events match on specific taxonomy levels, i.e., at a selected coarseness in the comparison of secondary event characteristics.

Averaging

When collapsing the information of matched entires, the default in MELTT is to retain all information—including the geo-locations, timestamps, and other event attributes—of those entries. Only one set of information, from the initial dataset, is used for a given pair of matched entries in subsequent comparisons performed as part of an integration. This approach has potential implications for the matches that the procedure can be expected to yield. One way that the user can test the sensitivity of results to this approach is to vary the sequence in which datasets are compared during the integration process, which will dictate the information that is used from matched entries for the purpose of performing subsequent comparisons. In addition, we designed MELTT so that a user can opt to specify that the average value across a pair of entries is retained, for certain event characteristics. This option may be worthwhile for event characteristics such as location and timing, in so

far as averages can afford greater accuracy and precision than the values that any single dataset reports. A downside of averages is the potential for "chain matches." For example, the average locations of paired entries in datasets 1 and 2 and of paired entries in datasets 2 and 3 may be within Δs of one another, but the locations of the entries from datasets 1 and 3 could actually be further apart than the bounds of the spatial fuzziness parameter the user specifies. For this reason, averaging event characteristics when integrating datasets is generally not recommended as a primary let alone exclusive strategy.

B.3 Generalization to Multiple Datasets

The discussion in the article explored integration and disambiguation of entries in two datasets. Yet users may wish to integrate data from more than two sources. Identifying potential matching entries across more than two datasets is straightforward using the protocol. For more than two datasets, however, a unique ordering by matching score, required for the identification of matching entries, is no longer guaranteed to exist. Therefore, a trivial generalization of the protocol described in the article to more than two datasets is not possible.

Instead, we generalize the protocol as follows. First, MELTT integrates datasets 1 and 2, then integrates the merger of those two datasets with dataset 3, the subsequent merger with dataset 4, etc. If the coding framework and quality are comparable across datasets, the sequence in which datasets are integrated should be largely irrelevant. In cases where the coding of events differs across datasets, we propose to select the coding of event attributes of the dataset that entered earlier into the integration sequence, for any given pair of entries.

As mentioned above, for maximum transparency, the protocol always retains in the output the full set of attribute information from all matching entries. The user can opt to review the information and gauge on this basis whether the identified matching entries are in fact consistent enough to capture the same event. If the coding quality or reliability differs across datasets, the protocol performs best if datasets are provided in order of decreasing reliability. We generally recommend to perform integration multiple times with different dataset orderings to ensure this element has no effect on the results. Further, if spatial precision codes vary strongly across datasets, we suggest to include those codes as an explicit taxonomy.

This way, matching pairs with different precision codes are penalized, favoring—in the case of ambiguity among matching entries—those with the same geospatial precision.

B.4 Generalization to Episodal Data

Certain event datasets, by construction, employ units of observation in which each entry corresponds to a single day. Such an entry could concern one day of a multi-day event—what we call "episodes"—or an event that occurred on just one day. The coding of ACLED data, for example, follows this logic. Yet other event datasets, again by construction, contain entries about episodes that are coded as starting and ending on different days. For example, this is true of both UCDP-GED and SCAD. These differences present an issue when comparing entries, since multiple entries in an event dataset with only day-length observations could actually correspond to a single entry in another event dataset that allows episodal observations. Because this issue arises with some regularity, we have generalized our data integration protocol to accommodate the treatment of episodal coding of events.

First, our protocol compares any entry recorded as associated with a single day (i.e., the start and end dates are identical) to all other entries that are likewise recorded as associated with a single day. Second, the protocol compares entries that qualify as episodes (i.e., the start and end dates are not identical), with the separate requirements that the start dates have to fall within Δt of one another, as do the respective end dates. Third, the procedure identifies any remaining single-day entry as a possible "episodal match" if a given single-day entry has a timing that falls within the period—or Δt of the start or end dates—of a given entry about an episode in another dataset, in addition to corresponding in terms of the recorded location and other event attributes. Due to the disparities in construction across the datasets, we only mark the single-day entries with possible episodal matches for subsequent manual inspection. Automatic integration of single-day entries with multi-day episodes is simply beyond the scope of an automated protocol because it is not possible without in-depth manual inspection to decide how many single-day entries aggregate to a multi-day episode. This may also be impossible even after manual comparison, due to lack of detail in the coding of events. The protocol at least minimizes the manual effort to be expended, by flagging the entries to review.

C MELTT Performance Tests

C.1 Synthetic Data

Our in-depth analysis of the performance of MELTT, as reported in the article, relies on synthetic data. In constructing these data, we first specify for each dataset J, the number of entries known to match to an entry in another dataset, M, and the total number of entries across datasets, N. We also specify a geographic bounding box, assigning entries only locations within this boundary. Similarly, we specify a temporal bounding box. Another possibility is to generate synthetic data with "incomplete" overlap, whereby only subsets of entries match up across datasets. Our approach has the advantage of simplifying the presentation and discussion of results, without exerting any impact on the findings. For simplicity, we assume that an entry in one dataset with a match has a match in every other dataset. Thus, we generate $J \times M$ matching entries—e.g., given four synthetic datasets and four known matches per dataset, a total of 16 entries would match up by design, producing 24 matched pairs of entries.

To generate entries that match, we begin by randomly assigning the location and timestamp of an entry in the first dataset. Entries in other datasets are then defined to match the entry in the first dataset, with randomly assigned locations and timestamps that fall within a specified spatio-temporal window of coding uncertainty. Spatial coding uncertainty of 1 km means that any matched entry can fall within a maximum radius of 1 km of the other entry to which it is matched. Note that this implies that if more than two datasets are generated this way, events for coding uncertainty 1 km can be up to 2 km apart if they are scattered in exactly opposite directions from the entry in the first dataset. Similarly, temporal coding uncertainty of one day means an entry may be recorded as much as a day before or after any matching entries—e.g., given an entry in dataset A reported as occurring on day 2, a matching entry in dataset B may occur on day 1, 2 or 3.

As a baseline, we assume a spatial coding uncertainty of 2 km, and 1 day of temporal uncertainty. In specific treatments, we vary the degree of uncertainty to systematically evaluate the impact on the performance of MELTT. We also assume that the other event attributes of entries that constitute matches are entirely identical. To specify these characteristics,

Table C.1 Schematic representation of default taxonomies used to generate artificial data

base category	category 1	category 2
taxonomy 1: event type		
event type 1	event category 3	broad event category 2
event type 2	event category 5	broad event category 1
event type 3	event category 1	broad event category 2
event type 4	event category 6	broad event category 2
event type 5	event category 4	broad event category 1
event type 6	event category 2	broad event category 1
event type 7	event category 1	broad event category 2
event type 8	event category 3	broad event category 2
event type 9	event category 5	broad event category 1
taxonomy 2: actor		
actor 1	actor type 1	
actor 2	actor type 2	
actor 3	actor type 1	
actor 4	actor type 2	
actor 5	actor type 2	
taxonomy 3: precision		
actor 1	actor type 1	
actor 2	actor type 2	
actor 3	actor type 3	
actor 4	actor type 2	
actor 5	actor type 3	

Note. We assumed that each entry records three other event attributes: event type, actor, and geo-spatial precision. The nine unique event types distribute across six event categories, which fall within two broad event categories. The five actors fall within two actor types.

we simply assume a given taxonomy structure—the number of other event attributes, the number of levels per attribute, and the number of distinct categories per level of attribute (see Table C.1)—and then randomly draw an attribute for each matched pair of entries. At the broadest taxonomy level, however, we ensure that matches we construct always agree. In other words, introducing uncertainty in the coding of other event attributes systematically varies the quality of the match, but not which events are matched.

