Article

cmps

Conflict Management and Peace Science
2015, Vol. 32(2) 222–242

DOI: 10.1177/0738894214559680
cmp.sagepub.com
SAGE

The MID4 dataset, 2002–2010: Procedures, coding rules and description

Glenn Palmer
Department of Political Science, Pennsylvania State University, USA

Vito D'Orazio
Institute for Quantitative Social Science, Harvard University, USA

Michael Kenwick
Department of Political Science, Pennsylvania State University, USA

Matthew Lane
Department of Political Science, Pennsylvania State University, USA

Abstract

Understanding the causes of interstate conflict continues to be a primary goal of the field of international relations. To that end, scholars continue to rely on large datasets of conflict in the international system. This paper introduces the latest iteration in the most widely used dataset on interstate conflicts, the Militarized Interstate Dispute (MID) 4 data. In this paper we first outline the updated data-collection process for the MID4 data. Second, we present some minor changes and clarifications to the coding rules for the MID4 datasets, as well as pointing out how the MID coding procedures affect several notable "close call" cases. Third, we introduce updates to the existing MID datasets for the years 2002–2010 and provide descriptive statistics that allow comparisons of the newer MID data to prior versions. We also offer some best practices and point out several ways in which the new MID data can contribute to research in international conflict.

Keywords

Conflict, data, MID, militarized interstate dispute

Corresponding author:
Matthew Lane, Department of Political Science, Pennsylvania State University, 203 Pond Laboratory, University Park, PA 16802, USA.
Email: mattlane25@gmail.com

Introduction

Examining the characteristics of war and peace continues to be at the center of international relations, and understanding the causes of conflict remains one of the main objectives of the international conflict literature. To that end, scholars have routinely relied on the collection and aggregation of real-world events into usable data through systematic and coherent coding rules and data-collection procedures. Among such datasets, the Militarized Interstate Disputes (MID) data continues to be one of the most heavily utilized datasets in the field.

In 2003, the MID3 dataset (Ghosn et al., 2003) built upon earlier versions of the MID data (referred to as "MID1" and "MID2.1", respectively; Jones et al., 1996) by updating the data through 2001 and by developing the new incident-level datasets. The MID4 project furthers this development by updating the MID and incident-level datasets through the year 2010 and cleaning older portions of the MID data. Additionally, the MID4 project updates the data-collection procedures first developed in MID3 toward developing a more rigorous and efficient data-collection method.

This paper proceeds in four sections prior to concluding. First, we outline the data-collection process utilized in the MID4 project and discuss the systematic benefits of this approach. Second, we provide some clarifications of the original MID coding rules and discuss how the rules affect the inclusion or exclusion of several salient instances of conflict. Next we present the updated MID4 data and provide some descriptive analyses on the most recent period of data collection, illustrating several points of comparison with previous iterations of the data. Following this descriptive analysis, we provide a series of best practices for proper use of the MID and incident-level data for research in international conflict.

MID4 data collection: an overview

The process used to collect the MID4 data is more automated than previous MID projects, enabling us to reduce the resources necessary to create the dataset. MID2 researchers, for example, manually searched through Keesings and the New York Times, among other sources. MID3 researchers primarily queried LexisNexis and manually searched through the results. While perfectly reasonable approaches, these types of searches are inefficient and, indeed, one of the conclusions drawn from the MID3 experience was that the identification of the "true positive" news reports was time-consuming and could be reduced.

The MID4 project addressed this source of inefficiency by utilizing automated document classification techniques to identify relevant news reports efficiently. LexisNexis was queried using general search terms to retrieve a document set containing large numbers of news stories. From this set, any document not containing the names of at least two geopolitical entities was discarded. Then a common method for classifying text documents, the Support Vector Machine (SVM) algorithm, was used to classify the documents automatically into two bins: those that are predicted to contain information about events relevant to a MID, and those that are predicted to be irrelevant. The resulting document set contains a manageable number of news reports to be read and analyzed manually.

News sources and search parameters

To retrieve the initial document set, LexisNexis was queried using search terms that are sufficiently general to be all-inclusive (Schrodt et al., 2008). The specific set of terms is reported

Table 1. MID4 search terms

(air base OR air strike OR airbase OR aircraft OR airstrike OR alert OR antiaircraft OR armed OR armo! OR arms OR army OR artillery OR attack OR batteries OR battery OR battle OR battleship OR block! OR bomb OR border OR buildup OR carrier OR casualties OR casualty OR cease OR ceasefire OR cease-fire OR clash! OR combat OR conflict OR crisis OR cruiser OR damage OR declare war OR defence OR defense OR defensive measures OR defian! OR deploy! OR destroy OR detained OR dispatch! OR display of force OR dispute! OR embargo OR erupt! OR fight! OR fire OR fired OR forc! OR fortification OR hit OR hostile OR incursion! OR infantry OR interstate OR invasion OR jet OR kill! OR launch! OR liberate OR line of control OR maneuver OR milit! OR missile! OR mobiliz! OR mortar OR naval OR nuclear OR occup! OR offensive OR operation OR patrol! OR peace declaration OR pullback OR radar OR raid! OR recon! OR reinforcement OR reprisal OR retali! OR rocket OR security OR seiz! OR shell! OR shoot OR shot down OR show of force OR shrapnel OR skirmish OR soldier! OR squadron OR stronghold OR subside! OR target OR tension! OR territ! OR threat! OR trade fire OR troop OR truce OR ultimatum OR USS OR vessel OR violat! OR violence OR vows to OR war OR warn! OR warplane OR warship OR weapon! OR weapons OR withdraw!)
AND NOT (sports OR business OR lifestyle OR tax cuts OR entertainment OR Wall Street OR budget OR baseball OR food OR weather OR health OR natural disasters)

