Known unknowns: media bias in the reporting of political violence

Nick Dietrich & Kristine Eck

To cite this article: Nick Dietrich & Kristine Eck (2020) Known unknowns: media bias in the reporting of political violence, *International Interactions*, 46:6, 1043-1060, DOI: [10.1080/03050629.2020.1814758](https://doi.org/10.1080/03050629.2020.1814758)

To link to this article: <https://doi.org/10.1080/03050629.2020.1814758>

View supplementary material [↗](#)

Published online: 26 Oct 2020.

Submit your article to this journal [↗](#)

Article views: 1303

View related articles [↗](#)

View Crossmark data [↗](#)

Citing articles: 3 View citing articles [↗](#)

Known unknowns: media bias in the reporting of political violence

Nick Dietrich ^a and Kristine Eck ^b

^aOhio Wesleyan University; ^bUppsala University

ABSTRACT

How does sourcing affect which events are included in international relations datasets? The increasing number of machine-coded datasets offers the promise of coding a larger corpus of documents more quickly, but existing automated processes rely exclusively on databases of news reports for coverage. We exploit source variation in the UCDP GED dataset, which includes events from media reports and non-media sources, to explore the bias introduced by including only media reports in international relations datasets. Unlike previous studies, our approach allows us to compare subnational and cross-national determinants of bias. We find that media sources severely underreport events in African countries, and coverage is also associated with country-level factors like international trade and subnational factors like access to communication technology. Non-media sources cover a significant number of events not included in media sources; their inclusion can expand coverage and reduce bias in datasets.

KEYWORDS

Political violence; media and politics; conflict; data

¿Cómo la causa afecta los acontecimientos que se incluyen en los conjuntos de datos de relaciones internacionales? La creciente cantidad de conjuntos de datos codificados por máquinas garantiza la codificación de un mayor corpus de documentos con mayor rapidez, pero los procesos automatizados existentes dependen exclusivamente de bases de datos de noticias para la cobertura. Utilizamos la variación de la fuente en el conjunto de datos de acontecimientos georreferenciados (Georeferenced Event Dataset, GED) del Programa de Datos sobre Conflictos de Uppsala (Uppsala Conflict Data Program, UCDP), que incluye acontecimientos de los informes de los medios de comunicación y fuentes que no pertenecen a los medios de comunicación, para analizar el sesgo que surge al incluir solo los informes de los medios de comunicación en los conjuntos de datos de relaciones internacionales. A diferencia de estudios anteriores, nuestro enfoque nos permite comparar los determinantes de sesgo a nivel subnacional y transnacional. Las fuentes de los medios de comunicación informan muy poco sobre los acontecimientos en los países africanos, y la cobertura también está asociada a factores nacionales, como el comercio internacional y factores subnacionales, como el acceso a la

CONTACT Nick Dietrich dietrich.nicholas@gmail.com Visiting Assistant Professor of Data Analytics, Ohio Wesleyan University, Delaware, OH, USA

Supplemental data for this article can be accessed on the [publisher's website](#).

© 2020 Taylor & Francis Group, LLC

tecnología de la comunicación. Las fuentes que no pertenecen a los medios de comunicación cubren un número significativo de acontecimientos que no se incluyen en las fuentes de los medios de comunicación; su inclusión puede ampliar la cobertura y reducir el sesgo en los conjuntos de datos.

Comment le sourçage affecte-t-il les événements inclus dans les jeux de données portant sur les relations internationales ? Le nombre croissant de jeux de données codés par machine offre la promesse de coder plus rapidement un plus grand corpus documentaire, mais les processus automatisés existants s'appuient exclusivement leur couverture sur des bases de données de reportages d'actualité. Nous exploitons la variation des sources du jeu de données GED de l'UCDP, qui comprend des événements provenant de reportages des médias et de sources non médiatiques, afin d'explorer le biais introduit en n'incluant que les reportages des médias dans les jeux de données portant sur les relations internationales. Contrairement aux études précédentes, notre approche nous permet de comparer les déterminants infranationaux et transnationaux du biais. Nous constatons que les sources médiatiques ne couvrent pas suffisamment les événements intervenant dans les pays africains et que cette couverture est également associée à des facteurs nationaux comme le commerce international et à des facteurs infranationaux comme l'accès aux technologies de communication. Les sources non médiatiques couvrent un nombre important d'événements non inclus dans les sources médiatiques, et leur inclusion peut élargir la couverture et réduire la partialité des jeux de données.

Introduction

Are news articles accurate sources of information about political events, and, if not, how can researchers improve the validity of their data? International relations researchers are increasingly using news articles to generate data, often using machines to process large quantities of information. There are significant advantages to this approach: machines can process the large corpus of news articles and can do so in near-real-time. Projects like Phoenix (Brandt et al. 2019) and TERRIER (Irvine et al. 2019) provide researchers with large samples of a wide variety of political actions. Parsing techniques make it possible to identify relevant actors and actions in international relations and create new automated data collection projects (Schrodt 2012; Solaimani et al. 2017). These techniques hold promise for real-time forecasting, and researchers have made progress predicting political phenomena like international and regional conflict (Brandt, Freeman, and Schrodt 2011; Hegre et al. 2019) and irregular leadership change (Ward and Beger 2017). Validation of these automated coding systems has found that they can be just as effective as human coders at extracting information from news databases (Bagozzi et al. 2018; Best, Carpino, and Crescenzi 2013).