In addition to the M matches, we generate a rich background of $N - M$ unrelated entries across the J datasets. These entries are placed randomly within the spatio-temporal bounding box and randomly assigned other event attributes. When generating the synthetic data, we mirror the spatio-temporal dimensions of our empirical illustration for our first

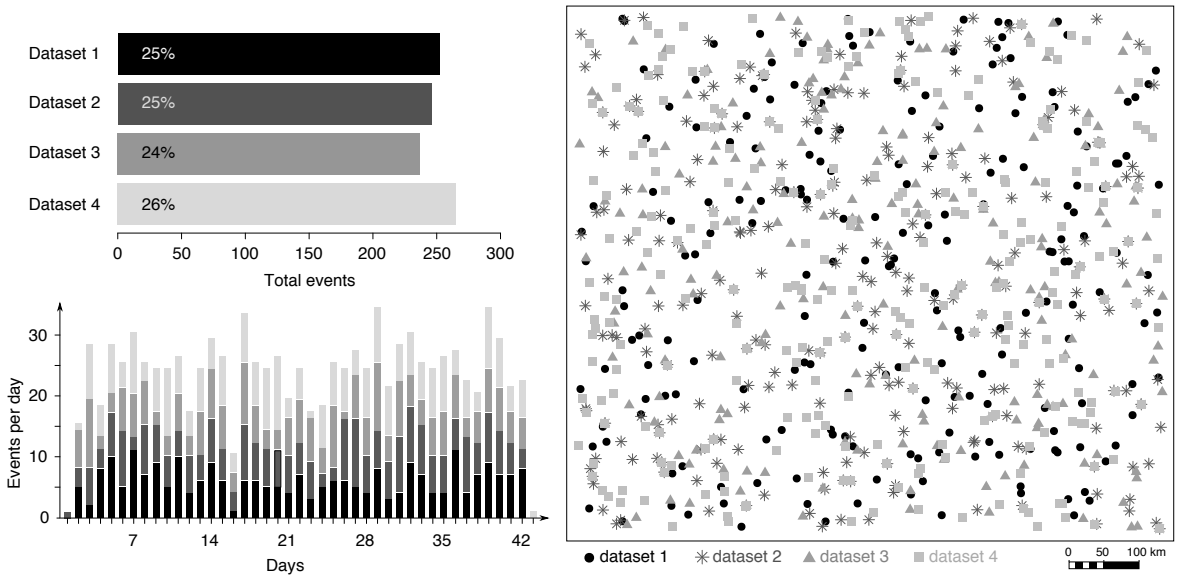

Fig. C.1 Illustration of synthetic event data used to test the performance of MELTT

case, Nigeria, and use the same number of datasets (4) and around the same number of total entries (1000).[1] The synthetic data also reproduces common challenges faced in empirical settings such as intrinsic coding uncertainty in the specification of locations and timestamps, multiple ambiguous matches, unique events that cluster with matches and may be very similar in their event characteristics, etc. Figure C.1 provides a graphical depiction of a representative example of the synthetic data that we use in our analyses.

C.2 Scenarios

This section details scenarios summarized in Table 1 of the article. We rely on four simulated datasets with the exception of one scenario, which varies the number of input datasets. If not explicitly stated otherwise, each scenario follows the same basic specifications. We generate a total of 1000 events with 50 known matches (45 of which are strictly events, 5 of which are strictly episodes), that fall within geographic boundaries similar to those of Nigeria and a time span that mirrors the one in that example case (1-1-2011 to 1-1-2012). In generating the data, we set the intrinsic coding uncertainty to 2 km for locations and 1 day for timestamps, i.e., potential matches can fall within a radius of 2 km and one day before or after one another. We utilize three simulated taxonomies to match the observations: one with depth

[1]Departing from the empirical case, our synthetic datasets each have similar numbers of entries.

(9,6), the others with depth (5,2) and (5,3). Unless stated otherwise, spatial fuzziness is set to 4 km and temporal fuzziness is set to 2 days to account for the full extent of the assumed intrinsic coding uncertainty.

C.2.1 Changing Order of Input Datasets

The first scenario varies the order of the input datasets to test whether this has any impact on the output generated by MELTT. Specifically, we iterate through every permutation of the ordering of the four dataset as an argument of MELTT.

C.2.2 Specifying Different Spatio-temporal Fuzziness

The second scenario tests the impact of the spatio-temporal fuzziness on MELTT matches. In this scenario, we assume intrinsic coding uncertainties of 1 km for locations and 1 day for timestamps and increase fuzziness parameters from the minimum value necessary to account for the intrinsic coding uncertainty. The tests proceed in three steps:

(a) We systematically vary the spatial fuzziness of MELTT matches, going from units of 2 to 21 km by intervals of 1. The temporal fuzziness is fixed at 2 days.

(b) We systematically vary the temporal fuzziness of MELTT matches, going from units of 2 to 21 days by intervals of 1. The spatial fuzziness is fixed at 2 km.

(c) We systematically vary both the spatial and temporal fuzziness of MELTT matches, going from units of 2 to 21 days/km by intervals of 1 days/km respectively.

C.2.3 Increasing the Number of Input Datasets

The third scenario increases the number of input datasets—from 2 to 10—that the procedure integrates. The number of known matches is reduced, relative to our baseline scenario, to accommodate the increase in datasets.

C.2.4 Varying the Number of Event Attributes

The fourth scenario varies the number of other event attributes evaluated in the matching procedure. We accomplish this by duplicating the same taxonomy (9,6,2) from 1 (e.g. 9,6,2) to 2 (e.g. 9,6,2,9,6,2) and so forth all the way to 10 times (ultimately yielding 10 event attributes, each with 3 categories).

