Article

Journal of Conflict Resolution
2015, Vol. 59(6) 1129-1149

DOI: 10.1177/0022002714530431
jcr.sagepub.com
SAGE

On the Accuracy of Media-based Conflict Event Data

Nils B. Weidmann[1]

Abstract

Empirical researchers of civil war rarely collect data on violence themselves and instead rely on other sources of information. One frequently used source is media reports, which serve as the basis for many ongoing data projects in the discipline. However, news reports rarely cover a conflict comprehensively and objectively and may therefore be prone to various reporting issues. This article provides an analysis of the accuracy of information given in news reports. In particular, if focuses on two types of "hard facts" that event data sets require: the location of an event and its severity. By linking media reports to firsthand accounts from a military database, the article does two things: (1) it analyzes the determinants of inaccuracy and confirms the expectation that events with a low number of observers tend to have higher reporting inaccuracies and (2) it assesses the magnitude of these inaccuracies and the implications for conducting empirical analyses with media-based event data.

Keywords

conflict event data, civil war, media bias, spatial analysis, data quality

One of the strongest trends in the study of civil war over the last couple of year has been the increased focus on microlevel analysis. By increasing the empirical resolution, researchers hope to be able to get closer to the micromechanisms responsible

[1]Department of Politics and Public Administration, University of Konstanz, Konstanz, Germany

Corresponding Author:
Nils B. Weidmann, Department of Politics and Public Administration, University of Konstanz, 78457 Konstanz, Germany.
Email: nils.weidmann@uni-konstanz.de

for the occurrence of political violence at the subnational level. Unavoidably, this trend has generated a need for better and more fine-grained data on violence. We can run cross-national comparisons by coding civil wars at the country/year level, but these data collections are useless if we are interested in the subnational patterns of violence. What we need instead are individual observations of violent incidents. That is what conflict event data sets provide.

How are such event databases created? Due to the difficulties and risks associated with data collection in conflict regions, this is a nontrivial question. In some instances, we can rely on reports from one of the actors involved, such as the military (Kalyvas and Kocher 2009; Berman, Shapiro, and Felter 2011) or the police (Sullivan 2013). In other cases, there exist extensive post-conflict data collections by third-party observers, such as nongovernmental organizations (NGOs; Lyall 2010; Nettelfield 2010). However, in most cases, these data collections are available for single conflicts only, and many other conflicts have no coverage at all. In order to be able to disaggregate to the local level, while at the same time being able to make comparisons across conflicts, we need a data source that is available more broadly across cases. This is why a lot of effort is being put into the development of event collections based on news reports. With the rapid flow of global news fueled by modern information and communication technology, news coverage of conflicts is increasing both in speed and in volume. Thus, news reports are a promising source of information also for scholarly work on violence.

Or are they not? The general criticism is that news reports do not objectively cover civil wars, but rather cater to a domestic audience. In other words, what is reported in the news, and how it is reported, may depend to a large extent on the people consuming these news. This insight is not new and has been shown to have strong effects (Galtung and Holmboe Ruge 1965; Davenport and Ball 2002; Davenport 2010). In particular, the two potential issues surrounding news reports are (1) the selection problem and (2) the veracity problem or description bias (Earl et al. 2004). The selection problem has to do with the fact that not all events that occur in a civil war show up in the international media. Rather, selection will be determined by supply-side factors (e.g., remoteness of an event may negatively impact selection into the news), but also demand-side factors (e.g., small-scale events may not be sensational enough to be reported). The veracity problem applies to events that are reported in the international news. Here, the question is whether the information provided about an incident in the news coverage is sufficiently detailed and accurate in order not to distort scientific results.

This article deals with the second problem. Veracity of reporting—or the lack of it—may have tremendous influence on our efforts to create comprehensive and accurate data collections on civil war. This may not be immediately obvious, since contemporary event data sets are not terribly demanding when it comes to the amount of information about an incident. Typically, what is included are the location and date of an event, the number of casualties, and the actors involved. All these are "hard" facts, which do not require much interpretation and are thus less likely to be

distorted by direct and indirect observers (Earl et al. 2004).[1] However, these seemingly ''hard'' facts are ''soft-reported'' by human observers and journalists until they eventually reach the final recipient, the news consumer. While we do expect some inaccuracies to be introduced during this process, is the resulting data quality still sufficiently high? In particular, can we detail the location of an event at a level precise enough to use it in our analysis of violence? Is the reported severity of an event anywhere close to the true death toll of an incident? What we need, and what has long been overdue, is an estimate of the error around the information provided in media-based conflict event data sets. This is what this article aims to provide.

In order to do this, this article takes advantage of a (rare) opportunity where in addition to a media-based conflict event data set, another event data set is available from a different source. Focusing on the ongoing conflict in Afghanistan, the data set used for comparison is based on military reports, so-called significant activities (SIGACTS).[2] Each of these SIGACTS events comes with short narrative, which permits matching a given event from a media-based event data set to the corresponding one in the military data set. Using the matched pairs of events, this article does two things. First, it tries to identify variation in the accuracy of reporting across events. In short, it is argued that the accuracy of media reporting should increase with the number of potential observers; thus, events in remote locations and more dangerous events should be surrounded by more uncertainty as regards the reported information. Second, the article aims to assess the overall magnitude of the differences we observe across data sets, in order to determine if media-based data sets are precise enough to be used in quantitative studies. The findings from this exercise are encouraging. At the same time, however, they highlight the limitations of media-based event data sets, in particular regarding the error in the reported location. Approaches using these data for fine-grained analyses (e.g., at the village level, or at the level of small grid cells) clearly overestimate the precision of locational information, and should therefore be avoided.

The article proceeds by discussing potential issues with media-based event data and derives expectations about the conditions affecting reporting accuracy. We then introduce the empirical approach and the data sets used in this study. The next section presents the results of the comparison, focusing first on the condition affecting quality and then on the overall magnitude of the inaccuracies found. The last section presents the conclusions and sketches avenues for future research.

Media-based Conflict Event Data

The recent years have experienced a surge of interest in the microanalysis of violent conflict, in particular civil war. The hope is that the disaggregation of the patterns and dynamics of violence can reveal the micromechanisms underlying these conflicts, which we fail to see when employing overly aggregated macroapproaches (Kalyvas 2008, 2012). These microlevel analyses typically study individual conflicts and focus on the variation of violence and its different forms across time, space, and actors. For

that reason, micro research is relatively demanding when it comes to the resolution of the data employed. Rather than coding start and end dates of civil wars at the country level—which is typically required for macroanalyses and can itself constitute a difficult undertaking—for micro studies we require information on individual incidents during civil war. This is why scholars have resorted to a different type of data collection: conflict event data sets.