Automated coding techniques gather a corpus of media reports from news databases, perform functions to identify actors and actions in those reports, and output a dataset of events. This research note addresses the first step of the process: gathering a corpus of articles for coding political events. The exclusive reliance on media reports differs from human-coded datasets like the Uppsala Conflict Data Program (UCDP) political violence datasets. UCDP codes from news reports and then supplements these news stories with reports from non-media sources. In the following sections, we review prior findings about bias in media reporting and list factors that may affect which events are reported. We then leverage UCDP's coding process to analyze the set of events that is included *only* in non-media sources, offering some insight into source bias in media-coded event data. The advantage of our approach, unlike previous studies of media bias, is that we can examine both subnational and cross-national media bias. Our findings suggest that media bias varies significantly by country, with African countries particularly likely to be underrepresented in media sources.

The Data Generation Process

How Events Become News

The study of international relations has turned increasingly to events-based data; this is particularly true of the sub-field interested in organized violence. Major data collection projects, such as the Uppsala Conflict Data Program (UCDP), the Armed Conflict Location Event Data (ACLED), and the Social Conflict Analysis Database (SCAD), have responded to the need for spatially and temporally disaggregated data by releasing the component events data. While UCDP, ACLED, and SCAD are manually coded, projects like TERRIER (Irvine et al. 2019) and Phoenix (Brandt et al. 2019) have attempted to automate this process and code additional-directed actions like protests and provisions of aid. These automated processes take a large number of news stories directly from a news database like Lexis-Nexis, apply parsing techniques (for example, Schrodtt 2012), and output a dataset of directed actions.

These projects recognize that they rely on publicly available sources that cannot comprehensively capture all political violence. Numerous post-conflict truth and reconciliation efforts have illustrated the reporting challenges posed by conflict; some percentage of events which take place are never publicly reported, and only careful field research – oftentimes of a forensic nature – can address this gap. What becomes news (Galtung and Ruge 1965) is a known unknown in this sense: we can be confident that not all conflict events are made public, but we cannot know what percentage is reported, nor do we know the processes which determine whether an event becomes public. As

a consequence, no media-based event dataset on armed conflict captures the true population of conflict events.

News media reports provide an indispensable, but incomplete, resource for identifying and documenting conflict since they provide the only viable basis for annual, global coverage. Media sources, however, cannot provide a complete accounting of political violence. International news sources may miss events because they lack coverage of certain areas, while local news might not achieve widespread circulation or may be subject to censorship or violence that interferes with reporting. Even in combination, local and international sources of news cannot provide a complete picture because they have limited resources and therefore make strategic decisions about allocation of coverage.

Take, for example, the Djugu region of DRC. Violence broke out at the end of 2017 and continued into 2018. Data from UCDP indicate that news media reported on 13 unique events which contained information about 134 deaths, while additional non-media sources reported on 118 events which involved 698 deaths. In this case, non-media sources were essential for gaining a more complete picture of the violence that occurred. In other cases, non-media sources provide little, if any, new information. What conditions explain when media coverage of political violence is more or less complete?

Researchers have reduced underreporting bias by using information from multiple media sources (Bagozzi et al. 2018) or using machine coding to expand the corpus of source articles for historical data (Althaus et al. 2019). These approaches do not, however, correct for systematic bias in which events are likely to be underreported in the first place. Previous research has found that media reporting of violence is positively correlated with the presence of communications technology (Croicu and Kreutz 2017; Weidmann 2016) and regime type (Baum and Zhukov 2015), while findings for press freedom (Drakos and Gofas 2006; Urlacher 2009), severity of violence (Croicu and Kreutz 2017; Davenport and Ball 2002; Price and Ball 2014) and urban areas (Croicu and Kreutz 2017; Davenport and Ball 2002) are mixed. Media reporting also varies by country: African countries are often excluded from international relations research because of cross-national differences in information availability (Lemke 2003, 120–124).¹

Media sources share journalistic norms and profit incentives when deciding what to cover. Non-media sources, however, work under a different incentive structure that is unlikely to be driven by either journalistic norms or an event's journalistic market value. Non-media sources constitute an eclectic mix of governments, international or regional organizations, and local organizations, all of which operate with different mandates and agendas. Davenport and Ball (2002) conduct one of the few studies on the data generation process of non-

¹Multiple systems estimation offers an alternative to media-based reporting (Lum, Price, and Banks 2013) but is only feasible for limited empirical domains (i.e. country or subnational levels). It is therefore not an option for researchers interested in large cross-national analyses.

media sources, finding that human rights organizations and victim interviews differed from newspapers in their coverage of violence in Guatemala. A subset of these sources are human rights organizations that operate with an agenda to change state behavior via “naming and shaming” (Davis, Murdie, and Steinmetz 2012; Hafner-Burton 2008). These organizations cover events that are not covered in media sources, often strategically selecting underreported events to fill in gaps in media coverage.