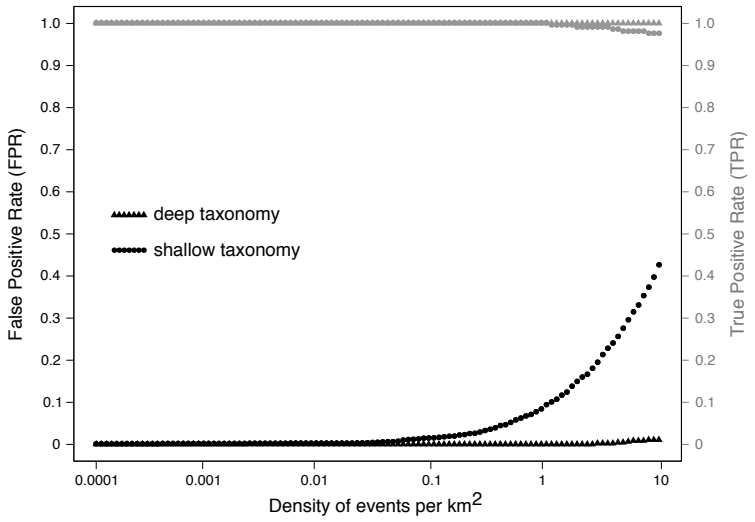

Fig. C.2 Performance of the data integration protocol as a function of the density of events

C.2.5 Increasing the Richness of a Single Taxonomy

The fifth scenario increases the richness of a single taxonomy. The scenario begins with a taxonomy of depth (3,2) and increases the number of distinct categories at a given taxonomy level (e.g. 4,2 then 5,2 etc.).

C.2.6 Increasing the Density of Observations

The sixth scenario generates a large number of events that occur in close geo-spatial proximity, such as what would occur with a concentration of conflict events in population centers. The test also explores the effect of shallow versus rich taxonomies on identifying matches under these conditions. We generate a set of rich taxonomies ((12,7,3),(10,6,2),(9,5,2)) and a single shallow taxonomy ((1,1)), then run the integration procedure under conditions with a large number of events within a constricted geo-spatial domain, achieved by reducing the spatial bounding box from $10km^2$ to $1km^2$. The results for this scenario are illustrated in Figure C.2.

C.2.7 Increasing the Proportion of Imprecise Entries

The seventh scenario simulates variation in location coding, closely corresponding to the precision categories used across most event level datasets. Four precision levels are considered: exact location, level 1 centroid (e.g. Administrative 2 Level, i.e., district), level 2 centroid (e.g. Administrative 1 Level, i.e., province or region), and level 3 centroid (e.g

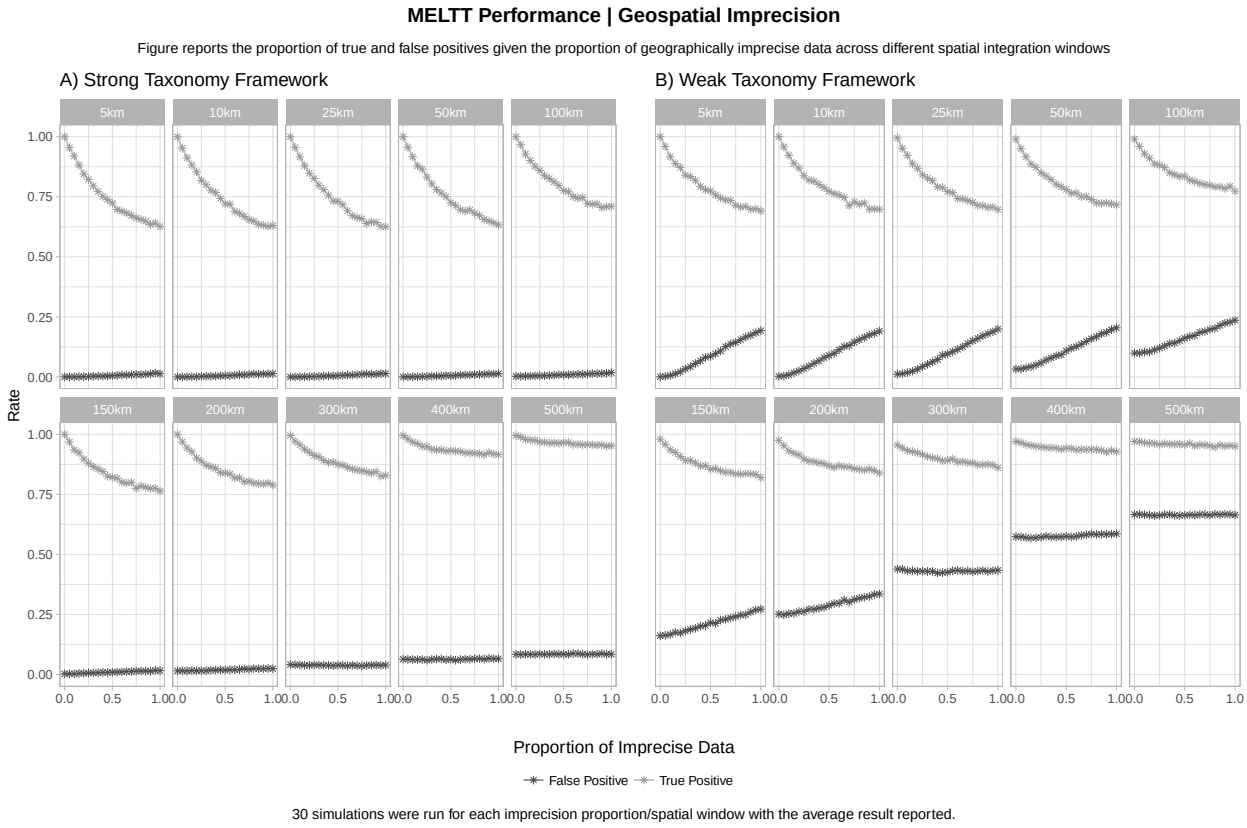

Fig. C.3 Performance of the data integration protocol as a function of data imprecision with a strong and weak taxonomy framework

Administrative 0 Level, i.e., country). We then randomly assign a set proportion of the input events to different levels of imprecision. The data generation corresponds with baseline conditions outlined above. As assignment of imprecise entries is random both with respect to the unit and the input dataset, we run the simulation 30 times and take the average. The simulation examines the impact of imprecision as the data goes from 0% of entries being imprecise to 100% of entries being imprecise. We then explore different spatial windows to see if performance improves.

In addition, we explore the algorithm's capacity to correctly identify matching entries when a strong versus a weak taxonomy is used to disambiguate between events—Table 1 in the article reports results from the strong taxonomy case. We generate a set of rich taxonomies ((12,7,3),(10,6,2),(9,5,2)) and a single shallow taxonomy ((1,1)), then run the integration procedure altering the different levels of geospatial imprecision in the data. For the strong taxonomy scenario, MELTT is sufficiently capable of disambiguating between

matching and non-matching entries; however, as the proportion of imprecise entries increases, the algorithm is unable to capture true positives that are offset by the spatial distortion. This is due to the fact that proximate entries are in essence relocated to centroid locations. To remedy this situation, one can increase the spatial window to effectively incorporated the centroid coordinates. Panel A of Figure C.3 demonstrates how expansion of the spatial window can lead to better recovery of true matches.

By contrast, Panel B of Figure C.3 reports the performance results of the weak taxonomy scenario. The results demonstrate that the FPR increases significantly as the proportion of imprecise entries increases. The figure highlights the importance of a robust taxonomy framework when expanding spatial windows during integration. In essence, the problem is similar to the spatial clustering issue outlined above. As more events are located within the same bin, the better the secondary metadata needs to be. If sufficiently detailed information is not available for disambiguation, we may still recover true positive matches for sufficiently large integration windows but the number of false positives is prone to increase substantially.

