The More The Merrier? Addressing Duplications in Automated Event Data. *

Sebastian Schutte, Howard Liu, & Michael D. Ward

We introduce a method for finding duplicates in machine-coded event data with potential applications in several high-profile data projects. First we select likely duplicates based on temporal and spatial proximity. In a second step, we rely on Natural Language Processing tools to generate various distance metrics for the text information associated with each event. These distances serve as a basis for multivariate classification models that are trained on a subset of human coded duplicates. We apply this method to two empirical samples from the ICEWS data collection and achieve strong classification performance at low rate of false positives.

Introduction

Recent years have seen a flurry of new data projects in political science that disaggregate interactions of political actors into single events. ACLED (Raleigh et al. 2010) and GED (Sundberg, Lindgren, and Padskocimaite 2010) code single conflict events within civil wars in Africa; SCAD (Salehyan et al. 2012) provides information on non-lethal political protests. The focus of the efforts in this paper, the Integrated Conflict Early Warning System (ICEWS) (Lustick et al. 2015; Boschee et al. 2015; Lautenschlager, Shellman, and Ward 2015) is a special type of data collection as it results from a largely automated coding procedure: ICEWS uses software to code politically relevant events in near real-time from natural language news sources. It features global coverage of events going back to 1990. Similar projects including the "Global Dataset on Events, Location, and Tone" (GDELT, see Leetaru and Schrodt 2013) as well as the Open Event Data project (see `http://openeventdata.org/`) have recently sparked significant interest.

There are a number of important differences between machine- and hand-coded data.

*Thanks to Andreas Beger for helpful comments. We would also like to thank the EU FP7 Marie Curie Zukunftskolleg Incoming Fellowship Program (Grant #291784) for financial support. Ward was supported in part by NSF Award SES 1259266.

Perhaps the most obvious difference is that machine-coding systems can collect and code events much faster and at a much lower cost. Additionally, machine-coding allows researchers to develop specialized databases for specific topics or locales. However, automated event databases come have some drawbacks. Generally, the massive increase in data *quantity* comes at the price of reduced data *quality*. While human coders can apply complex coding rules, machines rely on simpler decisions for event classification. Moreover, virtually all hand- and machine-coded databases rely on media reports about political events. Such reporting can be prone to reporting bias and selection effects (Weidmann and Rød 2015).

Building on the recent attempts to enhance the quality of event data, we focus on *event duplication*, i.e. the problem of a single political event that corresponds to multiple records in a database. While duplicated events can frequently be found in event datasets, they have received less attention than other issues.[1] Duplicates in event data pose a serious threat to all inferential studies: explaining the intensity of conflicts or protests, for example, crucially relies on accurately measured event counts. Our first analysis of the ICEWS data collection suggests that *approximately 30% of all events could be duplicates of previously coded events.*

In summary, the increased availability of event data presents researchers with a dilemma: on the one hand, a broad spectrum of political interactions that were largely resistant to quantitative analysis can now be analyzed in great detail. On the other hand, media biases and automated coding raise questions about the quality of contemporary event data. We use previous research on quality issues in media-based event data as a point of departure to identify and implement a remedy for one particularly important source of error: duplicated events.

The solution combines logical pre-selection with Natural Language Processing to finding duplicated observations in ICEWS. In the next section, we illustrate the magnitude of the problem. After that, we introduce an algorithm for identifying duplicated observations and demonstrate its efficiency in empirical applications.

Duplicates in ICEWS Event Data

ICEWS—International Crisis Early Warning System—began as a DARPA program led by Sean P. O'Brien. Subsequent funding came from the Office of Naval Research, and other agencies in the U.S. national security community. Initially, several different groups were in competition to develop the most accurate system, the team lead by Lockheed Martin

[1]For a discussion of this problem `https://dartthrowingchimp.wordpress.com/2014/06/06/another-note-on-the-limitations-of-event-data/`, received February 1, 2016.