Measuring Bias

Scholars have sought to gain empirical leverage over the study of conflict reporting bias in three ways. First, they have regressed the volume or nature of media coverage in a country with various attributes of that country (Baum and Zhukov 2015; Drakos and Gofas 2006; Urlacher 2009). This approach is similar to subnational research on social movements and media reporting that regresses the coverage by a certain news outlet with attributes of that outlet or the event (c.f. Davenport 2009); Oliver and Myers 1999).

Second, scholars have compared media coverage in a narrowly defined time and place (usually a single country) with another non-media source. These studies assume that the other source provides ‘ground truth’ data which can be used as a reference point for analyzing which events are not covered by the media. This is the current state of the art; prominent examples include Weidmann (2016), who uses US military’s Significant Activities (SIGACTS) data for Afghanistan and Davenport and Ball (2002), who use data generated by local human rights organizations and witness testimony in Guatemala.

Finally, a third approach exploits information gathered in the coding process to speak to the issue of data generation. We are aware of only one study which takes this tack: Croicu and Kreutz (2017) use the UCDP’s precision codes for the date of an event to proxy the quality of media-reported conflict data. Their assumption is that more detail provided by the source implies more information about an event and therefore better data.

Our empirical strategy adapts elements of the second and third approaches: we leverage information contained in non-media sources to inform us about media reporting. We do so using coding source information made public by the UCDP Georeferenced Event Dataset (GED) dataset (Sundberg and Melander 2013). The novelty of this approach is that it facilitates cross-national analysis. Most similar efforts (e.g. Davenport and Ball 2002; Loyle, Sullivan, and Davenport 2014; Weidmann 2016) use rich data from a single country. This approach is internally valid, but it is unclear the extent to which results from settings like Afghanistan or Guatemala travel to other contexts. Moreover, single-country analysis precludes modeling factors which vary at the national level, such as regime type or trade, and which may factor into calculations concerning journalistic coverage. Because we use global geocoded

events data in our analysis, we are able to include both national and subnational predictors.

Empirical Strategy

The departure point for our analysis is to exploit one of UCDP's coding practices to provide us with information. Since it streamlined its coding practices (beginning with 2013 data), the UCDP has consistently begun its annual coding with media sources.² Searches are done using the Factiva database to generate a body of news reports for coding; this process has been described extensively elsewhere (Eck and Hultman 2007; Sundberg, Eck, and Kreutz 2012; Sundberg and Melander 2013; UCDP 2018). Note that these news sources include both international news bureaus (like Reuters, Agence France Presse, or Xinhua) and local sources (many of which are translated from local languages and reproduced by the BBC Monitoring service).³

Coders determine whether an article contains an event which meets its criteria or whether it is unclear and warrants additional searches in both media and non-media sources. Thereafter, the UCDP turns to non-media sources which it has identified as high-value. Some sources are checked systematically every year – such as International Crisis Group, Amnesty International, and Human Rights Watch – while other reports are located using keyword searches. If an event is identified in a non-media outlet which has already been coded from a media source, then it is the media source attributed in the UCDP GED data. About 20% of events in UCDP GED are coded using non-media sources.

Because of UCDP's coding procedures, we can be sure that events which are attributed to non-media sources were not covered by the media in the news report corpus downloaded from Factiva.⁴ We exploit this variation in order to provide us with information about events which did not receive media attention.

An important caveat is in order. The non-media source data do not constitute population data.⁵ Non-media sources must also make strategic decisions on how to allocate resources, decisions which are shaped by their mandates.

²This is true with two exceptions: India and Syria. UCDP has determined that the non-media sources used in coding those countries — the South Asian Terrorism Portal and the Syrian Observatory for Human Rights, respectively — surpass media coverage in their detail and scope and therefore code these non-media sources first. These two countries are dropped from our analysis.

³We address the issue of source language in the online appendix.

⁴We are unable to make the opposite conclusion; events with only media sources may or may not also be reported in non-media sources. If an event has already been coded with a media source, no further action is taken if the same event is found in a non-media source.

⁵This problem is not unique to our data; previous research has also struggled with using non-media sources as a benchmark because they do not constitute true population data. Davenport and Ball (2002) is one of the few sources to acknowledge that coverage of conflict events by non-media sources is nonrandom. SIGACTS, for instance, includes events declassified by the U.S. Department of Defense, but not all events are equally likely to be declassified; it would be reasonable to expect that covert actions and special operations are likely to be classified at a higher level.

The premise of our analysis is not that we are able to provide a comprehensive analysis of media reporting compared to an ideal ‘ground truth.’ We instead compare the media-only sample of events against the expanded sample that includes non-media sources. We cannot infer that the expanded sample is the population of all events, but we are certain that the expanded sample is more complete than the media-only sample because the media-only sample is a proper subset of the expanded sample. Events that are included in the expanded sample but not the media-only sample – that is, events that were only reported by non-media sources – provide information about the kinds of events that media sources are likely to miss. Our results indicate associations between independent variables and the probability of exclusion from media sources.