in Table 1. By taking this inclusive approach, our team had an acceptable level of certainty that the initial document set would contain information on all events that could be classified as a militarized interstate incident (MII), the events that comprise a MID. However, doing so for *all* news sources in the LexisNexis universe would introduce an enormous amount of redundant information into the document set, and so we first experimented with queries on different news sources to determine which ones would be sufficient.

The goal of the source selection experiments was to produce a list of sources that limits redundant information in the document set, such as the same news story being reported by multiple sources. To do so, an initial list of 30 candidate sources was selected based on temporal coverage (2002–2010) and known contributions to earlier MID projects. These 30 were grouped into three categories based on their scope of coverage. The broadest category, *global*, contained sources that report news worldwide. The second category, *regional*, consisted of sources that predominantly report news from a set of countries confined to a particular geographic region. The third category, *national*, contained sources confined to an individual state or similarly small geopolitical unit.

Ten cases were then drawn randomly from the MID3 incident-level data.[1] Next, LexisNexis was queried for each group of news sources, and each group was evaluated based on the results of the query. Specifically, we wanted to know if the query returned documents containing information about the 10 MID3 incidents. We found that the *global* sources reported on events relevant for all 10 incidents; while *regional* and *national* reported relevant events as well, these documents contributed no new information. That is, the *global* sources contained at least as much information as the more localized sources. To reduce redundancy in the initial document set, we proceeded with just the *global* list. The sources used are reported in Table 2.

LexisNexis was queried using the search terms and source list specified above. These queries resulted in a set of 1,744,517 documents, or a yearly average of about 193,000. Automated methods were then used to pare this document set down to a more manageable number.

Table 2. MID4 news sources

Agence France Presse	Deutsche Presse Agentur	London Times
AFX News	Interfax	Montreal Gazette
Associated Press	ITAR-TASS News Agency	New York Times
British Broadcasting Corporation	Japan Economic Newswire	United Press International
CNN	Jerusalem Post	Xinhua General News Service

Table 3. Example of a term–document matrix

	military	meeting	aid
Document 1	0.4	0.2	0.4
Document 2	0.1	0.1	0.8
Document 3	0.3	0.3	0.4

Automated document classification

Using automated document classification, the initial set of 1,744,517 stories was reduced to 132,515. To summarize, this was accomplished in four successive steps: (1) the documents were pre-processed; (2) the documents were represented as data; (3) the set was classified using inductive SVMs; and (4) the set was classified again using transductive SVMs.[2]

In pre-processing, each document was formatted and searched in an effort to recognize any mention of a geopolitical entity.[3] If a document did not contain at least two such entities, it was discarded. In step 2, all remaining documents were represented as data in a term–document matrix, an example of which may be seen in Table 3. Each row corresponds to a document, and each column to a term that is scored based on its *normalized term frequency* (ntf). The ntf is intended to score the importance, or impact, of a term for a particular document. It is calculated as the number of times a term appears in a document, divided by the number of terms in the document. For example, if the term "military" appeared five times in a document of 100 words, the "military" term would be scored 0.05 for that particular document.[4]

Steps 3 and 4 used SVMs to predict which documents (observations in the term–document matrix) will contain information relevant to a MID. SVMs are commonly used algorithms for the classification of text documents (Abe, 2010; Dumais et al., 1998; Zhang and Oles, 2001). In its simplest form, the general idea behind SVMs is to estimate a hyperplane that best splits the data into two classes (Vapnik, 1995, 1998).[5] While there are many variations of SVMs available, we used the *inductive* and *transductive* algorithms for steps 3 and 4, which have been shown to be effective in applications similar to ours (Joachims, 1998, 2002).

In step 3, the inductive SVM model is trained using a set of 24,042 *labeled* documents spanning 1994–2001, consistent with what was used in the MID4 pilot study (Schrodt et al., 2008).[6] This model estimates the hyperplane that is used to classify the 1,744,517 observations in our term–document matrix. After classification, approximately 90% of these documents were predicted to be irrelevant.

While step 3 used a single model to classify a large portion of the data, step 4 was intended to be more precise. The idea behind step 4 was to split the remaining 10% of the documents, or those that were predicted to be "MID-relevant", by year and to conduct an additional classification on each year separately. We did this to allow for changes in the content of news stories, such as those pertaining to the Iraq War of 2003 and the Russia–Georgia conflict of 2008, to be directly reflected in the outcome of the classification. After grouping the remaining documents by year, we proceeded by randomly sampling approximately 250 documents from each year. These random samples were labeled as either "MID-relevant" or "irrelevant", and used as training data for the year-specific transductive models.[7] In step 4, approximately 21.6% of the remaining documents were removed, leaving the MID4 team to process 132,515 documents manually, about 15,000 per year.