A conflict event data set is a database that contains information about individual incidents in a conflict. To be sure, this type of data collection is not specific to the study of civil war but has been used in empirical investigations of related social phenomena. For example, Davenport (2010) gives an overview of the use of event collections in the study of contentious politics, where an "event" typically represents an instance of public protest or state repression. Event data sets have also been popular in the study of international crises and conflict (for a description, see e.g., Schrodt and Gerner 1994). Roughly speaking, these data sets include international events that can range between cooperative interaction and peaceful relations to outright war. The data sets this article deals with constitute a new generation of event collections. First, unlike their predecessors, they focus on violence in civil war, a topic that has become popular in recent years as already mentioned previously. Second, these data sets include precise information about the spatial location where an event happened, mostly in the form of spatial coordinates (longitude/latitude). The *Armed Conflict Location and Events Dataset* (Raleigh et al. 2010) was one of the first of these data sets, the Geo-referenced Event Dataset (GED) from the Uppsala Conflict Data Program (UCDP) is another more recent example (Sundberg and Melander 2013).

The addition of spatial coordinates to conflict event databases has enabled researchers to study a host of new questions and employ new approaches. First, the spatial dynamics of a conflict can themselves reveal a great deal about the actors and mechanisms involved. For example, it is now possible to analyze action–reaction dynamics between the government and a rebel group spatially, something that has so far been studied without taking into account the geographic dimension. Relatedly, event data sets can be used to look at the spatial diffusion of violence and the particular forms it takes (Schutte and Weidmann 2011). The second, and no less important, opportunity arises from the fact that events with spatial coordinates can be linked to other spatial data. By superimposing conflict events on spatial information of different kinds—for example, poverty, ethnicity, or accessibility—we can test a number of covariates of violence that we would otherwise have no access to. This strategy of data generation has been employed in a number of studies, for example, Buhaug and Rød (2006); Hegre, Østby, and Raleigh (2009); or Weidmann and Callen (2013).

There are different sources we can use for the creation of geospatial conflict event data sets, such as military and state records, or humanitarian organizations present in conflict regions. However, these sources are in most cases only available for a single conflict (or even only part of a conflict). Therefore, it is difficult to make

comparisons between conflicts, since the data collection mechanism differs radically. For that reason, developers of recent conflict event data sets have been resorting to media reports, just like their predecessors. Essentially, an event data collection is created by filtering out the relevant news reports for a particular conflict from a database such as Lexis-Nexis or Factiva, and hand coding the information contained in them (Öberg and Sollenberg 2011). In doing so, conflict event data sets are not very demanding as compared to earlier collections when it comes to the information required to code an event: the information extracted from the media reports is usually limited to "hard facts," such as the location of where an event occurred, the number of casualties (not distinguishing the type), and the actors involved. The limitation to these "hard facts" makes it possible to create comprehensive data sets with the same type of information across multiple countries and conflicts.

Despite the relatively sparse information contained in these event data sets, the reliance on media reports as only source of information may be problematic. Media reports are written with a specific audience in mind, so there is a risk that demand-side factors affect *what* is reported in the news and *how* it is reported. This problem has received extensive treatment in the sociology and political science literature. For example, Lichbach (1984) finds that media coverage of governability at the country level differs widely across different sources and world regions. Woolley (2000) provides a general overview of the use of media-based data in political science research. He cautions in particular against the uncritical use of event counts over time, since they can be significantly affected by reporting trends unrelated to the phenomenon under study. In the study of violent conflict, reporting issues have been central to the work of Patrick Ball and the Human Rights Data Analysis Group (HRDAG). The HRDAG specializes in the estimation of casualty counts in violent conflicts. In most cases, there is no single reliable source to report on conflict deaths; instead, Ball and collaborators use information from different sources to generate reliable estimates (see, e.g., Ball, Spirer, and Spirer 2000; Ball et al. 2003).

When it comes to media-based information, two problems have been identified (Earl et al. 2004): the problem of selection into the news media (which events are reported?) and the problem of the veracity of reporting (is the description of an event accurate?). The latter is the focus of this article. So far, there have been few (if any) attempts to apply the insights from the previous literature to more recent event collection efforts on violence in civil wars. Does the Rashomon Effect—the finding that different sources frequently tend to disagree in their reporting of an event (Davenport 2010)—also have major implications for event data collections on civil war? Granted, with their focus on "hard facts," these collections may not suffer from this problem to the same extent. As Earl et al. (2004) conclude, clearly verifiable information is less prone to biases in interpretation that are typically introduced during the reporting process.

Yet, this should not relieve us from the task of scrutinizing media-based event collections. Even the "hard facts" that these collections rely on could be subject to major inaccuracies. In the following, we focus on two of these "hard facts":

location of an event and number of casualties. The former is important as a key feature of the new generation of event data sets this article deals with. Geospatial conflict data set encode the spatial location of an event using *x*- and *y*-coordinates (longitude and latitude). Using this information, geographic information system (GIS) software can then be employed to analyze the patterns of violence or to generate other covariates. Casualty estimates are also an important variable in conflict event data sets, as they help us distinguish small skirmishes from major confrontations and thus provide a more nuanced picture of the dynamics of violence on the ground.

Both types of information may be subject to inaccuracies in reporting. These inaccuracies may stem from the fact that media reports rely on reports by (mostly civilian) people that have witnessed an incident. If the number of direct observers is low (or even zero), reporters rely on posterior information that only becomes available after an incident and cannot be verified. In general, we can assume that as the number of observers of an event decreases, the uncertainty surrounding an event report will increase. What does this mean for our two variables of interest, location and casualties? If a reporter can rely on a witness of an incident, it will be possible to obtain precise information of where an event occurred (e.g., close to a particular village). However, with few or no direct observers, reports will contain rough location descriptions, for example, referring only to the district where an event happened. The same applies to the number of casualties. Fewer or no direct observers mean that casualty figures will only be rough estimates and will be reported with a high level of error.