This design still leaves the possibility that there are events reported in no source, but is capable of identifying systematic differences in events that were covered by media sources versus other kinds of documentation. The value of this approach is twofold. First, it is useful from a data collection perspective to know how the choice of source information affects the violence events that are collected. Data projects that rely exclusively on news sources may be systematically missing events which have been reported elsewhere. Second, this approach provides some substantive insight into news reporting on political violence. For example, this approach can indicate whether bias is primarily associated with differences between individual events or broader differences in the media landscape between countries.

Data

We analyze events from the UCDP GED (Croicu and Sundberg 2017; Sundberg and Melander 2013). We use only events from 2013 to 2016 because of changes in UCDP coding procedure instituted in 2014 (which began with the coding of the 2013 events).⁶ We use this subset because source information for this period is streamlined and reliably reported in a consistent format and because we want to ensure that the process used to collect all events was the same. We cannot be certain that events collected before 2013 used the same systematic coding procedure.

We analyze GED event sourcing using random forests.⁷ We chose this method because it is capable of using information from a large number of covariates, which is particularly important for the large number of logistical, political, and economic factors that can affect media’s ability to report on events. Random forests focus on bivariate prediction, which means that we

⁶This is a sample of 25,180 events.

⁷Some readers may prefer a less complicated method. Although we believe that random forests are appropriate, we validate the findings of the random forests with a logistic regression in the online appendix. The results show significant differences at the 95% level for 32 of the 40 variables we tested and affirm the primary findings of the random forests.

need not have a fully specified model of the data generating process in order to correctly estimate relationships between the independent and dependent variables. This is critical for our analysis because political violence reporting is complicated and we are not confident that we can create a theoretically derived model for every aspect of the process. This technique has been used in other political violence applications to compare the correlates of repression (Hill and Jones 2014).

Random forests are predictive learning algorithms that aggregate classification trees to rank the predictive importance of covariates. Classification trees are created by randomly selecting variables at each node and choosing the variable split that best partitions the data. Random forests use a large number of classification trees to determine the aggregate classification importance of the predictors (see Breiman 2001; Hastie, Tibshirani, and Friedman 2009, 587–603). Random forests have a number of desirable properties: they are resistant to overfitting (Breiman 2001, 7), are capable of reliably measuring predictive importance even when covariates differ in scale (Strobl et al. 2007) or contain missing values (Hapfelmeier et al. 2014), and the substantive associations of individual variables can be evaluated using partial dependence plots (Jones and Linder 2016).

The overall importance of variables is determined by randomly allocating observations at nodes of the variable of interest and comparing the reduction in classification accuracy. We construct the forests using conditional inference trees to avoid bias introduced by covariates of different scales, as described by Strobl et al. (2007). This is necessary because international relations data vary dramatically in scale, including continuous variables (like GDP per capita), binary indicators (like whether violence targeted civilians), and count variables (like the number of fatalities in a given event).

Outcomes

The dependent variable for our analysis is the sourcing of GED events, specifically, whether a given event is attributed to a media source. Our random forests classify events as either having or not having a media source. We compare the covariates using permutation importance measures to identify the relative strength of predictors. We use a variable importance measure that accounts for missing data as described by Hapfelmeier et al. (2014).⁸

⁸We use this measure instead of the Strobl et al. (2008) conditional variable importance score for two reasons. First, patterns of missing data are a known problem in international relations (Lemke 2003, 120) and many of our covariates have a significant number of missing values. These missing values are unlikely to be missing completely at random, making complete case analysis inappropriate. Even if values were missing at random conditional on the other covariates, the Hapfelmeier scores would be more accurate than scores obtained from imputation because they account for the prevalence of missing data (Hapfelmeier et al. 2014, 21). Second, we prefer the Hapfelmeier et al. scores because our analysis is intended to be descriptive. Our goal is to inform practitioners about events most associated with media bias. Conditioning on other variables would change the interpretation of these scores and likely result in misspecification because we do not have a theoretically-derived model of the causal process.

The overall importance of variables is determined by making the variable of interest uninformative and measuring the reduction in classification accuracy. We construct the forests using conditional inference trees to avoid bias introduced by covariates of different scales, as described by Strobl et al. (2007). We also use partial dependence plots to visualize the substantive associations between covariates and media source classification (Jones and Linder 2016).

Predictors

We include a variety of predictors that previous research and common sense suggest may have the potential to influence media reporting on events. These are predictors that affect either (1) the logical difficulty of reporting a particular event (for example, location or censorship) or (2) the level of media interest in the event (for example, the type of violence or the country's international trade).

First, we use several event-level predictors from the UCDP GED dataset (Sundberg and Melander 2013). We include *estimated fatalities*, *type of violence* (whether the event is a state-based event, non-state event, or one-sided violence event),⁹ the *year* that the event took place, and the *region* of the world where the event occurred.

We incorporate a number of event-level variables from the PRIO-GRID subnational dataset (Tollefsen, Strand, and Buhaug 2012), including *distance to the capital*, *distance to the nearest international border*, *mountainous terrain*, *infant mortality*, *subnational population*, *subnational economic output*, and the *number of excluded groups*.