D Empirical Demonstration

D.1 Data Overlap for Nigeria 2011

In the article, we offer descriptive statistics for the degree of overlap when pooling events from the ACLED, UCDP-GED, GTD, and SCAD datasets. We find that in the Nigerian 2011 subset, 59.5% of all entries fall within one day and 3 km of an entry in another dataset. Using alternative thresholds of spatio-temporal proximity yields the following results:

(a) 5% of entries are recorded with the same day and longitude/latitude locations as an entry in another dataset.

(b) 5.6% of entries fall within ±1 day and have the same longitude/latitude locations as an entry in another dataset.

(c) 46.1% of entries fall within ±1 day and 1 km of an entry in another dataset.

(d) 67.5% of entries fall within ±3 days and 10 km of an entry in another dataset.

Examining clustering highlights the high degree to which the location and timing of events coincide across the four datasets. This underscores the need for a robust data integration approach that can reliably disambiguate among unique and matching entries using other event attributes.

D.2 Taxonomies

Here, we discuss the construction of the event, actor and precision taxonomies employed in the empirical analysis of Nigeria 2011. The levels work in increasing order, with the lowest category being the most granular and the highest category being the broadest.

D.2.1 Event Taxonomy

Our event taxonomy (Table D.2) is based on the original event attributes coded in the four datasets, which we then abstract to increasingly broader, plausible categories. All four datasets employ a categorical variable to differentiate types of event activity (ACLED = "EVENT_TYPE", UCDP-GED = "type_of_violence", GTD = "attacktype1", and SCAD

Table D.2 Event Taxonomy

Data Source	Original Event Type	Level 1	Level 2	Level 3
ACLED	Non-violent transfer of territory	Territorial Dispute	Nonviolent Possession	Nonviolent Event
ACLED	Headquarters or base established	Territorial Dispute	Nonviolent Possession	Nonviolent Event
ACLED	Riots/Protests - Protest (if fatalities == 0)	Protest/Demonstration	Nonviolent Displays	Nonviolent Event
ACLED	Non-violent activity by a conflict actor	Protest/Demonstration	Nonviolent Displays	Nonviolent Event
SCAD	Organized Demonstration	Protest/Demonstration	Nonviolent Displays	Nonviolent Event
SCAD	Spontaneous Demonstration	Protest/Demonstration	Nonviolent Displays	Nonviolent Event
SCAD	General Strike	Protest/Demonstration	Nonviolent Displays	Nonviolent Event
SCAD	Limited Strike	Protest/Demonstration	Nonviolent Displays	Violent Event
ACLED	Riots/Protest - Riots (if fatalities >= 1)	Violent Protest/Demonstration	Violent Displays	Violent Event
SCAD	Organized Violent Riot	Violent Protest/Demonstration	Violent Displays	Violent Event
SCAD	Spontaneous Violent Riot	Violent Protest/Demonstration	Violent Displays	Violent Event
GTD	Hijacking	Coercion	Violent Possession	Violent Event
GTD	Hostage Taking (Barricade Incident)	Coercion	Violent Possession	Violent Event
GTD	Hostage Taking (Kidnapping)	Coercion	Violent Possession	Violent Event
ACLED	Battle-Non-state actor overtakes territory	Territorial Dispute	Violent Possession	Violent Event
ACLED	Battle-Government regains territory	Territorial Dispute	Violent Possession	Violent Event
ACLED	Battle-No change of territory	Territorial Dispute	Violent Attack	Violent Event
GTD	Bombing/Explosion	Strategic Destruction	Violent Attack (Bombing)	Violent Event
ACLED	Remote violence	Strategic Destruction	Violent Attack (Bombing)	Violent Event
GTD	Assassination	Strategic Assault	Violent Attack	Violent Event
GTD	Armed Assault	Strategic Assault	Violent Attack	Violent Event
GTD	Unarmed Assault	Strategic Assault	Violent Attack	Violent Event
GTD	Facility/Infrastructure Attack	Strategic Destruction	Violent Attack	Violent Event
SCAD	Pro-Government Violence (Repression)	State-led Violence	Violent Attack	Violent Event
SCAD	Anti-Government Violence	Opposition-led Violence	Violent Attack	Violent Event
SCAD	Intra-government Violence	Within-Regime Violence	Violent Attack	Violent Event
GED	State-based conflict	State-led Violence	Violent Attack	Violent Event
GED	Non-state conflict	Opposition-led Violence	Violent Attack (No State)	Violent Event
SCAD	Extra-government Violence	Opposition-led Violence	Violent Attack (No State)	Violent Event
ACLED	Violence against civilians	Atrocity	Violent Attack (Against Civilians)	Violent Event
GED	One-sided violence	Atrocity	Violent Attack (Against Civilians)	Violent Event

= "etype"). We then generalize from these event types, defining categories that are broad enough to encompass two or more types spanning two or more datasets. From there, we generalize even further, ultimately arriving at the broadest taxonomy level, which delineates only between violent and nonviolent events.

D.2.2 Actor Taxonomy

Actor taxonomies (Table D.3) were created using key words of actor descriptions in ACLED and SCAD—by design, both datasets incorporate violent and nonviolent actor types. For UCDP-GED, we differentiate between state ("government") and non-state ("violent groups") actors in Level 1 and then categorize all as "violent groups" for Level 2. GTD, by design, is composed completely of violent non-state actors—thus both Level 1 and Level 2 are coded as "violent groups." The full actor taxonomy used is provided as part of the replication material.

D.2.3 Precision Taxonomy

The precision taxonomy (Table D.4) was built using a geo-spatial precision estimate present in each of the relevant datasets (ACLED = "GEO_PRECIS", UCDP-GED = "where_prec",

Table D.3 Actor Taxonomy

Key Words	Level 1	Level 2	Key Words	Level 1	Level 2
sect	religious groups	nonviolent groups	community	civilians	nonviolent groups
cult	religious groups	nonviolent groups	communities	civilians	nonviolent groups
islamists	religious groups	nonviolent groups	boys	civilians	nonviolent groups
catholics	religious groups	nonviolent groups	fans	civilians	nonviolent groups
nurses	religious groups	nonviolent groups	civilians	civilians	nonviolent groups
muslims	religious groups	nonviolent groups	villagers	civilians	nonviolent groups
christians	religious groups	nonviolent groups	magistrates	government	violent groups
movement	movement groups	nonviolent groups	police	government	violent groups
party	movement groups	nonviolent groups	military	government	violent groups
activists	movement groups	nonviolent groups	judicial staff	government	violent groups
protesters	movement groups	nonviolent groups	securtiy service	government	violent groups
demonstrators	movement groups	nonviolent groups	government	government	violent groups
supporters	movement groups	nonviolent groups	soldier	government	violent groups
bakers	civilian groups	nonviolent groups	council	government	violent groups
workers	civilian groups	nonviolent groups	militants	violent groups	violent groups
association	civilian groups	nonviolent groups	militant	violent groups	violent groups
unions	civilian groups	nonviolent groups	robbers	violent groups	violent groups
union	civilian groups	nonviolent groups	bandits	violent groups	violent groups
teachers	civilian groups	nonviolent groups	gunmen	violent groups	violent groups
doctors	civilian groups	nonviolent groups	bomber	violent groups	violent groups
drivers	civilian groups	nonviolent groups	bombers	violent groups	violent groups
vendors	civilian groups	nonviolent groups	kidnappers	violent groups	violent groups
ranchers	civilian groups	nonviolent groups	vigilantes	violent groups	violent groups
congress	civilian groups	nonviolent groups	militiamen	violent groups	violent groups
employees	civilian groups	nonviolent groups	attackers	violent groups	violent groups
journalists	civilian groups	nonviolent groups	assailants	violent groups	violent groups
nomads	civilian groups	nonviolent groups	gangs	violent groups	violent groups
tribe	civilian groups	nonviolent groups	rioters	violent groups	violent groups
herds	civilian groups	nonviolent groups	rebels	violent groups	violent groups
students	civilians	nonviolent groups	mob	violent groups	violent groups
youth	civilians	nonviolent groups	pirates	violent groups	violent groups
youths	civilians	nonviolent groups	armed group	violent groups	violent groups
women	civilians	nonviolent groups	thugs	violent groups	violent groups
citizens	civilians	nonviolent groups	fundamentalists	violent groups	violent groups
residents	civilians	nonviolent groups	boko haram	violent groups	violent groups
passengers	civilians	nonviolent groups	militia	violent groups	violent groups