What determines how many potential observers an event can have? Two effects should be relevant here. The first is remoteness of the incident's location. If an event occurs far away from the nearest populated place, the number of potential observers is low, and it will be unlikely that a news report can rely on information from a direct observer of the report. If spatial information is available, it will be in the form of larger geographic or administrative units. This information is imprecise and should lead to a larger spatial error in an event data set once the event is assigned spatial coordinates. A similarly high reporting error should apply to casualty estimates that in remote regions are unlikely to come from direct observers. This is exactly the opposite in the proximity of larger settlements. Here, a high population density increases the probability of event reports becoming available from observers. Correspondingly, the locational information and the casualty estimates we get from media reports about these events should be much more accurate. Therefore, for the events in our data set this means that

> **Hypothesis 1:** Inaccuracies in the reported location and casualty number of an incident will be higher the greater the distance to the nearest populated place.

Location may not be the only determinant of inaccuracies in reporting, however. Some types of incidents are associated with a high danger and may thus suffer from

more inaccurate reporting, because there are few or no observers present. Others constitute by their nature less of a threat to observers once they have taken place. Battle encounters between armed forces of both sides are of the first type. They consist of the simultaneous use of force by two sides, where bystanders in the incident's proximity can easily be killed, especially if these battle events last for a longer time. As a result, these events should be more prone to inaccuracies in reporting, because it is unlikely that independent witnesses will be present, and reports will be based on hearsay and guesses. In contrast, one-sided attacks using, for example, improvised explosive devices (IEDs) do not carry the same risk to observers *after* they have occurred, since attackers are typically no longer present. This means that information can be provided by possibly more than one observer, who had the chance to witness the consequences of an IED attack directly. Therefore, while controlling for the effect of locational inaccuracies discussed previously, we should expect that

> **Hypothesis 2:** Inaccuracies in the reported location and casualty number of an incident will be greater for battle encounters than for one-sided attacks.

Does Inaccuracy Matter?

Even if the previously mentioned expectations bear out empirically, it is unclear to what extent the inaccuracies we find question the general validity and usability of media-based conflict event data sets. At least initially, there may be reason to worry. Civil wars, by their very nature, often occur in remote locations (Kalyvas 2007). If the previously discussed error in the spatial location of an event is significant, geographic event data could simply be too inaccurate to be used for spatial analyses of the type described previously. By the same token, information about casualties presented in media reports could be so far off that it is too noisy (or biased) to be used in quantitative analyses. Therefore, in addition to shedding more light on the reporting mechanism, we need to assess the magnitude of potential inaccuracies in media-based event data. Essentially, what are the error bands around the information we obtain from media reports? This is the second, and no less important, task to be addressed subsequently.

Empirical Approach

In order to find out whether reporting accuracy is subject to the influences hypothesized previously, this article presents an empirical analysis comparing two different sources of information on violence in a civil war. This section first describes the two data sources, explains the process of matching events to each other and introduces additional variables required for the analysis. The next section then uses the matched data set to examine the two hypotheses using regression analysis.

Matching Media-based Conflict Reports with Military Records

We assess the accuracy of conflict event data by linking event reports from a media-based event data set to the corresponding ones from military records. The latter are rarely available for research purposes, which limits this analysis to the conflict in Afghanistan for the years 2008 and 2009. The media-based event data used here are drawn from the GED, created and published by the UCDP at the Department of Peace and Conflict Research at Uppsala University. Note that the UCDP GED is not generally based on media reports alone; instead, it also draws on other sources such as NGO reports. However, the Afghanistan coding is entirely based on global media sources such as AP, AFP and Reuters, and thus serves as valid approximation of the violence we can observe through media sources. A complete and fully documented prerelease of the Afghanistan data was kindly made available for the purpose of this research. The data set includes individual observations of civil war violence in Afghanistan along with the spatial and temporal coordinates of the event. The included events are those between organized rebel groups (in Afghanistan, the Taliban) and military forces fighting on behalf of the government. In Afghanistan, the latter include the Afghan National Army and the Afghan National Police, but also foreign forces such as the US military and International Security Assistance Force (ISAF) Coalition forces. Only events in which at least one death occurred are included in the data set. The data for 2008 to 2009 consist of 2,027 observations.

The military data set used for comparison is the SIGACTS database for Afghanistan. This database collects all information on SIGACTS from US and coalition forces in a standardized format, which is used across many deployments in contemporary conflicts such as Afghanistan and Iraq (Berman, Shapiro, and Felter 2011). A SIGACT can be anything from minor events such as detainee transfers or observations of (nonviolent) insurgent presence, to major lethal incidents such as shootings or IED explosions. For that reason, the 2008 to 2009 coverage is very comprehensive, with 42,398 events for the two years, many of which, however, did not involve lethal violence. Each event is coded with date and time, as well as precise geographic coordinates. The latter is important for our purpose; since these coordinates are measured with GPS technology, their precision is—not surprisingly—several orders of magnitude higher than what media-based event data can achieve. Each event is assigned a different category. The three major categories are confrontations initiated by the insurgents (Enemy Actions), events initiated by coalition forces (Friendly Actions), and "Explosive Hazards," which constitute mostly IED attacks. While the first two are two-sided confrontations (direct violence), the latter is a category of one-sided violence where the initiator is usually not present anymore at the time the device goes off. The events also come with a narrative of what happened during the reported incident, as well as casualty numbers, distinguishing between fatalities among foreign and Afghan forces, enemy (Taliban) fighters, and civilians. Since the latter were not consistently filled in the data set, we screened the narratives for

information about casualties and updated the casualty columns accordingly.[3] The SIGACTS version is not complete in the sense that it gives a full picture of the events on the ground (Carpenter, Fuller, and Roberts 2013), and we do not expect it to be. Upon closer inspection, it becomes obvious that the database fails to include, for example, many US- or coalition-initiated events, such as targeted raids or airstrikes.

Important for our purpose, the database was not created for public reporting about military activity on the ground. Rather, it represents a data collection effort internal to the military, undertaken for reporting and assessment inside the organization and subject to internal standards of accuracy. This alleviates major concerns about biased reporting on the side of the military: since the data were not collected to scrutinize military actions in the public sphere, there is a lower probability that, for example, absolute numbers of casualties are systematically underreported. Nevertheless, there could be a tendency for the *type* of casualty to be biased in one direction. For example, it may be that if in doubt about the identity of a person killed in fight, soldiers tend to assume that it is a member of the Taliban rather than a civilian. However, even if this was the case, this would likely not affect our analysis, since media-based data sets typically include only the total number of people killed in an incident rather than the type of casualties. Potential biases in the reporting of an event's location are even less likely to occur in the SIGACTS data set, as there is no incentive to misreport the location of an incident. Also, military units are usually tracked by GPS, so the location of an incident is recorded using electronic equipment rather than manual entry.