To ensure that MID-relevant documents were not being discarded in step 4, we conducted a post-hoc evaluation of documents removed from a single year, 2003. Since 2003 contains the largest number of removed documents, we believe that it is a sufficient subset to examine. This review *did* locate news stories containing information relevant to MIDs, but this same information was also contained in other news documents positively identified by the automated selection process and already coded in the MID data. As such, these additional news documents would not have changed the coding of the MID data.

Manual classification and coding

Although the automated document classification process removes a large percentage of irrelevant stories from the document set, additional processing is required to reduce the total number of classified stories into a manageable set for final coding. This additional processing requires the removal of irrelevant stories that were nevertheless selected by the automated classification, but could be removed prior to coding to limit the amount of time spent sorting through unnecessary documents. Not unlike previous iterations of the MID project, this process proceeded manually. Specifically, the entire set of 132,515 classified documents was reviewed individually by research assistants and determined to be either relevant or irrelevant for MID coding. This process was referred to as tilling. Tilling removed an additional 90% of remaining stories, leaving about 1300 documents per year, or about 12,000 documents in total, for coding.

Once the irrelevant stories are removed from the document set, the remaining stories are bundled into potential incidents based on common dates and actors. Bundling stories prior to coding improves efficiency and reduces the possibility of double coding, which can occur when multiple research assistants unknowingly code the same incident.

Coding the approximately 1300 news stories per year proceeded in much the same fashion as was the case for the MID3 Project (Ghosn et al., 2003). Coding the news stories was done mostly by graduate students at Penn State. Some of the coding was distributed to helpful colleagues at Rice University and the University of Alabama.[8] Each of the coded incidents was reviewed by the entire MID staff at Penn State, who then aggregated the coded incidents into disputes.[9] During this collective review, the MID staff read and discussed the news sources, narrative and coding of each individual MII and MID, making changes when necessary to reach a consensus about the accuracy of the data. As such, although MIIs and MIDs are coded by individual researchers, the final data represent the collective review and agreement of the entire MID staff.

Table 4. *DispNum4* geographic region identifiers

Region code	Geographic region
1	Continental Africa
2	Central and Eastern Europe
3	Far East and Oceania
4	South and Central America
5	Middle East
6	North America
7	South Asia
8	Western Europe

Coding the selected news stories represented a bottleneck in the process of generating the data, as was the case for the MID3 project (Ghosn et al., 2003). As a result of these experiences, research is currently underway to investigate the value of crowdsourcing to code the next iteration of the MID data (D'Orazio et al., 2014a). The basic premise of the crowdsourcing experiment is to evaluate the ability of non-expert coders to extract the same information from a document as an expert coder. If successful, future updates to the MID dataset may proceed much faster and at considerably lower cost.

Clarification of coding rules and ambiguous cases

A significant effort was made during the MID4 project to clarify the original MID coding rules and render their application in coding conflicts across a changing global landscape consistent with the intent of the original coding rules. In this section we provide the reasoning behind two new variables, clarification of the coding rules and examples of the application of those coding rules.

New variables

Users of the MID4 data will notice two new identification variables (*DispNum4* and *IncidNum4*). These identification variables are in addition to the traditional MID identifiers (*DispNum3* and *IncidNum3*) and exist exclusively for disputes and incidents in the period 2002–2010.[10] The purpose of these new identification variables is to add some substantive meaning to the dispute and incident identifiers.

The new dispute identifier (*DispNum4*) is a five-digit identifier. Table 4 provides a full list of all geographic identifiers. The first digit refers to the geographic region of the dispute, specifically to the location of the targeted state of the first incident in the dispute. Generally, disputes were placed into the region corresponding to the geographic location of the dispute. Thus, although the USA is a North American state, disputes between the USA and Pakistan are coded as being in the South Asia region. The second and third digits refer to the beginning year of the dispute. As such, a five-digit code beginning with 102 would refer to a dispute in continental Africa beginning in 2002, while a five-digit code beginning with 308 would refer to a dispute in the Far East beginning in 2008. With few exceptions, the final two digits of the identifier refer to the chronological order of disputes in that region-year. This count starts over with each new region year. Combining these five digits, a dispute with

identifier code #10205 is the fifth MID beginning in 2002 in continental Africa, while MID #10301 is the first dispute beginning in 2003 in continental Africa.

The new incident identifier (*IncidNum4*) follows the same five-digit nomenclature and adds a final three-digit incident identifier to the dispute. With some exceptions, this three-digit identifier refers to the chronological order of incidents in the dispute. As such, incident #20601006 is the sixth chronological incident of MID #20601.

Clarification of coding rules

There were several clarifications on the application of coding rules. The first coding rule requiring significant clarification centers on continuous military actions (blockades, border fortifications, occupations of territory and seizures) and concerns two seemingly contradictory clauses in Jones et al. (1996). In the absence of complete information concerning the end of continuous actions, Jones et al. (1996) allow for the MID to end either: (1) 6 months after the start of the continuous action or (2) 6 months after the end of any subsequent incidents in the MID. Generally in MID4, incidents involving continuous actions are coded as lasting one day. At the dispute level, the MID ends 6 months from the start of the continuous action, unless subsequent incidents in the MID extend the MID longer than 6 months from the continuous action. In such instances, the MID ends on the end date of the final incident in the dispute. An exception to this coding procedure concerns seizures. Following the coding rules laid out by the MID3 project (Ghosn et al., 2003), in the absence of complete information regarding the end date of seizures, seizures are coded as lasting 3 days at the incident level. At the dispute level, the MID is coded as ending 3 days from the start of the seizure or the end date of any subsequent incidents, had any taken place within the following 6 months.