was the only team that was continued past the first phase. Initially, this project focused on twenty-eight countries that fall under the US Pacific Command, but it expanded to include continuous monitoring over 250 news sources and other open source material covering 177 countries worldwide. ICEWS consists of several components, including a database of over 38 million multilingual news stories going back to 1990 and present to last week. These stories are parsed and coded with the ACCENT ontology which automatically identifies and extracts geo-located and dated events consisting of an actor, target, and action category. ICEWS itself consists of a suite of predictive programs that forecasts various types of domestic and international crises in the short term (one month ahead) as well as medium term (over 36 months). The accuracy of these predictions has been in the area of 85%. In 2015, ICEWS transitioned to a program of record in the US Defense Department, with primary sponsorship from the US Strategic Command, and has been actively used by a wide variety of commanders and planners. While the models and their uses are largely proprietary, through the efforts of the Office of Naval Research, the event database has been put in open source. The dictionaries, aggregations, ground truth data, and actor and verb dictionaries are publicly available with a one year lag at the ICEWS data repository `https://dataverse.harvard.edu/dataverse/icews`. In addition, for scholars who have access to news stories, the event coder has been made available publicly by the Office of the Director of National Intelligence.[2] In this article we focus on the event data that comprise part of the ICEWS database and address the issue of detecting duplicate events.

Duplicates can occur in machine-coded event data for three principal reasons. First, an original news report can be picked up by different news outlets. These “copy and paste" duplicates are especially frequent in the early phases of a breaking news story. As only scarce information is available from the Associated Press or Reuters, for example, identical text elements can be found across media outlets. Second, different news media can write up events independently, but rely on identical factual information. In this case, text elements between news articles may differ greatly, but the derived events would be identical. Finally, updates to previous stories oftentimes reiterate the original event which also leads to duplicates. In this first systematic attempt to address the problem, we focus on duplicates of type one and three, i.e. copies of existing news texts and updates to previously published stories. Empirically, we have found these types of duplicates to be most prevalent while they are comparatively easy to identify automatically.

To get a sense of the severity of the duplication problem in ICEWS, we hand-coded duplicates in a recent two-year time period (2010 to 2012) by reading all associated news stories of protest events in China and Mexico. For the China sample, 27.14% of all observations were identified as duplicates whereas 31.66% of observations in the Mexico sample were duplicates. These numbers suggest that repeated events in the ICEWS data

[2]Details at `http://bit.ly/2nS4nBU`.

collection pose a substantial duplication problem.

To illustrate the types of duplicates we have found in ICEWS, consider a 2010 protest event in Hong Kong: On September 19, a number of anti-Japanese protests were organized across Hong Kong–including one with an estimated 200 people at the Japanese consulate–to mark the anniversary of the Japanese invasion of Manchuria. The protests was reported and corresponding events were added to the ICEWS data collection. But because multiple sources reported on the protests two distinct records (events) were created. The two events both address the anti-Japanese protests in Hong Kong and use nearly identical language in reporting (see Table 1).

TABLE 1 *Duplicate with minor text changes example in China. Bold text indicates differences in the descriptions.*

Event 16396650	Event 16397971
*SCMP: Slogans, Shaved Heads Mark **Anti-Japanese** Protests in HK*	*Slogans, shaved heads mark protests in HK*
A number of anti-Japan protests were also staged across Hong Kong **to mark the 79th anniversary of the Japanese invasion of Manchuria.** More than 200 people marched from Victoria Park to the Japanese consulate and the central government's liaison office in Central.	A number of anti-Japan protests were also staged across Hong Kong. **In the biggest,** more than 200 people marched from Victoria Park to the Japanese consulate and the central government's liaison office in Central.

Despite the modifications, it is clear that either one of these stories was based on the other, or they were both based on a third independent text that was not captured by ICEWS. While some information might be lost by removing duplicate stories from individual studies (for example, if event 16396650 were deleted it would not be clear from the text of event 16397971 that the protests were marking the anniversary of the Japanese invasion) the vast majority of quantitative analyses will likely benefit more from accurate counts than they would benefit from the small amounts of additional information contained in a near-identical news story. In the following section, we introduce a software solution for finding duplicates in ICEWS.