GTD = "specificity", and SCAD = "geo_precis"). Like the event taxonomy, we first generate levels that are broad enough to capture two or more indicators, while still being mutually exclusive. The broadest category distinguishes only between precise and imprecise geo-codes. The precision taxonomy generates a penalty for geo-location codes that are less precise than an event with a greater degree of location precision.

D.2.4 Taxonomy Adjustment

Taxonomies are formalizations of criteria of comparison across datasets. Nuances within a given dataset's coding scheme can complicate efforts to generalize, especially if those nuances contribute to significant deviations from the other dataset(s) encompassed by a taxonomy. One such nuance appears within the UCDP-GED coding scheme. For all events coded as state-based conflicts, the actor reported as "side_a" within the dyadic set is always the government, by convention. This coding deviates from that of other datasets. Consider an event several datasets record as occurring on August 16, 2011 around latitude = 12.404 longitude = 4.65. ACLED, UCDP-GED, and SCAD report that gunmen fired upon a police

Table D.4 Precision Taxonomy

	Data Source	Base Category	Level 1	Level 2	Level 3
1	GED	1	exact town, area, city, village	Subregional	precise
2	GED	2	fuzzy space around town, city, village (outskirts, checkpoints)	Subregional	imprecise
3	GED	3	within a 2nd order admin region	Subregional	imprecise
4	GED	4	within a 1st order admin region	Subnational	imprecise
5	GED	5	within country	National	imprecise
6	GED	6	within country	National	imprecise
7	GED	7	unknown or other	Unknown	imprecise
8	ACLED	1	exact town, area, city, village	Subregional	precise
9	ACLED	2	within a 2nd order admin region	Subregional	imprecise
10	ACLED	3	within a 1st order admin region	Subnational	imprecise
11	GTD	1	exact town, area, city, village	Subregional	precise
12	GTD	2	fuzzy space around town, city, village (outskirts, checkpoints)	Subregional	imprecise
13	GTD	3	within a 2nd order admin region	Subregional	imprecise
14	GTD	4	within a 1st order admin region	Subnational	imprecise
15	GTD	5	unknown or other	Unknown	imprecise
16	SCAD	1	exact town, area, city, village	Subregional	precise
17	SCAD	2	exact town, area, city, village	Subregional	precise
18	SCAD	3	fuzzy space around town, city, village (outskirts, checkpoints)	Subregional	imprecise
19	SCAD	4	exact town, area, city, village	Subregional	imprecise
20	SCAD	5	fuzzy space around town, city, village (outskirts, checkpoints)	Subregional	imprecise
21	SCAD	6	within a 1st order admin region	Subnational	imprecise
22	SCAD	7	within country	National	imprecise
23	SCAD	-99	unknown or other	Unknown	imprecise

station, killing four policemen and two civilians. SCAD codes the actor as "Gunmen"; ACLED as "Boko Haram"; and UCDP-GED as the "Government of Nigeria". In UCDP-GED, an event can only fall into one of three types: state-based violence, non-state conflict, and one-sided violence. Since this event involves the state, the primary actor is always coded as the government. We adjust for this nuance by taking the side_b actor when composing the taxonomy for all events coded as state-based violence. Though not a perfect solution, we find through validation and adjustment that this yields the best recovery of matching events.

As noted in the article, taxonomy adjustment and refinement is an essential part of the MELTT procedure. The input assumptions must always be explored and validated to ensure that they are achieving what the researcher intended. As error is built into any comparison dimension, understanding the subtle differences in the variables encompassed by a taxonomy is central to generating results capable of sufficiently recovering potential matches.

D.3 Supplementary Analysis

This section provides further details for two supplementary analyses referenced in the article: (1) testing the impact of the ordering of datasets on the performance of MELTT, and (2) exploring the effect of variation in the spatial and temporal fuzziness on the number of matches obtained in the empirical data.

D.3.1 Sequence of Dataset Comparison

Our analysis using synthetic datasets did not find any significant effect on the performance of MELTT associated with the sequence in which datasets were compared. We repeat the analysis for all integration sequences of the four conflict event datasets used in the empirical analysis for Nigeria 2011 (ACLED, UCDP-GED, GTD, and SCAD). Specifically, we run MELTT for every potential sequence of datasets, holding the spatiotemporal fuzziness constant at 3 km and 1 day (Figure D.4). In keeping with the finding about the synthetic data that we report in the article, no significant disparity in results is observed across different sequences. We find only slight variation in the number of unique events with a minimum of 684 (S-A-GT-G) and a maximum of 688 (A-G-GT-S, A-G-S-GT, A-S-G-GT, GT-A-G-S, GT-A-S-G, GT-G-A-S, GT-G-S-A) events identified.

D.3.2 Impact of Fuzziness Parameters

To explore the effect of different degrees of spatio-temporal fuzziness on the extent of matches, we run the MELTT procedure across a series of specifications ranging spatially from 0 to 10 km and temporally from 0 to 3 days. Figure D.5 reports the total number of events across datasets that MELTT identified to be matches. In addition, we track the number of unique changes in the matches as we expand the fuzziness parameters. For example, as the spatial window is expanded from 1 to 2 km (holding the temporal window at 1 day), 178 matches of the previously located 272 matches are altered either through the addition of a new match (e.g. event 3 in ACLED was matched to event 1 in GED and event 1 in GTD was added to the match) or the omission of a prior matching entry. Lastly, we describe the number of new matches that emerge from each manipulation.