The easiest method of explaining this nuance is perhaps through two distinguishing cases. During the 2002 conflict between Ethiopia and Somalia (MID #4345), Ethiopian military forces occupied large swaths of Somali territory in the Puntland and Gedo regions. None of these occupations had any clear end dates. Looking at the incident-level data, all of these occupations are coded as single-day incidents. Following these occupations, there are three additional incidents that extend the MID longer than 6 months from the last occupation. As such, the end date of the MID is coded as the end date of the last incident (Incident #4345008). Similarly, the 2002 occupations of Rwanda by Congolese military forces (MID #4356) have no clear end date, and so are coded as single-day incidents. In contrast, however, the final incident of the MID (Incident #4356005) occurs within 6 months of the last occupation, so the end date of the MID is coded as being exactly 6 months from the start of the occupation.

Uncertainty has also arisen with respect to whether airspace violations are classified as shows of force or border violations. The MID Incident Coding Manual states that a show of force is "a public demonstration by a state of its military forces intended to intimidate another state but not involving actual combat operations", and notes that these include "the intentional violation of another state's territorial waters or airspace". A border violation, on the other hand, is "the crossing of a territorial land boundary ... by military forces of one state without any significant damage to the territory or population of the violated state". Consistent with the MID3 project, these rules were interpreted to mean that the intentional violation of maritime and airspace borders could potentially constitute a show of force, but never a border violation.

Consistent with the MID coding rules, some actions taken by states, although occurring contemporaneously, are grouped into separate incidents based on their geographic locations. While such incidents might encompass similar actions over similar times, that they take place in clearly different geographic areas requires that they be identified as distinct incidents. Of particular note are the Israeli aerial bombings of Lebanon (MID #4182 and MID #4533). In each of these MIDs, we distinguish between Israeli bombings of southern Lebanon, particularly targeting Hezbollah positions in the Shebaa Farms region, from Israeli bombings of Beirut and the surrounding area. Similar in this regard are the US and British actions enforcing the "no-fly zones" over northern and southern Iraq (MID #4273). While the resulting shows of force over these two areas occurred simultaneously, we distinguish between actions over the northern and southern locations and place them into separate incidents.

The increasing prevalence of violent non-state actors in international affairs also requires clarification of state revision types. In several instances in the data, states take militarized action against non-state groups located in other states. In most cases, these actions take the form of border fortifications, border violations or attacks. This presents a question of when states undertaking such actions are attempting to revise the status quo relationship with the target state and what, precisely, this revision is. Generally, incidents involving actions taken against non-state actors abroad are coded with policy as the revision type. The defining characteristic for whether or not a policy revision is coded is the degree of cooperation between the non-state group and the target state. For instance, Senegalese border fortification along the border with Gambia (MID #4351) is not coded as having a revision type, since there is no cooperation between the Gambian government and the Casamance rebels. In contrast, almost all actions taken by Israel against Hezbollah forces in Lebanon are coded as being a policy revision, since there is a high degree of congruence between Hezbollah and the Lebanese government.

Along these same lines, an increasingly prevalent feature of the international system is the use of unmanned aerial drones, particularly by the USA. Actions taken by such unmanned actors do not necessarily fit the conventional mold of militarized disputes, since they are not technically undertaken by uniformed troops. However, the decision was made that such "actors" still constitute a direct extension of a state's military and the decisions of uniformed military units, and so incidents involving unmanned aerial drones are included in the data.[11]

There was also extensive discussion regarding the coding of American use of armed drones to strike suspected Taliban and Al Qaeda positions in northern Pakistan. Several sources point out that such drone strikes occurred hundreds of times between 2005 and 2010. The MID coding rules direct that these attacks should be coded as distinct incidents if there is a three-day gap between them. Following that rule, these events were combined into 15 incidents, starting on 8 May 2005 and ending on 17 December 2010 (see MID #4568, MID #4571 and MID #4575). The incidents coded from the drone strikes were separated from incidents that did not involve drones, such as attacks by helicopter gunships, because of their unique features, such as high-altitude positioning and distanced control by military personnel.

Close calls and ambiguous cases

Several disputes in the MID4 data were difficult to code as they related to the historical record. Furthermore, several notable conflicts from the historical record were left out of the MID4 data. Additional care and attention was afforded these cases to make sure whether

they should or should not, by the MID coding rules, be included in the data and, if included, that they accurately adhere to the incident and dispute coding rules.

Perhaps the most notable absence from the MID4 data is any disputes between the USA and Afghanistan, even though President Karzai and other top Afghan officials protested several US attacks on Afghan towns and against the broader Afghan population in operations against Taliban and other rebel forces. Even given these explicit protests, the decision was ultimately made that, following the conclusion of the 2001 interstate war in Afghanistan (MID #4283), all US operations in Afghanistan occurred because of sustained cooperation between the two states and the continued invited presence of US forces by the Afghan government. Given this sustained cooperation, even protested actions were deemed non-conflictual by the dispute coding rules.