Matching of the two data sets was done by going through the UCDP GED and retrieving the media report upon which the coding is based. The matching event in the SIGACTS database was then located by using event date and location as a starting point and then gradually increasing the spatial and temporal search radius until the corresponding event was found. In order for a match to be established, the information in the news report and the SIGACTS narrative had to agree on the nature of the event. As mentioned previously, due to the incompleteness of the SIGACTS database, we do not expect to find matches for each and every event included in the UCDP GED. Of the 2,027 GED events for 2008 to 2009, we were able to match slightly more than half ($N = 1{,}077$; 53 percent). This clearly indicates that both data sets overlap only to a certain extent. However, since in this article we are not interested in the different selection mechanisms into each of the data sets but rather the accuracy of reporting, this is less relevant for the purpose of our study. For our comparison, all we need is a large enough sample of *matched* events, which is what the previously mentioned procedure generated. A short example helps illustrate the type of information contained in the data sets, as well as the matching procedure.

Table 1 presents a typical match of two incidents from the UCDP and SIGACTS data sets. The left column shows the UCDP coding of a Reuters report from May 23, 2008. According to the report, on that day a suicide attack

Table 1. Comparison of UCDP GED and SIGACTS Reports of a Suicide Bomb in Khost Province on May 23, 2008.

Data set	UCDP GED	SIGACTS
Identifier	AFG-2008-1-327-210	166506D5-EF51-B533
Date	May 23, 2008	May 23, 2008
Location	Khost (Matun) district Khost province	Mando Zayi district Khost province
Casualties	6	7
Report	Khost, Afghanistan, May 23 (Reuters) A suspected Taliban suicide bomber killed one child and four Afghan soldiers in an attack targeting an army convoy on Friday, a provincial governor's spokesman said. [. . .] Five others, including four Afghan soldiers, were wounded in the latest attack, which came in the eastern province of Khost, bordering Pakistan, the spokesman Khaiber Pashtun said.	UNIT: ANA TYPE: SUICIDE BOMBER TIMELINE: PCC REPORTS THAT ANA STRUCK A IED AT WB 7855 8769 AT 0430z, AWAITING BDA UPDATE: AT 0520 WAS NOT A IED IT WAS A SUICIDE BOMBER WITH A VEST UPDATE: 0526 2/D/2-506 REPORTS THAT THE ANA STOPPED AT WB 7855 8769, A SUICIDE BOMBER; PUSHED A CHILD IN FRONT OF THE CONVOY TO GET IT TO STOP OR SLOW DOWN (UNCONFIRMED) WALKED UP TO THIER VEH, AND DETONATED HIMSELF [. . .] SUMMARY: 7×ANA KIA 3×ANA WOUNDED TAKEN TO SAL 3× CIV LN WIA,: (2× KIDS, 1× ELDER) TAKEN TO KHOWST HOSPITAL

Note: ANA = Afghan National Army; BDA = battle damage assessment; IED = improvised explosive devices; VEH = vehicle.

targeting a convoy occurred in Khost Province. The report mentions that in addition to the attacker, four Afghan soldiers and a child died as a result of the attack, bringing the death toll up to six. The report also mentions the source of information for the report, in the case the administration of Khost Province where the attack happened. The military report has many more details about the incident, but also shows the process of updating the report as more information is obtained. It starts with a specification of the military unit involved (Afghan National Army [ANA]) and the type of incident. The first report on the incident describes it as an IED attack and mentions that further information on damage (battle damage assessment [BDA]) is expected. This update arrives about an hour later and mentions that the incident was a suicide attack. Also, a more detailed description of the incident is given. The report ends with a summary

of the casualties: seven members of the ANA were killed (killed in action [KIA]), three wounded (wounded in action [WIA]), and another three civilians were also wounded.

Overall, the reports agree to a large extent. Date and province correspond. However, the UCDP GED places the event in the district around the provincial capital, Khost (Matun) district. The correct location as given in the SIGACTS report is in the nearby Mando Zayi district, about ten kilometers from where the UCDP GED places the event. There is a small difference in the number of casualties. Whereas the media reports talks about six casualties (including the attacker), the SIGACTS report puts this number at seven (excluding the attacker). Remarkably, according to the military report, the child was not killed in the incident, as the media report claims. We do not know, however, if this difference is due to the media coverage making the event sound more sensational than it actually is, or the military's omission to accurately record the casualties. In sum, we see slight differences in reporting, both as regards the location and the number of casualties. The quantitative analysis presented in the next section will help us determine how these inaccuracies vary with remoteness of location and type of events and, ultimately, what the magnitude of these inaccuracies is.

Variables

From the matched pairs of events described in the previous section, we create a set of variables for the statistical analysis presented subsequently. The dependent variables, inaccuracy in location and casualty estimates, are computed as follows. The first one is simply the (logged) distance between the actual location of the incident as given by the SIGACTS and the location that was assigned to the event by the UCDP GED—in other words, the spatial error in the media-based conflict data set. Inaccuracy in casualty estimates is computed by comparing the GED's casualty numbers to those from the SIGACTS data set. It is important to note that the GED does not only provide a single estimate of this number (called the best estimate in the data set) but also an interval of this estimate, ranging from a *low* to a *high* number. This interval is included to capture the uncertainty in some media reports regarding casualties. Using this interval, we create a new variable, ''casualties misreported,'' which takes a value of one if the SIGACTS casualty count falls outside the *low–high* range.

There are two types of explanatory variables we need. First, the remoteness of an event is measured in two ways: first, by its (logged) distance from the nearest major populated place, and second, by the population density at the incident location. The former is measured both for towns (population of at least 5,000) and cities (population of at least 25,000), relying on a data set of Afghan settlements released by the Central Statistical Office and revised by Jason Lyall. According to the first hypothesis, this distance should have a positive effect on each of the two dependent variables. Alternatively, we also use population density for operationalizing

Table 2. Linear Regression Results. Dependent Variable: Logged Distance between Reported and Actual Location of an Event.