The degree of matching between datasets stabilizes around 3 km for all four temporal settings. These results indicate that increasing spatial fuzziness beyond 3 km would result

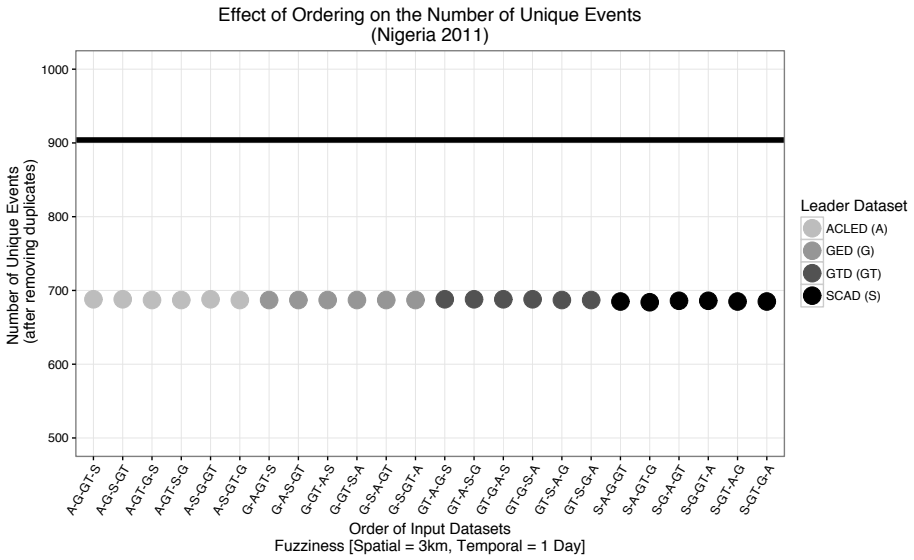

Fig. D.4 Number of unique events identified for every potential combination of the ACLED, UCDP-GED, GTD and SCAD datasets using MELTT. Spatial and temporal fuzziness are set s = 3km and t = 1 days respectively. Solid black bar denotes the number of unique events when pooling all four datasets and not accounting for potential matching entries (904). The figure demonstrates ordering does not substantially affect MELTT matches.

in only a marginal increase in the number of matching events. In turn, increasing fuzziness from zero to one day (holding the spatial window at 3 km) increases the share of matching events by roughly 6 percentage points. The number of matches stabilizes after 1 day, indicating only marginal gains when expanding the window past this point. We therefore set the fuzziness parameters to $\Delta s = 3$ km and $\Delta t = 1$ day.

In practice, researchers can determine, in a systematic, tractable way, the sensitivity of MELTT to choices about degrees of spatial and temporal fuzziness. Researchers can then use the results to find optimal specifications for the dataset integration.

D.4 Performance of MELTT for Nigeria 2011

To systematically evaluate the performance of MELTT, two coders manually analyzed the sample of 904 entries for Nigeria in 2011 from ACLED, SCAD, UCDP-GED, and GTD. Each entry was coded either as a correctly identified matching entry (true positive), correctly identified unique entry (true negative), incorrectly identified matching entry (false positive), or incorrectly identified unique entry (false negative).

This manual assessment of entries as either matching or unique highlights the challenges

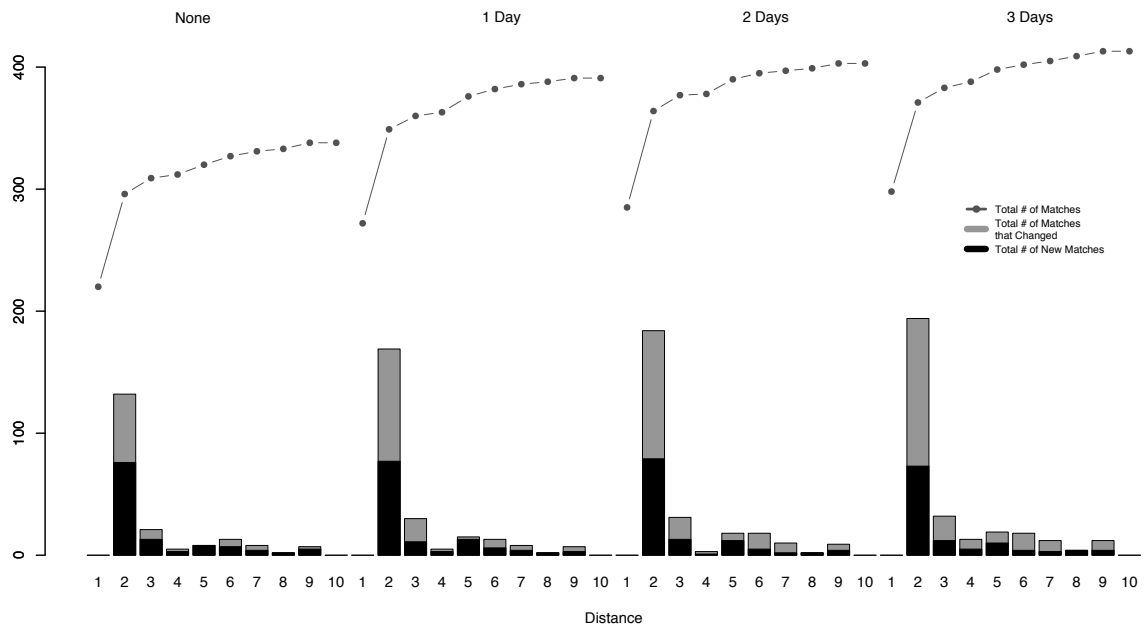

Fig. D.5 Total Number of Located Matches (altered and new matches) for ACLED, UCDP-GED, GTD, and SCAD in Nigeria in 2011.

of disambiguating event data. Consider the following example. The four datasets have a total of 9 entries on April 7 and 8 about temporally proximate events on the eve of parliamentary elections. Militants attacked officials distributing election materials near a police station, killing at least four and wounding many. On the same day, militants bombed an election office, killing at least 12. Finally, one militant was killed and another wounded when militants moved an explosive device.

While the spatial proximity parameter of 3 km separates these descriptively similar entries into sets of possible matching entries in Borno, Niger and Kaduna states, the other metadata from the event datasets both helps and hinders our efforts to establish which could be matches, if any. For example, event types are coded differently across datasets, hindering our efforts, but actor types are similar, referring to unknown militants or Boko Haram across datasets, suggesting potential matches.

The validation exercise does not involve comparing each entry to every other entry, which would require a large volume of human assessment (408,608 comparisons of pairs of entries). Instead, we examine entries that are sufficiently proximate, i.e., events that occur within

3 km and 1 day of another event. This blocking technique mirrors the process formalized within the MELTT procedure. Our logic is that only those entries considered proximate per our choice of spatial and temporal windows should be considered when evaluating whether the procedure correctly or incorrectly identifies a given entry as having a corresponding entry in another dataset.

For each set of entry that was identified as a "positive"—containing one or more entries that the procedure identifies as being potential matching entries—by MELTT, we examine the corresponding subset of proximate entries. Coders then compared the other attributes of these collections of entries, including event type, actor type, type of tactics, and extent of casualties, as well as any descriptions of named individuals, organizations, or settings and motives (e.g., reprisal attacks, assassinations, etc.).