Somewhat less notably, the historical record shows several instances of border fortifications along Ethiopia's border with Somalia.[12] These border fortifications followed rebel operations in eastern Somalia, and were conducted to prevent a spillover of the ongoing Somali civil conflict into Ethiopia. In these instances, the decision was made not to code any militarized incidents, since, at the time, the Somali government was non-functioning and, in several cases, Ethiopian reports suggest that the motivating concern for the border fortifications was precisely the lack of any meaningful Somali government; Ethiopian forces were acting as an alternative to Somali action. Because of these two factors, the lack of a functioning target state and the suggested, if underlying, cooperative characteristics of the actions, such cases are excluded from the dataset. This is in contrast to several other instances of border fortifications against rebels, since in these other cases the rebels were operating in a functioning state and the militarized action occurred in spite of actions by the target government.

Also not included in the data are several purported instances of South Korean aggression against North Korea, particularly over supposed South Korean military movements over the Northern Limit Line in the Yellow Sea. These reports were made either by the North Korean government or the North Korean news agency. Given the particular lack of credibility about these claims, the decision was made to not include such incidents unless further verification could be found via additional news agencies.

Users of the MID4 data will notice a significant difference in the time-span of the Russian–Georgian conflict over South Ossetia and Abkhazia (MID #4436). The dispute culminated, in terms of hostilities, in what is colloquially referred to as the "2008 Russia–Georgia War". While this series of incidents were extremely destructive, they never produced enough military fatalities to be coded as an interstate war. Additionally, while Russia clearly

Table 5. Distribution of MIDs by region

Region	Number of MIDs
Africa	64
Central Europe	35
Far East	50
Latin America	15
Middle East	40
North America	3
South Asia	48
Western Europe	0

attained military victory over Georgia during the the 2008 "war", militarized incidents continued almost uninterrupted following the cessation of military confrontations between the two states and the underlying issue of the political status of Abkhazia and South Ossetia remains unresolved. As such, the dispute is coded as ending in a stalemate in 2010.

Finally, it is important to note several challenges in verifying the validity of some reported incidents from the available news documents. Several news stories provide conflicting reports of events. For instance, some purported incidents and actions are protested by target states, but completely denied by initiator states. In these cases, coders err on the side of the target state's protest in an effort to capture the most incidents possible and avoid throwing out actual incidents based on states' denials. Additionally, several accounts of incidents and actions come from state-sponsored news agencies which provide conflicting reports of events, as was often the case during the ongoing conflict between Armenia and Azerbaijan. In these instances coders made every attempt to locate third-party news documents to acquire and verify unbiased accounts of events. When third-party news stories were unavailable, manual coders reconciled the details of such incidents to the best of their abilities. The coding of these incidents was later reviewed and verified by the entire MID staff.

The MID4 dispute and incident data

The MID4 project coded 255 MIDs that began between 1 January 2002 and 31 December 2010. When including MIDs that continue into this period from 2001 or earlier, the number of "active" MIDs between 2002 and 2010 increases to 262. Disputes coded during this period continue the pattern of some geographic regions being more contentious than others. Table 5 provides a breakdown of the number of MIDs for each region of the world. Continental Africa was, based solely on the total number of MIDs between 2002 and 2010, the most disputatious region of the world, followed by South Asia and the Middle East. North America and Western Europe were very peaceful in comparison, with three and zero MIDs, respectively.

Generally, the number of MID onsets appears to be decreasing over time. Figure 1 provides the yearly breakdown of new MID starts since 2002. The yearly average of new MID starts between 2002 and 2010 is 28.3, down slightly from the mean number of 32.9 starts between 1993 and 2001. There is a rather significant decrease in MID starts, however, beginning in 2006, as the average number of new MIDs decreases to 22.4. There are several possible explanations for this decrease in MID starts.[13] Among other explanations, it is possible that the international system is indeed somewhat more peaceful following 2005. It is also possible that the duration of active MIDs is increasing, such that existing MIDs are occurring for longer periods, rather than ending and beginning again as new disputes.

Following this insight, Table 6 provides the duration of disputes. The average duration of disputes following 2002 is about 122 days, lower than the average duration of 144 days in disputes prior to 2002. There is also a clear trend between the three periods of the MID project of an increasing proportion of disputes being shorter in duration. While this does not reveal anything about the underlying causes of the decrease in MID starts following 2006, it does imply that, continuing the findings of MID3, the MID project is doing a better job of locating and coding shorter and possibly less intense disputes.