Proposed remedy: the DupDe algorithm

To alleviate this a problem for ICEWS and similar data projects, we have developed an algorithm that identifies duplicates in three steps:[3]

1. *Logical preselection:* First, suspected duplicates are identified logically. Duplicates always occur after an initial event within a certain time frame. Based on the manual coding, we have found that more than 95% of duplicates occur within a seven day period and most of them on the same day as the original report. Moreover, duplicates share the multivariate signature of the initial event: they involve the same actors and actions and take place in the same country.

2. *Multivariate scores of language similarity:* For each suspected duplicate and the corresponding initial events, the news reports and headlines are analyzed. Specifically, different distance metrics are used to express lexicographic and phonetic similarity between the reports.

3. *Probabilistic classification based on multivariate models:* Finally, probabilistic classification models are used to predict whether or not suspected events are duplicates. Based on the predicted probabilities, applied researchers can then use customized thresholds for excluding suspected duplicates from their samples.

In the following sections, we will detail the individual steps of the duplicate detection algorithm and then demonstrate its abilities and limitations in a series of benchmarks.

Logical preselection of suspicious events

The first obvious feature of a duplicated event is that it is temporally preceded by a virtually identical event. To identify suspicious records, we therefore identify all events that have an almost identical event preceding them. For our application to ICEWS, we selected suspected duplicates based on the following criteria:

1. Posterior events are preceded by an event that happened up to seven days prior to them.

2. Posterior events have a different event ID from the preceding event, but they might rely on the same underlying news story.

[3]We decided to craft our own algorithm, because we are confronted with a unique data situation. Instead of only relying on natural language information, we can exploit the structural information in ICEWS including actor ids and locations. This fundamentally sets our study apart from related research on identifying plagiarism or near-identical duplicate detection in large text corpora.

3. Both events share IDs for location, type of interaction, as well as source and target actors.

This logical selection can be implemented in a nested SQL statement for event data that are stored in a relational database. The key to building a corresponding selection is to select two sets of events from the main event table which we call here *pre* and *post*. Selecting these two sets is demonstrated in the stylized SQL source code below.

```
SELECT
pre.<event attributes>,
post.<event attributes>
FROM events AS post,
(
        SELECT <event attributes>
        FROM   events
) AS pre
WHERE pre.event_id != post.event_id
AND   post.event_date -7 <= pre.event_date
AND   <other attributes of pre and post identical>
```

Database records that are selected based on the above criteria are then associated with corresponding news stories and headlines for further analysis as described below. All events within the *post* sample are treated a suspected duplicates and the final classification is performed probabilistically based on the similarities in the natural language description of the events as described below.

Calculating multivariate similarity scores

Once suspected duplicates and the corresponding original record are logically selected, we can quantify the similarity of the natural language event descriptions. Both computer science and computational linguistics offer a variety of tools that can be used for this purpose. The problem that we are facing, however, differs slightly from the more ambitious attempt to extract factual information from natural language corpora or to cluster texts according to topics. Instead, we merely need to measure how similar two texts are to determine whether or not they refer to the same real-world event. To accomplish this, we rely on a series of distance measurements for both the stories and their headlines.[4]

[The next two paragraphs need more work]

Two statistics have proven to be reliable predictors for duplicates and are included for probabilistic classification in different versions. The **Levenshtein distance** is a classic

[4]The distance measures are implemented and documented in the R package “stringdist" (van der Loo 2014) which we used in our implementation.

measure of string similarity that encodes how many characters in a text string have to be changed in order to turn it into another text string (see Levenshtein 1966). For this application, higher distances indicate that stories or headlines associated with events differ of a character-by-character basis. Of course, this can also be the case if stories or headlines simply differ in length.[5]

The second measure central to the classification is based on similarity rather than difference. The **Q-gram distance** is based on the length of common substrings. Q-grams are q consecutive characters present in both stories. Clearly, two stories can have multiple such q-grams which can be represented as vectors of q-gram lengths x and y. The Q-gram distance is the sum over the absolute differences $|x_i - y_i|$.

As we have no strong theoretical expectations about which distance measures work best in this specific application, we include several measures and chose a final classification model in a data-driven fashion.[6] In a final step, we use the multivariate distance measures for suspected duplicates for probabilistic classification. This step is described below.