We also include access to communication technology using the geolocated position of IP addresses (see Croicu and Kreutz 2017), calculating the *distance to nearest IP6 node* for each event in our analysis using the GeoLite2 IP address database (Maxmind 2018). The distance to IP nodes, while approximate, is a good proxy for access to communication technology at a given location.

We also use several national-level indicators, including economic indicators from the (World Bank Development Indicators 2015). We include *international trade* as a percentage of GDP to capture the level of involvement of a state in the international system. This may be important for international demand for news about a country, and can also serve as an indicator of the country's openness to international reporters. We also include *GDP per capita* as an indicator of a country's economic performance, which may affect local media markets and incentives to produce news about political violence events.

National-level conflict indicators are included based on the UCDP-PRIO Armed Conflict Dataset (Gleditsch et al. 2002; Pettersson and Eck 2018). We include variables for *international conflict*, *civil conflict*, and *internationalized*

⁹For information on these categories, see <http://pcr.uu.se/research/ucdp/definitions/>

civil conflict. These variables are counts of the number of such conflicts ongoing in a given country-year. Ongoing conflict may affect journalists' ability to access information safely. It may also affect the newsworthiness of information about political violence, either by making violence routine or by affecting demand for news about a country.

We include a number of relevant media indicators from the V-Dem Project (Coppedge et al. 2018). These include *media censorship*, *internet censorship*, *media diversity*, *self-censorship*, *media bias*, *internet access*, *media criticism*, *media corruption*, and *journalist harassment*. Each of these variables is measured as a latent variable according to the V-Dem aggregation methodology. For comparison, we also include the Freedom House media freedom measure (Freedom House 2017). Conditions for media operations in a country have the potential to affect the ease of accessing information and the cost of reporting for journalists.

Governance variables from the Polity project (Marshall, Gurr, and Jaggers 2016) are also incorporated. These are components of the frequently used democracy and autocracy indices and measure several facets of a country's domestic political institutions. We include the *competitiveness of participation*, *regulated participation*, *competitiveness of executive recruitment*, *regulations for executive recruitment*, and *constraints on the executive*. In addition to the components of the polity index, we include the aggregate *polity* score frequently used in international relations research as an indicator of democracy. We also include an additional country-level indicator for legal right to information (Global Right to Information Rating 2018). The right to information captures de jure right to request information from the government, and does not highly correlate with other domestic political institutions.

Results

Figure 1 shows the variable importance scores for the predictors described in the previous section. Variable importance permutation scores indicate the change in the predictive accuracy of the random forests when the variable of interest cannot be used to classify the outcome. While these scores do not have intuitive interpretations in an absolute sense, they can be interpreted relative to one another: a higher variable importance score indicates that the variable is more important for predicting the outcome.

The variable importance scores in Figure 1 demonstrate that location is an important predictor of differences in reporting. In particular, whether an event took place in Africa is the most informative factor for whether that event was reported in media sources. Other regional variables are of moderately high importance. The distance from an internet node is another informative

Variable importance scores indicate the mean importance of the variable over 200 random forests. Confidence bars indicate that the importance score was within the range in 95% of random forests.

Figure 1. Media source variable importance (Random forest of GED events, 2013–2016).

predictor of differential reporting on an event, corroborating previous findings that the availability of communication technology influences the ability to report on events (Weidmann 2016).

Beyond the location of events, political context is associated with media reporting. International trade is a strong predictor of media reporting, as is the level of participation in the political system. These variables capture both international and domestic aspects of politics that have the potential to affect reporting.

A final category of important predictors contains media-specific variables. Several media freedom indicators, including V-dem’s media censorship and media corruption and Freedom House’s media freedom index, are associated with media reporting. Whether an event is a one-sided violence event is also informative for media coverage.

In order to visualize the substantive relationships between the predictors and media sourcing, we use partial dependence plots. Partial dependence plots indicate the predicted probability that an event has a media source for values of the predictor. Partial dependence plots, unlike classical regression methods, are capable of detecting nonlinear relationships. Note that because upward of 80% of cases in the GED have a media source, the predicted probability will generally be centered at upwards of 80%.

The partial dependence plots for the regional variables are shown in Figure 2. These results demonstrate that events occurring in Africa are less likely to be reported in media sources compared to non-media sources in our sample. These results comport with existing research demonstrating that data from Africa are more likely to be missing and likely to be of lesser quality when not missing (Lemke 2003, 120–124). A problem with reporting on Africa relates to the overall lack of coverage; with relatively few media stories filed in Africa, UCDP is left with too little and often too imprecise information to include events without additional details gleaned from non-media sources. In the online appendix, we discuss this problem in greater detail and illustrate the lack of news coverage in African conflict countries compared to other conflict regions. We also provide a number of examples of the different kinds of texts coders encounter across media and non-media sources.

Given that the most important variables in our analysis are country- or region-level variables, we show the proportion of media-reported events by country in Figure 3. This figure shows the proportion of events with a media source for every country with more than 100 events between 2013 and 2016

Points indicate the mean increase in the probability that an event occurring in a region has a media source over 200 random forests. Confidence bars indicate that the difference was within the range in 95% of random forests.