In terms of the actors in MIDs, a majority of dispute actors did not seek to change the status quo between 2002 and 2010, as only about 38% of actors were revisionist, or seeking

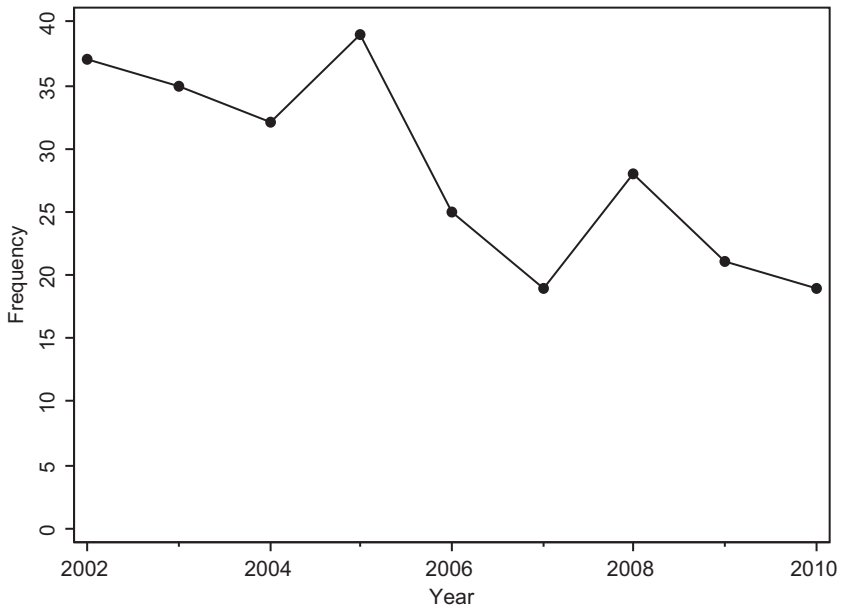

Figure 1. Annual frequency of MID onset, 2002–2010.
Note: MID3 covers the period 1993–2001, while MID4 covers 2002–2010.

Table 6. Dispute duration

	1816–1992	1993–2001	2002–2010
One week or less	38.72%	49.32%	53.33%
8–31 days	11.74%	9.12%	5.09%
32–61 days	8.21%	6.76%	4.31%
62–100 days	6.93%	5.07%	5.88%
101–365 days	24.37%	18.92%	24.31%
More than 365 days	10.02%	10.81%	7.06%

to change the status quo, in some way. However, most dispute initiators were revisionist, with over 55% of dispute initiators seeking to revise the status quo in some way. In contrast, a clear majority of Side B, or target, states in disputes were not revisionist. When a state is identified as revisionist, the most common revision type is policy, as roughly 81% of all revisionist actors sought to change the policy of the other state in the dispute. Furthermore, about 83% of revisionist initiator states were policy revisionists, as opposed to about 15% who were territory revisionists and under 3% who were regime revisionists. To illustrate this further, Table 7 offers a comparison of the breakdown of revisionist states between the 2002–2010 period and the previous two periods of the MID project.

Between 2002 and 2010, there was a clear pattern in which most disputes were either revisionist only on Side A of the dispute or not revisionist at all. In fact, the plurality of disputes was not revisionist by either side in the dispute. The smaller percentages of revisionist actors

Table 7. Revisionists and sides

	Side A	Side B	Neither	Both
2002–2010	102 (40%)	13 (5.10%)	104 (40.78%)	36 (14.12%)
1993–2001	140 (47.3%)	44 (14.9%)	66 (22.30%)	46 (15.54%)
1816–1992	1241 (61.0%)	206 (10.12%)	364 (17.89%)	224 (11.01%)

Table 8. Hostility levels and dispute participants

	1816–1992	1993–2001	2002–2010
No Militarized Action	24.46%	29.52%	29.70%
Threat of Force	4.77%	3.77%	1.32%
Display of Force	19.88%	30.28%	29.70%
Use of Force	44.57%	34.67%	38.91%
War	6.33%	1.76%	0.38%

correlates with the lack of reciprocation of disputes during this same period, as roughly 65% of disputes after 2002 were not reciprocated, that is, the target state took no military action against the dispute initiator, up from about 50% prior to 2002. Between the three periods of the MID project, the proportion of Side A revisions consistently decreased, while the proportion of disputes with no revisionist state increased significantly after 2001.

The overall level of hostility of disputes has remained fairly constant over time. Table 8 reveals a decline in Uses of Force and Threats to Use Force since 2002. However, the proportion of Displays of Force in the data has significantly increased since 2002. Adding support to the the notion that more disputes are unilateral and not reciprocated by the target state, the proportion of dispute participants taking no militarized actions increased slightly after 2002, while the proportion of Uses of Force has increased since 2002. Perhaps most notable is the absence of war in the MID4 data; aside from the invasion of Iraq by coalition forces in 2003, no disputes reached the level of hostilities or fatalities to be coded as interstate wars.

The MID4 Project maintained the incident coding rules developed in MID3, such that incidents are defined by being similar actions in similar locations within a 3 day time period. Incidents may have multiple actors; however, each incident may, by definition, only have one target. For disputes in which several incidents occurred in overlapping time periods, great care was taken to distinguish, when necessary, actions in disparate locations into separate incidents and, when necessary, to combine similar actions in proximate times and locations into single incidents. Table 9 provides a simple breakdown of the distribution of incidents across disputes in the MID4 data, and offers a comparison with the same breakdown in the MID3 data. About half of disputes between 2002 and 2010 are single-incident MIDs, and over 85% of disputes comprise four or fewer incidents. With some slight variation, the distribution of incidents per MID is relatively constant when compared with MID3, with the majority of MIDs in both time periods consisting of a single or relatively few incidents.