	Model 1	Model 2	Model 3
Distance nearest town (log)	0.41*** (0.03)		
Distance nearest city (log)		0.39*** (0.04)	
Population density (log)			−0.19*** (0.02)
Explosive hazard	−0.24** (0.09)	−0.27** (0.09)	−0.33*** (0.09)
(Intercept)	8.47*** (0.20)	8.39*** (0.21)	10.32*** (0.24)
N	1,077	1,077	1,077
BIC	4,256.50	4,333.03	4,323.61
log L	−1,681.41	−1,719.67	−1,714.96

Note: Province fixed effects not shown. BIC = Bayesian Information Criterion.
*Significant at $p < .10$. **$p < .05$. ***$p < .01$.

remoteness, using the spatial population raster data set LandScan (Oak Ridge National Laboratory 2008). Fewer people at the incident location should be related to higher inaccuracies, so we expect population density to have a negative effect. Second, we use the event category from the SIGACTS data set, which distinguishes between indirect violence (Explosive Hazards) and battle events (Enemy Action and Friendly Action). Using the former as the baseline, we expect the accuracy of reporting to be higher (i.e., lower error) for events of indirect violence.

Results

The following sections report on the results of our analyses with the matched pairs of events from the two data sets. First, we present regression results showing whether our theoretical expectations about the accuracy of reporting in conflict regions bear out empirically. Second, we conduct a descriptive investigation of the magnitude of the inaccuracies found in media-based event data set, in order to address potential concerns about the general usability of these data sets in empirical investigations.

Explaining Inaccuracy

Using regression analysis, we test whether the GED's inaccuracies in the spatial location and the number of casualties are systematically related to the remoteness of the incident's location and its type. We use linear regression with spatial error as the dependent variable (models 1–3), and logit models with the "casualties misreported" variable (models 4–6). All models include province fixed effects (not shown) to net out potential effects of uneven reporting across the country. Table 2 presents the results.

Models 1 through 3 provide support for Hypothesis 1. The first two models show that distance to the nearest major settlement (our first indicator for remoteness) is

Table 3. Logit Regression Results. Dependent Variable: Casualties Misreported (0/1).

	Model 4	Model 5	Model 6
Distance nearest town (log)	0.01 (0.05)		
Distance nearest city (log)		−0.04 (0.06)	
Population density (log)			−0.00 (0.03)
Explosive hazard	−0.68*** (0.14)	−0.70*** (0.14)	−0.69*** (0.14)
(Intercept)	0.44 (0.32)	0.52 (0.32)	0.49 (0.37)
N	1,077	1,077	1,077
BIC	2,125.91	2,125.44	2,125.96
Log L	−616.11	−615.88	−616.14

Note: Province fixed effects not shown. BIC = Bayesian Information Criterion.
*Significant at $p < .10$. **$p < .05$. ***$p < .01$.

positively related to the spatial error in the GED. This effect applies independently of what size of settlement we look at (towns or cities). In other words, the farther away from the nearest major settlement an incident is located, the higher the spatial error: roughly tripling the distance to the nearest town or city increases the spatial error by about 1.5 kilometers. Model 3 uses a different operationalization, population density, as the independent variable. According to the results, a higher population density decreases the error, which also supports Hypothesis 1. Tripling population density roughly decreases the spatial error by about 1.2 kilometers. Models 1 through 3 also support Hypothesis 2, since events of indirect violence (Explosive Hazard) exhibit a significantly lower spatial error than direct encounters (the baseline category). This difference is relatively large, reducing the error by about 30 percent. In sum, the results provide evidence that supports the impact of the number of potential observers on reporting accuracy: if location and type of event permit more observers to witness an incident, the error in the reported location is significantly lower.

Models 4 through 6 (Table 3) use logistic regression with a dummy for misreported casualties as the dependent variable. Again, different operationalizations of remoteness as well as event type are used as independent variables. Surprisingly, remoteness seems to have no effect on the quality of casualty numbers in event reports: none of the three variables has a significant effect in the expected direction. However, Explosive Hazard events are again subject to significantly lower errors in the reporting of casualties, as we expected.[4] Why would remoteness of an event affect the reported location, but not the accuracy of casualty numbers? One reason could be that the casualty number, in contrast to location, is politically sensitive information. So when journalists report about an event, they may actively seek confirmation of this sensitive information from other sources (including the military), but not for relatively unimportant information (location). This is why remoteness may affect locational accuracy (as we have shown previously), but not the accuracy of casualty reports.[5]

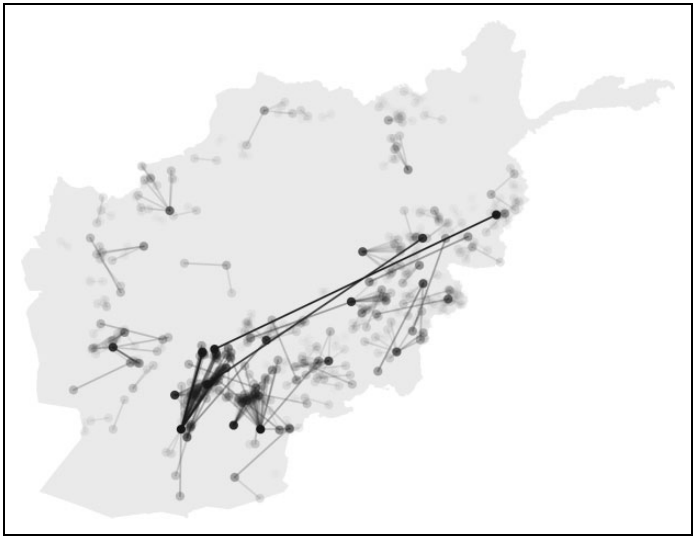

Figure 1. Spatial error map for georeferenced conflict events.
Note: The lines connect referenced and actual locations. Color intensity is scaled by length, such that longer lines (larger error) are more visible.

The Margins of Error in Media-based Data Sets

In the previous section, we have shown how remoteness and the type of an incident affect the accuracy of media reports, and we have seen that systematic effects do exist. However, the overarching question is how the identified variation in accuracy affects the usability of media-based event data sets for quantitative analyses. As argued previously, event data set generation based on news reports is a powerful and frequently employed approach. In the light of the previously mentioned findings, is this approach viable? This section uses the matched sample of 1,077 events from the UCDP GED and the SIGACTS data set to find out. Our comparison starts with the first measure of error introduced previously, the spatial error between the reported and the actual location of an event. Is the information provided in news reports sufficiently accurate to attach precise spatial coordinates to an event? What is the magnitude of error we should expect? Figure 1 shows an error map of the matched event pairs in the data set. Each line connects the location that the UCDP GED assigns to an event to the actual location obtained from the SIGACTS data set. The lines are shaded such that longer ones appear better visible.