Probabilistic classification

In a final step, probabilistic classification models are used to classify duplicates and original events based on similarity scores. These models are fitted on a subset of the event data that have been classified by human coders and then applied to the classification of a test sample.

For the empirical test of the DupDe algorithm, we have limited ourselves to three candidate models that we briefly discuss here. We started with two logistic regression models that were fitted on different subsets of the explanatory variables. The full regression model (model 1 hereafter) contains all the above mentioned explanatory measures. A second logistic regression model (model 2) was chosen algorithmically via step-wise variable selection based on the full model. The selection criterion was the AIC statistic (Akaike 1974). AIC essentially rewards high goodness-of-fit while penalizing on the number of explanatory variables. Minimizing AIC values through variable selection is a frequently chosen approach to preventing overfitting and generating models with good predictive capabilities.

[5]Please note the we rely on a variant of the Levenshtein distance that has proven to be a more reliable predictor. The **weighted, restricted Damerau-Levenshtein distance, with optimal string alignment)** is the distance between strings as measured by the minimum number of operations (deletion, substitution, insertion, transposition) required to convert either of the strings into the other. We incorporate both the raw measure as well as a version that is normalized by the length of the preceding stories and headlines.

[6]Please refer to section XYZ in the supplementary information for an overview and a formal introduction of all the distance measures involved in the analysis.

Model 3 uses a Support Vector Machine (SVM) to classify duplicates (see Cortes and Vapnik 1995).[7] The advantage that SVMs offer over regression-based classification models is that they do not require an ex-ante expectation about the functional form between predictors and dependent variable. In section , we discuss the performance of these three models for subsets of duplicates in the ICEWS data collection.

Case selection: Protests in ICEWS

To generate a manageable subset of event types for manual coding, we decided to focus on protest events within ICEWS. As the approach easily generalizes to related projects, such as Open Event Data, we will make the code fully available. In recent years, there has been an emerging interest in studying protests and nonviolent campaigns in the conflict literature (Chenoweth and Cunningham 2013; Cunningham 2013; Celestino and Gleditsch 2013; Sutton, Butcher, and Svensson 2014; Gleditsch and Rivera 2015; Chenoweth and Ulfelder 2015).

Within ICEWS, we focus on two cases for studying the severity of the duplication problem and the effectiveness of our remedy: China and Mexico. Protest reports in China are a hard test case for duplication as the problem is particularly severe in recording Chinese local demonstrations. Without presentation of local correspondents and informants, international media often rely on regional new agencies such as South China Morning Post, Xinhua News, or Hong Kong Information Centre for Human Rights and Democracy reports to reveal indigenous protest information. The news they release is then reused by other newswires such as Taipei News and AFP, creating multiple reprints of the same story. The prevalence of this news reprinting and regenerating process makes protest events in China a hard test for the approach. In creating the training data, we hand coded 200 Chinese protest events collected by ICEWS in the span of 2010-2012 where data quality is generally superior to older ICEWS observations. In order to test the effectiveness of the approach beyond one case. We perform out-of-sample classifications for protest events in Mexico.

[7] In essence, an SVM automatically chooses observations that “demarcate" the boundary between two classes (duplicates and non-duplicates in our case). Kernels are fitted to predict an area around the chosen observations that belongs to the same class as these observations.

RESULTS

Logical preselection. As described in Section , we first select suspicious events logically. By comparing locations, actors, and types of interactions between events in close temporal proximity, 80 out of the 193 hand-coded events for the China sample were flagged as suspicious – 41.5% of all observations.

In principle, these suspicious events could just be excluded from data analysis. Table 2 shows counts for correctly and incorrectly suspected observations only based on logical selection. Counts for true duplicates and true non-duplicates were obtained from the human coding while the suspected duplicates came from the logical selection. The main diagonal shows counts for correctly suspected events and holds about 75% of all observations.

	Duplicates	Non-duplicates
Suspected duplicates	44	36
Suspected non-duplicates	13	100

TABLE 2 *Counts of suspected and real duplicates for the China sample. The percentage of events on the main diagonal (which are correctly classified) is 74.6.*

For the Mexico sample, the results for the preselection are comparable as shown in Table 3. Of the 183 classified events, 130 are classified correctly (0.71%). This result is very encouraging in that it shows that simple processing of machine-readable event information goes a long way in identifying unwanted duplicates.