Table 9. Number of incidents per MID

Number of incidents	1993–2001	2002–2010
1	147 (49.66)	128 (50.20)
2	36 (12.16)	51 (20)
3	28 (9.46)	28 (10.98)
4	14 (4.73)	14 (5.49)
5	16 (5.41)	5 (1.96)
6	4 (1.35)	7 (2.75)
7	2 (0.68)	2 (0.78)
8	7 (2.36)	2 (0.78)
9	4 (1.35)	1 (0.39)
10	3 (1.01)	2 (0.78)
11	0 (0)	2 (0.78)
12	1 (0.34)	1 (0.39)
>12	34 (11.49)	11 (4.31)

Note: Cell entries report incident frequency and incident proportion (in parentheses).

Overall, there were 1224 incidents between 2002 and 2010. This is a sharp decrease from the 2092 incidents that occurred between 1993 and 2001. Although the percentage of single-incident MIDs remained fairly constant between MID3 and MID4 (about 49 and 50%, respectively), there were several highly incidentious MIDs in MID3, the scope of which was not repeated in MID4. For instance, the ongoing conflict between Israel, Lebanon and Syria (MID #4182) contained 391 incidents prior to 2002, and only 93 following 2002. Additionally, the dispute culminating in the Iraq War of 2003 contained 158 incidents prior to 2002 and only seven incidents following 2002. Overall, the distribution of incidents per MID is much more concentrated toward fewer-incident disputes in MID4 than in MID3, but when controlling for these outlier disputes, the distribution of incidents per MID between the two periods becomes significantly more congruent.

As Table 10 illustrates, the distribution of hostility levels at the incident level closely mirrors that of the dispute level. Once again, there was an increase in displays of force, and there was a decrease in the number of threats and uses of force. The reduction in observed threats to use force is by far the most drastic, with only about 13% as many threats recorded for MID4 as were recorded for MID3. A primary reason for this reduction is that a small number of disputes were responsible for a large proportion of the total observed threats in MID3. Specifically, four disputes account for 77 of the 171 (45%) of the threats recorded in MID3.[14] Even after accounting for these cases, there were still considerably fewer threats

Table 10. Distribution of incident hostility levels

Action	1993–2001	2002–2010
Threat to Use Force	171 (8.17)	23 (1.88)
Display of Force	1075 (51.39)	567 (46.32)
Use of Force	839 (40.11)	633 (51.72)
War	7 (0.33)	1 (0.08)

Note: Cell entries report Hostility Level frequency and proportion (in parentheses).

Table 11. Most incidentious states (ranked by number of incidents involved)

Country	Number of incidents involved	Percentage of incidents involved	Country	Number of incidents initiated	Percentage of incidents initiated
Pakistan	206	8.34	Israel	156	12.52
Israel	175	7.09	India	93	7.46
India	173	7.00	USA	92	7.38
Lebanon	157	6.36	Armenia	90	7.22
Azerbaijan	112	4.53	Pakistan	73	5.86
Armenia	108	4.37	Turkey	72	5.78
USA	102	4.13	Russia	64	5.14
Turkey	101	4.09	North Korea	50	4.01
North Korea	96	3.89	Iran	47	3.77
South Korea	90	3.64	South Korea	37	2.97

observed during MID4 when compared with MID3. We took several steps to ensure that this reduction was not caused by biases in our data collection procedures.

First, we sought to determine whether the research team for the MID4 project was coding threats differently from the MID3 project. To do this, we recovered the set of 1993–2001 news stories MID3 positively identified as threats and re-evaluated the coding decisions. We determined that, while there was usually agreement between each iteration of the project, the MID4 team may have adopted a more stringent application of the coding rules.[15]

Specifically, the language in 23 incidents identified as threats from the MID3 project was determined to be too ambiguous to be classified as a threat by the MID coding rules. These incidents were subsequently removed from the data.[16] Overall, however, this difference in the application of the coding rules appeared to be relatively mild, and did not account for the substantial reduction in the number of threats between MID3 and MID4.

Second, we considered the possibility that, when reducing the set of news stories to be manually coded, SVM may have done a particularly poor job in identifying stories about threats as relevant documents. One potential reason for this would be that SVM was trained on relatively few documents on threats, as these stories are relatively rare to begin with. To address this possibility, we added an additional 67 news stories about threats from MID3 to

the set of documents the SVM was trained on to identify relevant documents. The re-trained SVM did not, however, produce any additional news stories about previously uncoded threats.

Third, we re-examined the news stories containing the word 'threat' that were removed during the tilling and manual coding process. In doing so, we uncovered a handful of news stories that contained information about threats that had not been previously identified. Again, however, this number was quite small and did not seem to indicate any systematic error. After having taken these additional steps, we came to the conclusion that the reduction in threats appears to be a manifestation of an empirical reality—states seem to have issued fewer military threats (as defined by our coding rules) from 2002 to 2010 than they did from 1993 to 2001—and not a result of the data collection procedures undertaken here.

Table 11 provides the states that initiate and are involved in the most incidents between 2002 and 2010. Not surprisingly, Pakistan was involved in the most incidents during that period, being involved in ongoing disputes with both India and the USA, and accounting for almost 10% of the total incidents during the MID4 period. However, only about 35% of disputes in which it was involved were initiated by Pakistan, revealing that, while it is a very disputatious state, it is often the target of hostilities rather than the initiator. Similarly, Azerbaijan initiated only about 17% of the incidents in which it was involved, often being the target during its ongoing dispute with Armenia. By contrast, the USA and Israel each initiated about 90% of the incidents in which they were involved.