At first glance, the map draws our attention to the fact that for some events, the spatial error in their location is considerable. The long lines connect locations in completely different parts of the country, in one case up to a distance of more than 600 kilometers apart. However, the map does not reveal the fact that the great majority of lines is actually very short and does not really show up on the map. Therefore,

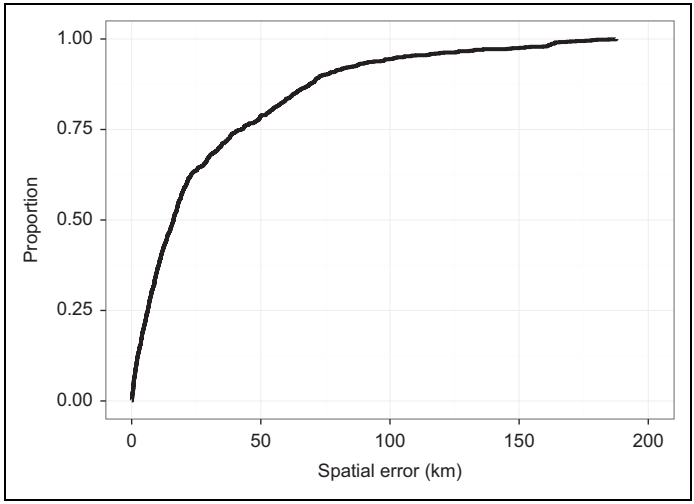

Figure 2. Cumulative distribution of the spatial error in the media-based event data set.

a plot of the distribution of the spatial error (the distance between the referenced and the actual location) is more revealing. Figure 2 shows the cumulative distribution of the spatial error in the data set of matched pairs.

Figure 2 presents a much less concerning picture of the spatial error in the media-based data set. About 80 percent of the events are located within 50 kilometers of the where they actually occurred, which roughly corresponds to the diameter of the average district in Afghanistan (area of about 2,000 sq. km). Thus, we note that the spatial error of media-based event data sets is roughly in the same order of magnitude as the size of a district. In other words, we should expect the accuracy of media-based event data to be approximately correct when it comes to identifying the district where an event occurred, but not more precise than this. An additional test confirms this. If we compare the administrative district an incident actually occurred in to the one where the media-based data set puts it, we find that slightly more than 80 percent are referenced to the correct (63.4 percent) or one of the neighboring districts (18.6 percent). At a slightly larger scale, the UCDP GED is almost perfect when it comes to locating events in the correct province: in 98.14 percent of all matched pairs, the actual and referenced provinces correspond.

Based on the previously mentioned results, it seems that we can trust media-based event data sets approximately down to the level of districts, but not below. Thus, spatial analyses using these data should take this account by limiting the spatial resolution at which they operate. Using the district level seems to be within a reasonable range (Weidmann and Ward 2010). Also, spatial approaches that use artificially created cells (a so-called fishnet) should not attempt to go below a size of about 50 kilometers by 50 kilometers when using media-based event data. Many of these fishnets employ this resolution or stay well below (Buhaug and Rød 2006;

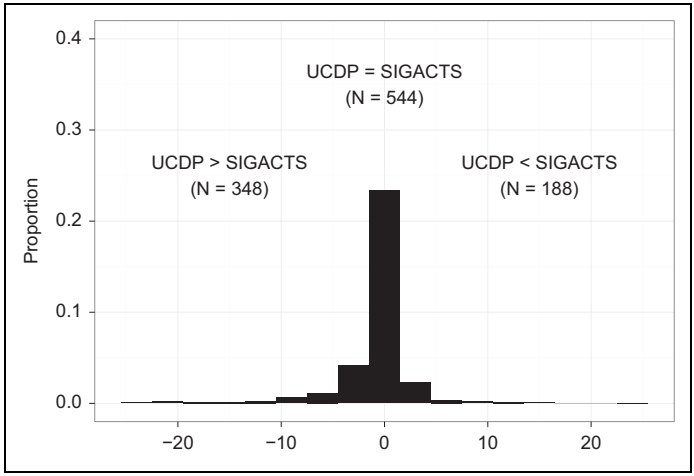

Figure 3. Comparison of casualty numbers from the UCDP GED and the SIGACTS data set (left area of the histogram: UCDP reports a higher casualty number than the SIGACTS, center: the numbers correspond, right: UCDP reports a lower casualty number).
Note: UCDP = Uppsala Conflict Data Program; GED = geo-referenced event data set; SIGACTS = significant activities.

Tollefsen, Strand, and Buhaug 2012), leaving us with a high level of confidence that an event referenced to a cell actually occurred there. Others, however, seem to be too fine-grained in comparison to the spatial error inherent in media-based data sets. For example, a cell size of 8.5 kilometers by 8.5 kilometers is probably too small (Hegre, Østby, and Raleigh 2009; Raleigh and Hegre 2009): according to Figure 2, less than 30 percent of all events will be referenced to the cell in which they actually happened. In sum, these results can alleviate potential concerns to a great extent: if we take into account that media-based event data come with a certain amount of locational uncertainty—which is roughly within a range of 50 kilometers—the spatial information is precise enough for most analyses.

The Severity of Events

A second possible concern discussed previously is the distortion of casualty numbers in different reports about a single event. Using the matched pairs of events from the UCDP GED and the SIGACTS, we can find out to what extent these potential issues arise. Similar to the approach introduced previously, casualty numbers are coded as "corresponding" if the SIGACTS casualty count is within the *low–high* interval given by the GED, as an "underestimation" if the SIGACTS count is higher than the UCDP's *high* estimate, and as an "overestimation" if the SIGACTS count is lower than the GED's *low* estimate.

Figure 3 displays a histogram of the differences in the casualty numbers.[6] In about 50 percent of all cases ($N = 544$), the casualty estimates correspond, such that the "true" casualty number reported in the SIGACTS data set is within the interval given in the GED.[7] We also see that there are a number of under- and overestimations. As we would have expected, there is a slight tendency for the media-based data set to report higher casualty estimates, but it is not very pronounced in the light of the high number of correspondences we find. However, Figure 3 not only shows the direction of reporting, that is, whether the GED reports the same casualty numbers as the SIGACTS or whether it over or underreports. But, what is the magnitude of misreporting? Are the underreported number off by a much larger amount as the overreported ones? A quick look at Figure 3 does not confirm this suspicion. The magnitude of misreporting (the difference between media-reported and military estimates) is roughly the same, regardless of whether we see under- or overreporting: in both cases, the difference has a median of two, and roughly the same average (4.33 and 3.01).