	Duplicates	Non-duplicates
Suspected duplicates	31	40
Suspected non-duplicates	13	99

TABLE 3 *Counts of suspected and real duplicates for the Mexico sample. The percentage of events on the main diagonal (which are correctly classified) is 71.1.*

However, the cross tables also indicate the limitations of logical preselection: the top right cells in both tables show counts of 36 and 40 observations respectively. Extrapolating from these counts, more than 20% of original observations in ICEWS would be incorrectly flagged as duplicates. Clearly, deleting original observations has to be avoided as it could introduce problems of non-random missing data in subsequent statistical analysis. To address this issue, the we incorporate probabilistic classification based on stories and headlines associated with the events in an additional step.

Performance of classification models. The probabilistic classification models utilize different string distance measures as discussed in section . The full regression model utilizing all available string similarity measures (model 1), and optimized regression model

(model 2), and a Support Vector Machine (model 3) were trained on the hand-coded China sample. The fitted models were then used to classify observations in the Mexico sample. To assess the performance of the probabilistic classification models, we calculated ROC statistics and plotted corresponding curves (see Figure 1). In a first step, the models were used to classify the observations they were fitted on. In this exercise, all three models performed equally well. The SVM classifier (Model 3) performed best overall with an area under curve (AUC) of 0.78. Both the full regression model as well as the optimized one (models 1 and 2) achieve very similar classification performances.

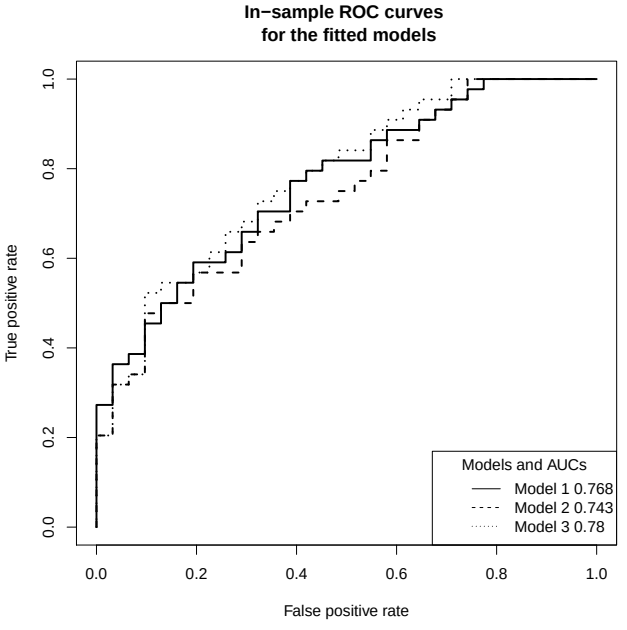

Figure 1. In-sample ROC curves for the fitted models. The legend on the bottom right also shows corresponding area under the curve (AUC) values. Please note that the performance is especially good in the left half of the plot: a true positives rate of 0.5 can be achieved at a 0.1 false positive rate.

The real test for the algorithm, however, is out-of-sample classification. To this end, we rely again on the second hand-coded database for Mexico.

Applying the fitted models to Mexico

The models that were fitted on the China sample were used to predict duplicates in Mexico. This test allowed us to assess the performance of the approach for a larger sample of the

ICEWS data. Beyond increasing the absolute classification performance, the predicted probabilities can also be used to reduce the false positive rate: when applied researchers are concerned about non-random data deletion, they can simply pick a high predicted probability as a cutoff for selecting their sample of events.

Performance of classification models. As visible in Figure 2, the performance of all three models is substantively reduced in comparison to the in-sample test. Interestingly, the AIC–optimized regression model yields the best AUC at 0.66.