Figure 2. Media source partial dependence (Random forest of GED events, 2013–2016).

Proportion of events with at least one media source. Countries with fewer than 100 events are excluded.

Figure 3. Proportion of media-sourced events by country, 2013–2016.

in our sample. The list of countries with the most non-media reported events validates the algorithm's reliance on region. Most of the countries with a small proportion of media-reported events are on the African continent, with the notable exceptions of Mexico and Myanmar. That the majority of events in some conflicts were never reported in media sources underscores the importance of non-media sources for building a complete picture of conflict events.

Figure 4 shows the effects of international trade, distance from an internet node, and polity. Although some of polity's component indicators have relatively high variable importance scores, polity itself is a largely uninformative measure when it comes to reporting biases. States with very low international trade are less likely to be covered in the media sample compared to the non-media sample, but the relationship plateaus once a particular threshold is reached. Finally, events that are close to an internet node are more likely to be reported in media sources compared to non-media sources in our sample.

Figure 5 shows the effects of media-specific variables and one-sided violence events on media reporting. One-sided violence events are less likely to be reported by media sources in our sample, probably because these events are more difficult to observe and are more likely to draw investigations by NGOs

Figure 4. Media source partial dependence (Random forest of GED events, 2013–2016).

Figure 5. Media source partial dependence (Random forest of GED events, 2013–2016).

or international organizations. Media censorship, according to both the V-dem and Freedom House scores, is associated with a lower chance of media reporting on events. A higher score on the Freedom House measure indicates less media freedom while a higher score on the V-dem measure indicates more media freedom. In V-dem's media corruption score, a higher score indicates less corruption. Surprisingly, more corrupt media are associated with a higher chance of media reporting. This is likely because media corruption implies that there is a functioning media in the country that operates with government support and has the resources to report on events. Although our analysis examines only whether events are reported at all, media characteristics could lead to differences in accuracy or tone of coverage.

Concluding Discussion

What do these results tell us about reporting of political violence events? First, media coverage of events varies with location. Coverage of violence events is especially poor in Africa compared to other regions. Subnationally, the proximity to communication technology is highly associated with media coverage of political violence events. These results make clear that media coverage is not uniform across space, but highly dependent on demand for and access to information. Previous studies of media bias have drawn similar conclusions within conflicts (Croicu and Kreutz 2017; Weidmann 2015), but our results suggest that future work should pay more attention to cross-national differences in reporting quality.

Our findings cannot speak to the broader question of how many conflict events go completely undocumented and are therefore missing from conflict datasets (Price and Ball 2014), and as a result the full extent of the measurement error is not quantifiable (Gohdes and Price 2013).¹⁰ But they do suggest that relying exclusively on media reports excludes information correlated with politically important factors, and is especially likely to exclude information in African countries.

Although researchers are working on statistically modeling the reporting process (Cook et al. 2017), there is currently no single-shot scientific solution to correct for different sources of reporting bias. One promising approach is exploiting differences in individual reports (Cook and Weidmann 2019); our results suggest that using reports from different sources could be useful for these bias-correction approaches. We find that the majority of events in some conflicts were reported only in non-media sources, meaning that these sources can fill in some of the gaps in media coverage.

As machine-coded data become more common in International Relations, it is possible to code from a larger set of sources in a reasonable timeframe. Increasing the number of sources has the potential to reduce bias in the resulting data, but only if researchers adapt these approaches to include non-media sources. The tools for analyzing text describing political events already exist; now researchers need only expand the corpus used as input for these approaches beyond the standard news database APIs. Including non-media sources in international relations datasets will reduce bias, yield more new events per source, and result in a more complete and accurate picture of political phenomena than using media sources alone.

Acknowledgments

We would like to thank Mihai Croicu, Anita Gohdes, Nils Metternich, Nils Weidmann, and Yuri Zhukov for excellent advice. Replication materials are available on Dataverse at <http://dvn>.

¹⁰Our intuition, however, is that the risk set for this uncertainty is circumscribed to conflict zones: we have high confidence that most countries truly have no conflict events (e.g. Sweden, Norway), but in countries with ongoing armed conflict there is uncertainty regarding the scope of reporting accuracy.

iq.harvard.edu/dvn/dv/internationalinteractions. Please direct questions about replication to Nick Dietrich at dietrich.nicholas@gmail.com.