In sum, our results can also alleviate some concerns about casualty reporting at the event level across different sources. We see an almost perfect correspondence in about half of all cases, and if deviations exists, they are mostly small. Also, we see no clear trend that media-based sources consistently report higher casualty numbers than the military. While there is a slightly higher number of cases with overreporting, the magnitude of this effect remains moderate and is no different from the error in the opposite direction. However, these results are based on a subset of cases that exist in both data sets. Therefore, what our results suggest is that casualty numbers in media reports do not seem to be too far off *if an event is reported*. We *cannot* conclude that media reports pick up most of the violence, and can therefore serve as reliable sources to estimate overall conflict severity. If we do not have a data set with comprehensive coverage—as may be the case for the majority of conflicts—selective inclusion of casualties across data sets remains a major problem (Ball, Spirer, and Spirer 2000).

Conclusion

This article has attempted to achieve two goals. First, it aims to identify systematic variation in the quality of media reporting on conflict. Focusing on "hard facts" of an incident, it argues that remoteness and type of an event affect the accuracy of reporting. Second, it tries to assess the implications this has for the creation of event data collections based on news reports. Since a number of ongoing projects rely on this data source, it is necessary to scrutinize the potential problems that could arise from using it. The empirical analysis relies on matching event reports from a media-based data set to one based on military records. In the sample of matched events, differences across sources were computed both as regards the location of an incident and the number of casualties. Statistical analysis largely supports not only the expectation that remoteness negatively affects accuracy, but also that incidents of

indirect violence are subject to a lower error in reporting as compared to those of direct violence. However, additional analyses on the magnitude of these errors show that it remains within reasonable limits. Locational information in media reports is accurate roughly at the level of districts, but not below. Casualty numbers are reported with a high level of precision, and under- versus overreporting is roughly balanced and low in magnitude. In sum, this suggests that media-based event collections can serve as valuable bases for empirical research, if these limits of precision are taken into account.

The availability of both a media- and a military-based event data set makes the Afghanistan conflict one of the few opportunities for a data set comparison such as the one presented here. In fact, Afghanistan may even serve to establish a lower boundary on the error present in the former: due to the large presence of international forces and the continuously high levels of violence, the conflict has been in the spotlight for many years. Also, the apt use of international media by all sides in the conflict means that what is portrayed in the media may be an unusually accurate picture of the conflict, as compared to other wars with much less international presence and therefore, media coverage. If media reporting in Afghanistan is, on average, better than for other conflicts, we should assume that reporting inaccuracies in other conflicts are of similar magnitude or higher.

Nevertheless, there are several points of caution. First, as mentioned previously, our analysis applies to "hard facts," that is, relatively well-defined characteristics of violent incidents. Reporting biases may be much more severe if we are interested in less clearly recognizable features, such as the initiator of an event or the nature of violence used. For these characteristics, information contained in media reports may either suffer from considerable interpretation on the reporter's side or may simply be too sparse to let the coder verify that a particular characteristic applies or not. The usual advice given under these circumstance—to triangulate different sources (Davenport 2010)—is particularly difficult to follow in the context of civil war, where different reports are oftentimes impossible to obtain. Therefore, the limitation to "hard facts" employed by many event data sets is reasonable. Second, no source is without problems, which certainly applies to the SIGACTS used as reference category in this analysis. As argued previously, the SIGACTS database was never designed for public distribution, so the incentives for biased reporting in favor of the military may be low. Still, bias could exist in particular when it comes to casualty estimates, but should be low (or almost zero) when it comes to the reported location of an event. For that reason, the results with respect to location may stand on firmer ground than results on casualty numbers. Finally, our analysis is based on two years' worth of coverage for a single conflict, which may raise concerns about generalizability. However, the quality of casualty and location reporting given in news reports should hardly differ across cases, which should give us some confidence that the findings hold more generally. Still, limitations could arise from our use of a military data set for comparison. Therefore, alternative validations of this kind should be conducted using firsthand accounts from other sources.

While encouraging for the development of media-based event collections, more research will have to be done into the quality of media-based event data. First and foremost, research has to address the selection problem: what is reported in the news, and what is left out? While we have shown that if an incident is reported, the quality of this reporting is, on average, surprisingly good, there may still be many events that are simply left out of the picture. Thus, future research has to determine how events are selected into the news and what effects this selection may have on the results we derive from media-based event data. However, with more and more attempts to assess data quality in conflict research, this important effort should soon be underway.

Author's Note

Jason Lyall kindly gave permission for his updated Afghanistan settlements data set to be used for the purpose of this analysis. Replication data are available from http://thedata.harvard.edu/dvn/dv/nilsw.

Acknowledgment

Thanks to Magnus Öberg, Jesse Hammond, and participants at the "Ethnicity and Conflict" Workshop at Uppsala for comments and suggestions and to Sabine Otto for excellent coding assistance. The author gratefully acknowledges support from the UCDP by making an early version of their Geo-referenced Event Data set (GED) available to the author.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is supported by the Alexander von Humboldt Foundation (Sofja Kovalevskaja Award) and by the EU FP7 Marie Curie Zukunftskolleg Incoming Fellowship Program, University of Konstanz (grant no. 291784).

Notes

1. By "hard facts," I mean those that are less subject to interpretation by an observer.
2. The data set was released by WikiLeaks as "Afghanistan War Diary."
3. While declassified versions of the SIGACTS database exist, they do not include the incident narratives, which are central to this project. This is why we resort to the SIGACTS version published by WikiLeaks in June 2010.
4. An alternative explanation for this finding could be that Explosive Hazard events involve more civilian casualties, which may raise media attention. However, this suspicion does not bear out empirically; including a dummy for $\geq$1 civilian casualties does not alter the results of models 4 through 6.
5. The lack of variation in remoteness due to clustering of matched events in urban areas is not an explanation for the lack of an effect. Matched events vary considerably as regards

remoteness; the median matched event is located about eleven kilometer away from the nearest town, and almost fifty kilometer away from the nearest city.