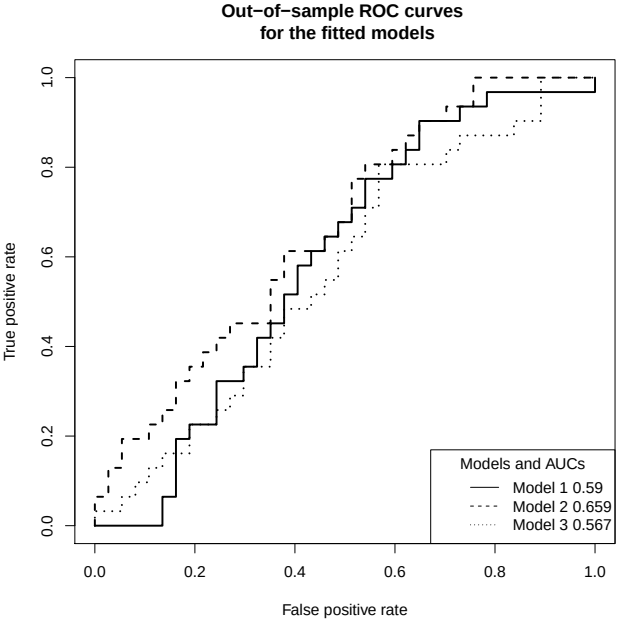

Figure 2. Out-of-sample ROC curves for the fitted models. The legend on the bottom right also shows corresponding area under the curve (AUC) values. Please note out-of-sample prediction performance is weaker than the in-sample performance. The optimized GLM (Model 2) achieves highest scores.

But how can the probabilistic models be used to prevent the deletion of original observations? When choosing a cutoff at .5 for the predicted probabilities for observations being duplicates with model 2, the counts for correctly and incorrectly classified observations change. Table 4 shows the corresponding counts. Only 19 (as opposed to 40) of the 193 observations are incorrectly classified as duplicates. At the same time, the number of correctly classified duplicates has gone down from 31 to 22 while the number

of correctly classified original observations has gone up from 99 to 120. In summary, the biggest problem of the logical selection approach –false positive classifications– has been mitigated by combining multivariate string-similarity measures with probabilistic predictions. The overall classification performance has been increased from 71% to 77.6% out-of-sample.

	Duplicates	Non-duplicates
Suspected duplicates	22	19
Suspected non-duplicates	22	120

TABLE 4 *Cross-classification Results for Suspected Duplicates and Non-duplicates using a cutoff of* 0.5 *for Model 2, based on logical pre-selection.*

A closer look at the most successful prediction model (2) also reveals which predictors contribute to classification. Table 5 shows that the logical preselection also contributes to probabilistic classification: the *times suspected* variable indicates how many times an observation has been flagged as a possible duplicate in the logical selection step. Widely visible real-world events generate multiple media stories. In such cases, highly clustered sets of duplicates enter the ICEWS data collection and the logical preselection identifies trailing events as duplicates multiple times.

For the headlines, both the q-gram and Jaccard distances are reliable predictors of duplicates. For the news stories, the optimal string alignment and q-gram distances perform best. As with the logical preselection, these results suggest that comparatively simple metrics and models can be used to achieve good classification performances. After all, OSA is very similar to the classic Levenshtein distance (see Wagner and Fischer 1974), but it features an important extension. Unlike the original measure, OSA is not just the number of character replacement steps necessary to turn one string into another, but it also allows for transposition of adjacent characters. The q-gram variable expresses the length of the longest common substring in two strings. Interestingly, the more complex string kernel similarity and the phonetic similarity measure are not among the most reliable predictors. In summary, logical preselection, an optimized GLM, and fairly basic measures of string similarity go a long way in classifying unwanted duplicates in machine-coded event data.

Discussion and conclusion

In this research note, we have argued that duplicates in machine-coded event data pose a serious problem and corroborated this claim with a simple coding exercise. Based on human coding, approximately one third of all protest events reported for China and Mexico between 2010 and 2012 were identified as duplicates. Very likely, this situation will lead to substantive claims based on biased data and overly optimistic predictive results.