ORCID

Nick Dietrich <http://orcid.org/0000-0001-8380-0203>

Kristine Eck <http://orcid.org/0000-0002-4998-7964>

References

- Althaus, Scott, Joseph Bajjalieh, John F. Carter, Buddy Peyton, and Dan A. Shalmon. 2019. "Cline Center Historical Phoenix Event Data Variable Descriptions." Cline Center for Advanced Social Research. v.1.2.0. University of Illinois Urbana-Champaign.
- Bagozzi, Benjamin E., Patrick T. Brandt, John R. Freeman, Jennifer S. Holmes, Alisha Kim, Agustin Palao Mendizabal, and Carly Potz-Nielsen. 2018. "The Prevalence and Severity of Underreporting Bias in Machine- and Human-Coded Data." *Political Science Research and Methods* 7 (3):641–49. doi:10.1017/psrm.2018.11.
- Baum, Matthew A., and Yuri M. Zhukov. 2015. "Filtering Revolution: Reporting Bias in International Newspaper Coverage of the Libyan Civil War." *Journal of Peace Research* 52 (3):384–400. doi:10.1177/0022343314554791.
- Best, Rebecca H., Christine Carpino, and Mark J.C. Crescenzi. 2013. "An Analysis of the TABARI Coding System." *Conflict Management and Peace Science* 30 (4):335–48. doi:10.1177/0738894213491176.
- Brandt, Patrick T., John R. Freeman, and Philip A. Schrodtt. 2011. "Real Time, Time Series Forecasting of Inter- and Intra-state Political Conflict." *Conflict Management and Peace Science* 28 (1):41–64. doi:10.1177/0738894210388125.
- Brandt, Patrick T., Jennifer S. Holmes, Latifur R. Khan, Ng Vincent, Benjamin Bagozzi, John R. Vito D'Orazio, Jill Irvine Freeman, Javier Osorio, and Philip A. Schrodtt. 2019. "Phoenix Real-Time Event Data." <http://eventdata.utdallas.edu>. Accessed 1 May 2019.
- Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1):5–32. doi:10.1023/A:1010933404324.
- Cook, Scott J., Betsabe Blas, Raymond J. Carroll, and Samiran Sinha. 2017. "Two Wrongs Make a Right: Addressing Underreporting in Binary Data from Multiple Sources." *Political Analysis* 25 (2):223–40. doi:10.1017/pan.2016.13.
- Cook, Scott J., and Nils B. Weidmann. 2019. "Lost in Aggregation: Improving Event Analysis with Report-Level Data." *American Journal of Political Science* 63 (1):250–64. doi:10.1111/ajps.12398.
- Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, et al. 2018. "V-Dem Codebook V8." Varieties of Democracy (V-Dem) Project.
- Croicu, Mihai, and Joakim Kreutz. 2017. "Communication Technology and Reports on Political Violence: Cross-national Evidence Using African Events Data." *Political Research Quarterly* 70 (1):19–31. doi:10.1177/1065912916670272.
- Croicu, Mihai, and Ralph Sundberg. 2017. "UCDP Georeferenced Event Dataset Codebook Version 17.1." Department of Peace and Conflict Research, Uppsala University.
- Davenport, Christian. 2009. *Media Bias, Perspective, and State Repression: The Black Panther Party*. Cambridge: 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. doi:10.1177/0022002702046003005.
- Davis, David R., Amanda Murdie, and Coty Garnett Steinmetz. 2012. "Makers and Shapers': Human Rights INGOs and Public Opinion." *Human Rights Quarterly* 34 (1):199–224. doi:10.1353/hrq.2012.0016.
- Drakos, Konstantinos, and Andreas Gofas. 2006. "The Devil You Know but are Afraid to Face: Underreporting Bias and Its Distorting Effects on the Study of Terrorism." *Journal of Conflict Resolution* 50 (5):714–35. doi:10.1177/0022002706291051.
- Eck, Kristine, and Lisa Hultman. 2007. "One-sided Violence against Civilians in War: Insights from New Fatality Data." *Journal of Peace Research* 44 (2):233–46. doi:10.1177/0022343307075124.
- Freedom House. 2017. "Freedom of the Press 2017 Methodology." <https://freedomhouse.org/report/freedom-press-2017-methodology>. Accessed 29 September 2017.
- 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):64–90. doi:10.1177/002234336500200104.
- Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand. 2002. "Armed Conflict 1946 to 2001: A New Dataset." *Journal of Peace Research* 39 (5):615–37. doi:10.1177/0022343302039005007.
- Global Right to Information Rating. 2018. "Global Right to Information Rating." <http://www.rti-rating.org>. Accessed 25 July 2018 .
- Gohdes, Anita, and Megan Price. 2013. "First Things First: Assessing Data Quality before Model Quality." *Journal of Conflict Resolution* 57 (6):1090–108. doi:10.1177/0022002712459708.
- Hafner-Burton, Emilie M. 2008. "Sticks and Stones: Naming and Shaming the Human Rights Enforcement Problem." *International Organization* 62 (4):689–716. doi:10.1017/S0020818308080247.
- Hapfelmeier, Alexander, Torsten Hothorn, Kurt Ulm, and Carolin Strobl. 2014. "A New Variable Importance Measure for Random Forests with Missing Data." *Statistics and Computing* 24 (1):21–34. doi:10.1007/s11222-012-9349-1.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*, 2nd ed. New York: Springer.
- Hegre, Håvard, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyle, Lisa Hultman, Stina Höglbladh, Remco Jansen, et al. 2019. "ViEWS: A Political Violence Early-warning System." *Journal of Peace Research* 56 (2):155–74. doi:10.1177/0022343319823860.
- Hill, Daniel W., Jr., and Zachary M Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *The American Political Science Review* 108 (3):661–87. doi:10.1017/S0003055414000306.
- Irvine, Jill, Christian Grant, Andrew Halterman, Khaled Jabr, and Yan Liang. 2019. "TERRIER (Temporally Extended, Regular, Reproducible International Event Records)." <http://terrierdata.org>. Accessed 1 May 2019.
- Jones, Zachary, and Fridolin Linder. 2016. "Edarf: Exploratory Data Analysis Using Random Forests." *Journal of Open Source Software* 1 (6):1–4. doi:10.21105/joss.00092.
- Lemke, Douglas. 2003. "Review: African Lessons for International Relations Research." *World Politics* 56 (1):114–38. doi:10.1353/wp.2004.0005.
- Loyle, Cyanne E., Christopher Sullivan, and Christian Davenport. 2014. "The Northern Ireland Research Initiative: Data on the Troubles from 1968 to 1998." *Conflict Management and Peace Science* 31 (1):94–106. doi:10.1177/0738894213501974.