6. The difference is zero if the SIGACT count c(SIGACTS) is within the [low, high] interval given by the GED; c(SIGACTS) − low if c(SIGACTS) < low, and c(SIGACTS) − high if c(SIGACTS) > high. Three outliers truncated to improve graphical presentation.
7. In 456 of the 544 events, the UCDP's point estimate (*best*) corresponds perfectly to the SIGACTS casualty estimate.

References

Ball, Patrick, Jana Asher, David Sulmont, and Daniel Manrique. 2003. "How Many Peruvians Have Died? An Estimate of the Total Number of Victims Killed or Disappeared in the Armed Internal Conflict Between 1980 and 2000." *Report, American Association for the Advancement of Science (AAAS)*. Available at https://www.hrdag.org/wp-content/uploads/2013/02/aaas_peru_5.pdf.

Ball, Patrick, Herbert F. Spirer, and Louise Spirer, eds. 2000. *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis*. Washington, DC: American Association for the Advancement of Science (AAAS) Science and Human Rights Program.

Berman, Eli, Jacob N. Shapiro, and Joseph H. Felter. 2011. "Can Hearts and Minds Be Bought? The Economics of Counterinsurgency in Iraq." *Journal of Political Economy* 119 (4): 766-819.

Buhaug, Halvard, and Jan Ketil Rød. 2006. "Local Determinants of African Civil Wars, 1970-2001." *Political Geography* 25 (3): 315-35.

Carpenter, Dustin, Tova Fuller, and Les Roberts. 2013. "WikiLeaks and Iraq Body Count: The Sum of Parts May Not Add Up to the Whole–A Comparison of Two Tallies of Iraqi Civilian Deaths." *Prehospital and Disaster Medicine* 28 (3): 1-7.

Davenport, Christian. 2010. *Media Bias, Perspective, and State Repression: The Black Panther Party*. New York: Cambridge University Press.

Davenport, Christian, and Patrick Ball. 2002. "Views to A Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977-1995." *Journal of Conflict Resolution* 46 (3): 427-50.

Earl, Jennifer, Andrew Martin, John D. McCarthy, and Sarah A. Soule. 2004. "The Use of Newspaper Data in the Study of Collective Action." *Annual Review of Sociology* 30:65-80.

Galtung, Johan, and Mari Holmboe Ruge. 1965. "The Structure of Foreign News: The Presentation of the Congo, Cuba and Cyprus Crises in Four Norwegian Newspapers." *Journal of Peace Research* 2 (1): 64-91.

Hegre, Håvard, Gudrun Østby, and Clionadh Raleigh. 2009. "Poverty and Civil War Events: A Disaggregated Study of Liberia." *Journal of Peace Research* 53 (4): 598-623.

Kalyvas, Stathis. 2007. "Civil Wars." In *The Oxford Handbook of Comparative Politics*, edited by Carles Boix and Susan C. Stokes, 416-34. Oxford, UK: Oxford University Press.

Kalyvas, Stathis N. 2008. "Promises and Ptifalls of an Emerging Research Program: The Microdynamics of Civil War." In *Order, Conflict and Violence*, edited by Stathis Kalyvas, Ian Shapiro, and Tarek Masoud, 397-421. Cambridge, UK: Cambridge University Press.

Kalyvas, Stathis N. 2012. "Micro-level Studies of Violence in Civil War: Refining and Extending the Control-collaboration Model." *Terrorism and Political Violence* 24 (4): 658-68.

Kalyvas, Stathis N., and Matthew A. Kocher. 2009. "The Dynamics of Violence in Vietnam: An Analysis of the Hamlet Evaluation System (HES)." *Journal of Peace Research* 46 (3): 335-55.

Lichbach, Mark Irving. 1984. "The International News about Governability: A Comparison of the New York Times and Six News Wires." *International Interactions* 10 (3-4): 311-40.

Lyall, Jason. 2010. "Are Coethnics More Effective Counterinsurgents? Evidence from the Second Chechen War." *American Political Science Review* 104 (1): 1-20.

Nettelfield, Lara J. 2010. "Research and Repercussions of Death Tolls: The Case of the Bosnian Book of the Dead." In *Sex, Drugs and Body Counts: The Politics of Numbers in Global Crime and Conflict*, edited by Peter Andreas and Kelly Greenhill, 159-87. Ithaca, NY: Cornell University Press.

Oak Ridge National Laboratory. 2008. "LandScan Global Population Database." Electronic Resource. http://www.ornl.gov/landscan/. Accessed Feb 22, 2010.

Öberg, Magnus, and Margareta Sollenberg. 2011. "Gathering Conflict Information Using News Resources." In *Understanding Peace Research: Methods and Challenges*, edited by Kristine Höglund and Magnus Öberg, 47-73. New York: Routledge.

Raleigh, Clionadh, and Håvard Hegre. 2009. "Population Size, Concentration and Civil War: A Geographically Disaggregated Analysis." *Political Geography* 28 (4): 224-38.

Raleigh, Clionadh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset." *Journal of Peace Research* 47 (5): 651-60.

Schrodt, Philip A., and Deborah J. Gerner. 1994. "Validity Assessment of a Machine-coded Event Data Set for the Middle East, 1982-92." *American Journal of Political Science* 38 (3): 825-54.

Schutte, Sebastian, and Nils B. Weidmann. 2011. "Diffusion Patterns of Violence in Civil Wars." *Political Geography* 30 (3): 143-52.

Sullivan, Christopher. 2013. *Undermining Resistance: Mobilization, Repression, and the Enforcement of Political Order*. Working paper, Department of Political Science, University of Michigan, Ann Arbor.

Sundberg, Ralph, and Erik Melander. 2013. "Introducing the UCDP Georeferenced Event Dataset." *Journal of Peace Research* 50 (4): 523-32.

Tollefsen, Andreas Forø, Håvard Strand, and Halvard Buhaug. 2012. "PRIO-GRID: A Unified Spatial Data Structure." *Journal of Peace Research* 49 (2): 363-74.

Weidmann, Nils B., and Michael Callen. 2013. "Violence and Election Fraud: Evidence from Afghanistan." *British Journal of Political Science* 43 (1): 53-75.

Weidmann, Nils B., and Michael D. Ward. 2010. "Predicting Conflict in Space and Time." *Journal of Conflict Resolution* 54 (6): 883-901.

Woolley, John T. 2000. "Using Media-based Data in Studies of Politics." *American Journal of Political Science* 44 (1): 156-73.