To alleviate this problem, we have demonstrated that logical pre-selection already

TABLE 5 *General Linear Model estimates and standard errors for the simple logit model of duplicate event detection using six story characteristics.*

	Dependent variable:
	Is Story Duplicate?
Times suspected	0.918
	(0.431)
Headline *q*-gram	0.042
	(0.023)
Headline jaccard	−6.064
	(3.554)
Story OSA	−0.003
	(0.001)
Story *qgram*	0.002
	(0.001)
Story OSA (normalized)	−0.007
	(0.005)
Constant	−0.177
	(0.583)
Observations	75
Negative Log Likelihood	−42.352
Akaike Information Criterion	98.703

offers a good remedy. However, while pre-selection can filter duplicates, it is also prone to producing false positives, i.e. original observations that should be preserved in the data collection. The successfully tested remedy for this problem is a combination of generating multivariate string distances for both headlines and stories of suspected duplicates and inductive statistics. We have found that a small number of predictors in a logistic regression model can be used to eliminate almost all of the false positives from the pre-selected sample. We hope that the introduced procedure will be used in machine-coded event data projects. To this end, we will make the source code for this project available.

References

Akaike, Hirotugu. 1974. "A new look at the statistical model identification." *IEEE Transactions on Automatic Control* 19 (6): 716–723.

Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael D. Ward. 2015. "ICEWS Coded Event Data." Harvard Dataverse Network (http://dx.doi.org/10.7910/DVN/28075), March.

Celestino, Mauricio Rivera, and Kristian Skrede Gleditsch. 2013. "Fresh Carnations or All Thorn, No Rose? Nonviolent Campaigns and Transitions in Autocracies." *Journal of Peace Research* 50 (3): 385–400.

Chenoweth, Erica, and Kathleen Gallagher Cunningham. 2013. "Understanding Nonviolent Resistance: An Introduction." *Journal of Peace Research* 50 (3): 271–276.

Chenoweth, Erica, and Jay Ulfelder. 2015. "Can Structural Conditions Explain the Onset of Nonviolent Uprisings?" Doi: 10.1177/0022002715576574, *Journal of Conflict Resolution.*

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-vector networks." *Machine Learning* 20 (3): 273–297.

Cunningham, Kathleen Gallagher. 2013. "Understanding Strategic Choice: The Determinants of Civil War and Nonviolent Campaign in Self-determination Disputes." *Journal of Peace Research* 50 (3): 291–304.

Gleditsch, Kristian S., and Mauricio Rivera. 2015. "The Diffusion of Nonviolent Campaigns." Doi: 10.1177/0022002715603101, *Journal of Conflict Resolution.*

Lautenschlager, Jennifer, Steve Shellman, and Michael D. Ward. 2015. "ICEWS Coded Event Aggregations." Harvard Dataverse Network (http://dx.doi.org/10.7910/DVN/28117), March.

Leetaru, Kalev, and Philip Schrodt. 2013. *GDELT: Global Data on Events, Language, and Tone, 1979 - 2012.* Paper presented at the International Studies Association Annual Conference, April 3rd - 6th, 2013, San Francisco, CA.

Levenshtein, Vladimir I. 1966. "Binary codes capable of correcting deletions, insertions, and reversals." *Soviet Physics Doklady* 10 (8): 707–710.

Lustick, Ian, Sean O'Brien, Steve Shellman, Timothy Siedlecki, and Michael D. Ward. 2015. "ICEWS Events of Interest Ground Truth Data Set." Http://dx.doi.org/10.7910/DVN/28119 Harvard Dataverse Network [Distributor] V1 [Version], March.

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): 651–660.

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

Sundberg, Ralph, Methilda Lindgren, and Ausra Padskocimaite. 2010. *UCDP GED Codebook version 1.5-2011.* Available online at http://www.ucdp.uu.se/ged/.

Sutton, Jonathan, Charles R Butcher, and Isak Svensson. 2014. "Explaining Political Jiu-jitsu." *Journal of Peace Research* 51 (5): 559–573.

van der Loo, M.P.J. 2014. "The stringdist package for approximate string matching." *The R Journal* 6 (1): 111–122. `http://CRAN.R-project.org/package=stringdist`.

Wagner, Robert A., and Michael J. Fischer. 1974. "The String-to-String Correction Problem." *ACM* 21 (1): 168–173.

Weidmann, Nils B., and Espen Geelmuyden Rød. 2015. "Making Uncertainty Explicit: Separating Reports and Events in the Coding of Violence and Contention." *Journal of Peace Research* 52 (1): 125–128.