- Lum, Kristian, Megan Emily Price, and David Banks. 2013. "Applications of Multiple Systems Estimation in Human Rights Research." *American Statistician* 67 (4):191–200. doi:10.1080/00031305.2013.821093.
- Marshall, Monty G., Ted Robert Gurr, and Keith Jagers. 2016. "Polity IV Project Dataset Users' Manual, V.2015." Polity IV Project, 1–86.
- Maxmind. 2018. "GeoLite2 IP Database." <https://dev.maxmind.com/geoip/geoip2/geolite2/>. Accessed 24 July 2018.
- Oliver, Pamela E., and Daniel J. Myers. 1999. "How Events Enter the Public Sphere: Conflict, Location, and Sponsorship in Local Newspaper Coverage of Public Events." *American Journal of Sociology* 105 (1):38–87. doi:10.1086/210267.
- Pettersson, Thérèse, and Kristine Eck. 2018. "Organized Violence, 1989–2017." *Journal of Peace Research* 55 (4):535–47. doi:10.1177/0022343318784101.
- Price, Megan, and Patrick Ball. 2014. "Big Data, Selection Bias, and the Statistical Patterns of Mortality in Conflict." *SAIS Review of International Affairs* 34 (1):9–20. doi:10.1353/sais.2014.0010.
- Schrodt, Philip A. 2012. "CAMEO: Conflict and Mediation Event Observations Event and Actor Codebook." Pennsylvania State University. Accessed 1 May 2019. <http://eventdata.parusanalytics.com/cameo.dir/CAMEO.Manual.1.1b3.pdf>.
- Solaimani, Mohiuddin, Sayeed Salam, Latifur Khan, Patrick T. Brandt, and Vito D'Orazio. 2017. "APART: Automatic Political Actor Recommendation in Real-Time." In *Social, Cultural, and Behavioral Modeling*, edited by D. Lee, YR. Lin, N. Osgood, and R. Thomson, vol. 10354, 342–48. LNCS Cham: Springer.
- Strobl, Carolin, Anne Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. "Conditional Variable Importance for Random Forests." *BMC Bioinformatics* 9 (307):1–11. doi:10.1186/1471-2105-9-307.
- Strobl, Carolin, Anne Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics* 8 (25). doi:10.1186/1471-2105-8-25.
- Sundberg, Ralph, Kristine Eck, and Joakim Kreutz. 2012. "Introducing the UCDP Non-state Conflict Dataset." *Journal of Peace Research* 49 (2):351–62. doi:10.1177/0022343311431598.
- Sundberg, Ralph, and Erik Melander. 2013. "Introducing the UCDP Georeferenced Event Dataset." *Journal of Peace Research* 50 (4):523–32. doi:10.1177/0022343313484347.
- Tollefsen, Andreas Forø, Håvard Strand, and Halvard Buhaug. 2012. "PRIO-GRID: A Unified Spatial Data Structure." *Journal of Peace Research* 49 (2):363–74. doi:10.1177/0022343311431287.
- UCDP. 2018. "Methodology." Uppsala University, Accessed 1 July 2018. <https://pcr.uu.se/research/ucdp/methodology/>.
- Urlacher, Brian R. 2009. "Wolfowitz Conjecture: A Research Note on Civil War and News Coverage." *International Studies Perspectives* 10 (2):186–97. doi:10.1111/j.1528-3585.2009.00369.x.
- Ward, Michael D., and Andreas Beger. 2017. "Lessons from near Real-time Forecasting of Irregular Leadership Changes." *Journal of Peace Research* 54 (2):141–56. doi:10.1177/0022343316680858.
- Weidmann, Nils B. 2015. "On the Accuracy of Media-based Conflict Event Data." *Journal of Conflict Resolution* 59 (6):1129–49. doi:10.1177/0022002714530431.
- Weidmann, Nils B. 2016. "A Closer Look at Reporting Bias in Conflict Event Data." *American Journal of Political Science* 60 (1):206–18. doi:10.1111/ajps.12196.
- World Bank Development Indicators. World Bank Open Data. Accessed 2 November 2017. 2016. <http://data.worldbank.org/indicator>.