Taking this descriptive evidence into account, the coders were directed to arrive at a judgment as to whether an entry from the other datasets, if any, could be reasonably judged as matching to the given entry. If this matching entry identified by the coders is the same as the matching entry identified by MELTT, this judgment is reflected as a true positive for both the given entry and the matching entry/entries. If the matching entry identified by the coders differs from the matching entry identified by MELTT, this judgment is reflected as a false positive for both the given entry and the matching entry. If no matching entry is identified, this judgment is also reflected as a false positive for the given entry.

The coders also examined the corresponding subset of proximate entries that were not identified as "positives." These entries are those that the procedure identified as being potential unique events. Following the same procedure as the one followed for establishing true and false positives, the coders were directed to arrive at a judgment as to whether an entry from the other datasets, if any, could be reasonably judged as matching the given entry. If no matching entry is identified by the coders, this judgment is reflected as a true negative for the given entry. If the coders identify a matching entry, this judgment is reflected as a false negative for both the given entry and the matching entry.

Table D.5 MELTT Performance for the Nigeria Conflict Event Datasets in 2011

	Indicator	Entries	% of Sample	Total %
correct	True Positives	344/ 345	38.01/ 38.16%	92.10/ 93.91%
	True Negatives	489/ 504	54.09/ 55.75%	
incorrect	False Positives	16/ 15	1.77/ 1.66%	7.85/ 6.08%
	False Negatives	55/ 40	6.08/ 4.42%	

Note: The numbers and fractions shown correspond to coder 1/ coder 2 throughout the table.

D.5 Validation Results and Intercoder Reliability

The results of the manual validation are summarized in Table D.5. The coding procedure yields the four nominal categories for assessing the performance of MELTT: correctly identified matching entry (true positive), correctly identified unique entry (true negative), incorrectly identified matching entry (false positive) and incorrectly identified unique entry (false negative).

Two coders analyzed the sets according to the scheme above. We calculated two reliability coefficients, percent agreement (91.8%) and weighted (squared) Cohen's Kappa (0.98), using the R package *irr*[2] for categorical data. The results indicate a high level of inter-coder reliability beyond the very close qualitative agreement evident in Table D.5. The accuracy rate of MELTT independently reported by both coders is very high for an untrained procedure. Iteratively refining the various taxonomies could further improve performance. Aside from accounting for the particularities of the actor coding in UCDP-GED (see Section D.2.4), however, we refrained from iteratively adjusting taxonomies, instead opting to demonstrate the MELTT protocol's "out-of-the-box" performance.

[2]Matthias Gamer, Jim Lemon and Ian Fellows Puspendra Singh (2012). "irr: Various Coefficients of Interrater Reliability and Agreement." R package version 0.84. `https://CRAN.R-project.org/package=irr`

D.6 Manual Verification of No Overlap Between Protest and Riot Events

We assessed in depth the finding of no matches between datasets in the protests event category. An initial step was to consider whether this result was due to inherent features of the MELTT protocol. Our manual review of Nigeria 2011, however, yielded only three instances in which one or both coders identified a false negative—the MELTT protocol should have identified a matching entry, but did not—for a protest event in ACLED and SCAD. Thus, the lack of matches appears to be largely a function of the respective data-generating processes for those datasets.

First, ACLED captures a lot more activity than SCAD overall. Therefore, the feasible extent of matching between the datasets is much smaller than the universe of all protest events that the datasets capture.

Second, ACLED often codes protest events at more precise geo-locations than does SCAD. As a consequence, MELTT is prone to overlook potential matches across the datasets, unless the spatial fuzziness parameter used in the blocking procedure is specified to be large. Due to the differing precision of location coding, the manual review could not determine with much certainty whether certain pairs of entries concern the same event as events located further apart are more likely to be unique.

Third, we encountered artifacts arising from the original coding schemes of the datasets, even after transforming them along the same time, space, and categorical lines—one of the main difficulties of integrating event data. Specifically, in order to integrate the data we had to split the "protests and riots" category in ACLED into two categories, separating protest from riots, where the latter are defined as protests in which at least one casualty occurs. Therefore, an ACLED protest event may have the word riot in its event description, but if there are no casualties, it is coded as a protest. In contrast, SCAD codes anything beyond an organized demonstration as a spontaneous violent riot in its original coding scheme. Thus, many protest events are characterized as riots in SCAD even if no casualties occurred.

In sum, the manual review indicates that a majority of the events are either unique to one of the datasets, or coded by SCAD at less precise geo-locations. Ultimately, we are

dealing with the subtle and often blurry transition from non-violent to violent behavior—a distinction which becomes much less subtle if the event data are then merged using higher levels of generality in the event taxonomy, e.g., non-violent displays vs. violent displays or violent actions, as discussed below.

D.7 Manual Analysis of Effect of Broad Taxonomy Levels on FPR

We also assessed the possibility that the broadest level of the event taxonomy (i.e., the distinction between violent and non-violent events) could generate excessive false positives. Due to this aspect of the event taxonomy, events can be deemed as matches if they are coded as involving any of many types of violent events across the four datasets, ranging from riots to acts of war. In order to exclude the possibility that these broad taxonomy levels lead to over-identification of matches, we manually analyzed all cases of false positive identification (FPR) after integration.

We found that the broadest level of the event taxonomy was the sole driver of such identification errors in only three of the cases of false positives. Meanwhile, the broad level of the event taxonomy was a factor, but not the sole factor, in about half of the false positives. Upon review, many false positive identifications concerned events that appear to be related, but not clearly matches, or instances of remote violence, bombing/ explosions, and armed assaults that occurred close enough in space and time to be considered proximate, but in fact transpired sequentially over the course of one or two days.

Different datasets may further differ in how they delimit individual entries. Even though our analysis strives for consistency in this definition across datasets, there remain subtle differences in how, for example, isolated incidents are delimited from events encompassing the use of violence of a given group, on a specific date, in a given location. This can further contribute to false positives. Ultimately, most of the misidentifications we found arose from such inconsistency or imprecision in the original coding and/or the absence of reliable information for disambiguation—reasons that are equally detrimental for MELTT as they are for comparable manual approaches to the integration of event datasets.

D.8 Degree of Missingness

To calculate the degree of missingness, we leverage a Latent Class Modeling approach for Multiple Systems Estimation (Manrique-Vallier, Price and Gohdes 2013). Unlike standard log-linear or Bayesian model averaging approaches to capture-recapture, the latent classifier approach treats each record as its own latent class and explicitly models the probabilities underpinning recapture as being heterogeneous in the population (see `https://hrdag.org/tech-notes/basic-mse.html` for a cursory review of these techniques and their implementation in *R*).

The capacity of the Latent Class Modeling approach to deal with the underlying heterogeneity in the population allows flexibility in recovering population estimates in conceptually variant environments. For instance, when the latent recapture rate is heterogeneous across the different lists, in addition to the sub-categories contained within those lists, the population estimates produced by the model can differ markedly given how one chooses to "slice" the data. In using MSE to model conflict events, we acknowledge that bias in media coverage (across event types, regions, and time) can influence the estimates. Moreover, the conceptual ambiguity in what qualifies as an "event" across the different datasets implies uncertainty in the units themselves. This generates challenges when employing methods that require closed populations, where there is no ambiguity in the unit being sampled.

We demonstrate this variability in the estimated true population of events for Nigeria 2011 data when considering individual violent event types, time periods, and geographic regions. Table D.6 offers a breakdown by each respective stratification. The event types are restricted to only include violent activity to maintain conceptual consistency, and the time periods are aggregated to three-month units to increase variation in the event counts. The estimated counts reflect the median value of the posterior distribution, the standard deviations record the standardized variance from the MCMC posterior, and the confidence intervals are reported at the 95% level. The "estimated coverage" is calculated as the proportion of observed events over the estimated count: values closer to one hundred percent indicate low to no missingness, whereas values closer to zero percent indicate higher levels of missingness.[3]

[3]The estimated coverage is offered as point estimate but the range of these values fall along a distribution.

Table D.6 MSE Estimates by Strata for Nigeria 2011

Strata	Observed Counts	Estimated Counts	Std. Dev.	Confidence Intervals	Estimated Coverage
		Type			
Violent Attack	411	2,027.90	644.65	(1,221 - 3,617)	20%
Violence Against Civilians	17	281.20	106.79	(169 - 523)	59%
Riots	19	20.7	2.42	(19 - 27)	92%
		Time			
Jan-Mar	150	481.4	219.16	(261 - 1,026)	31%
Apr-June	173	563.6	163.64	(355 - 965)	31%
July-Sept	119	235.5	58.98	(162 - 379)	51%
Oct-Dec	154	410.2	114.34	(268 - 692)	38%
		Region			
North Central	168	622	187.03	(380 - 1,080)	27%
North East	254	448.5	84.63	(331 - 655)	57%
North West	78	212.6	64.94	(135 - 371)	37%
South East	10	20.9	11.63	(11 - 50)	48%
South South	66	295.7	132.17	(151 - 626)	22%
South West	20	57.5	32.7	(26 - 139)	35%
		Pooled			
Total	596	2,318	707.93	(1,491 - 4,078)	26%

Note: Counts estimated using the LCMCR package in R. See Manrique-Vallier, Price and Gohdes (2013)

The results demonstrate that non-homogeneous coverage of different types of events, time periods and/or regions affect the underlying estimates regarding the degree of missingness in the data. Given how the data is sliced, markedly different pictures emerge with respect to the potential "true" population. That said, these differences can help inform us about data sensitivity given the particular strata of interest, which could have important implications for quantitative analysis that seeks to explore sub-regional, temporally disaggregated, or concept-specific questions.

Specifically, when stratified by type, estimates regarding Riots appear largely complete. This makes sense as riots often indicate mass political instability and are more likely to be reported. Thus, coverage of this type of contentious activity is missed less often than violent events that are more heterogeneous in severity, locations, tactics and targets. This same logic is likewise reflected in the higher coverage rate for violence against civilians, as this form of contentious activity is more likely to be reported and picked up by mainstream news outlets and therefore more likely to be covered by the datasets.

When stratifying by time, most temporal windows reflect around the same degree of missingness; however, the coverage increases from July through September. The reason for this is unclear, as one would expect the period of April through June — which corresponds with

the timing of the national election in Nigeria — to be covered more readily. One potential explanation is that the subtype categories that are more likely to be picked up (Riots and Violence Against Civilians) occurred in greater frequency during this time period. These results demonstrate the degree to which heterogeneity in recapture rates can present different stories when stratifying conflict event data. Lastly, coverage appears to vary substantially when stratifying by region. The area with the highest coverage also corresponds with the rise of Boko Haram in the North East. Though rural, this area received greater coverage as the insurgency deepened, increasing the probability of an event being picked up in media reporting and eventually recorded in the datasets.

D.9 Selecting and Validating Alternative Cases

Case Selection

We locate cases where the data quality differs from that of Nigeria 2011 by generating a distribution for overlapping country-years using the geographic precision variable reported across all four of the datasets. For country-years where there was insufficient conflict activity to merit inclusion in the UCDP, no data is reported in the UCDP-GED. As such, all country-years are excluded where there is not mutual overlap in coverage of event activity (e.g. Libya 2013). The geo-precision variable captures the degree to which the geo-located coordinates reflect the actual location of an event: some entries in the data are *precise*, providing coordinates near or around the actual location of the event, and other entries are *imprecise*, providing coordinates at the centroid of the first administration unit or country. We broadly categorize the precision codes as precise/imprecise such that the proportion of precise entries is equal to one minus the proportion of imprecise entries.

Using the total proportion of precise events, we generated the empirical distribution for all overlapping country-years. We found that Nigeria 2011 was located in the middle of this distribution with roughly 83% of entries coded precisely. We then located the country-years at the tail ends of the distribution. On the lower end of the distribution, we identified South Sudan 2015, for which approximately 58% of entries are coded as precise. On the upper end, we identified Libya 2014, for which approximately 95% entries are coded as precise. These country-years provide two instances for which the data-generating process differs in

quality from that of Nigeria 2011. Thus, we leverage these two alternative cases to assess the performance sensitivity given differing levels of precision.

Validation

As noted in the main text, we established two spatial windows when validating the South Sudan and Libya entries: 3 km and 50 km. We integrated the events for each country separately and used the taxonomy assumptions outlined in the article and appendix.[4] The integration was then validated by two RA coders using the framework outlined in detail in Section B. The coders reviewed a validation sample constituting of 50% of located matches for the Libya 2014 3km/1day integration, and 25% of located matches for the Libya 2014 50 km/1 day integration. The differences in sample proportions were motivated by the overall number of matching entries. The objective was for coders to validate roughly equal representative sample sizes (in terms of the number of entries each reviewed). MELTT located a larger proportion of matching entries in the second integration given the expanded 50 km spatial window—correspondingly, only a smaller proportion was sampled. Likewise, for South Sudan we ultimately sampled 100% of matching entries for both spatial windows given the limited number of located matches. The results of the validation exercise are reported in the article.

[4]Note that the actor taxonomy was adjusted to accommodate the relevant set of actors for both country-years.

References

Ashlagi, Itai, Yash Kanoria, and Jacob D. Leshno. 2017. Unbalanced Random Matching Markets: The Stark Effect of Competition. *Journal of Political Economy* 125(1): 69–98.

Gale, David, and Lloyd S. Shapley. 1962. College Admissions and the Stability of Marriage. *The American Mathematical Monthly* 69(1): 9–15.

Manrique-Vallier, Daniel, Megan E. Price, and Anita Gohdes. 2013. Multiple Systems Estimation Techniques for Estimating Casualties in Armed Conflict. In *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict*, eds. Taylor B. Seybolt, Jay D. Aronson and Baruch Fischhoff. Oxford, UK: Oxford University Press.

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

Raleigh, Clionadh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing ACLED–Armed Conflict Location and Event Data. *Journal of Peace Research* 47(5): 651–660.

Salehyan, Idean, Cullen S. Hendrix, Jesse Hamner, Christina Case, Christopher Lineberger, Emily Stull, and Jennifer Williams. 2012. Social Conflict in Africa: A New Database. *International Interactions* 38(4): 503–511.

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