In addition to examining which specific states were involved in the most militarized action, it is also useful to examine how MII and MID participation varies over time as a function of major power status. As previously stated, MID4 differs from previous iterations of the MID project in its predominant reliance on global news sources for data collection. While our initial analyses produced no evidence of this, the concern remains that these sources may be biased in favor of reporting on powerful or influential states at the expense of information about minor power interactions. To assess this possibility, we report the proportion of total MIIs and MIDs that feature only minor power participants in Figure 2. While there is considerable annual variation in this proportion, it does not appear to differ meaningfully between the years covered by MID3 and MID4. We interpret this as additional evidence that reliance on global news sources does not lead to a systematic underreporting of incidents or disputes fought among minor powers.

The use and misuse of MIDs

In this section, we provide a brief overview of many of the research design choices that scholars are confronted with when utilizing the MID and MII data and make some recommendations regarding best practices. Many of the recommendations made in this section have been expressed in previous works (Beck et al., 1998; Bennett and Stam, 2000; Ghosn et al., 2003). We restate many of these issues here because these research design choices are often overlooked or under-discussed in analyses using the MID Data.

Fishing boat MIDs are MIDs nonetheless

The presence of fishing disputes in the MID data has sparked debate among users of the data. Disputes commonly referred to as "fishing boat MIDs" or "tuna boat chases" are

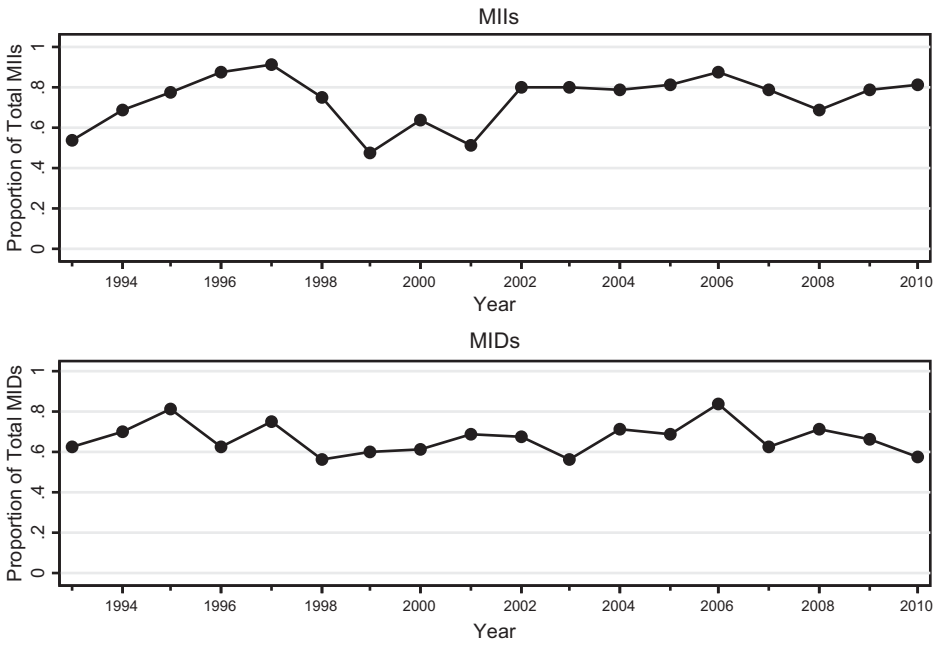

Figure 2. Proportion of MIIs and MIDs featuring only minor powers.
Note: MID3 covers the period 1993–2001, while MID4 covers 2002–2010. Major powers during this time period include China, France, Germany, Japan, Russia, the UK and the USA. All other states are classified as minor powers.

cases where states take militarized actions against civilian vessels, often because the latter are conducting economic activities in disputed waters. Some researchers have trivialized these disputes and argued in favor of their removal from the data under the contention that these disputes do not reflect serious issue incompatibilities between states. Others, more interested in investigating factors associated with escalation, may exclude fishing disputes as unlikely to result in escalation by either party. Both of these points are misguided. First, the fact that hostilities have arisen over fishing rights does not indicate the absence of a salient issue incompatibility between the states involved. Territorial and maritime disputes are fought primarily over where citizens from the involved countries have the rights to carry on private or economic activities. A "fishing boat" MID between Canada and Portugal in 2005 (MID #4550) was part of a protracted salient disagreement between those two countries that involved diplomatic intervention by the European Union. The use of force entailed by the Canadian seizure was neither casual nor trivial; it was a use of force that communicated effectively a political position. Similarly, some MIDs include boat seizures as part of a larger set of disagreements between states. For example, many of the incidents in the protracted dispute between Greece and Turkey over maritime and territorial holdings in the Aegean take the form of seizures of fishing boats. Much the same can be said of the dispute over the Senkaku Islands, with states using the seizure of civilian vessels as deliberate strategy to manage the underlying issue. As these cases demonstrate, it would be a mistake to assume that maritime relations are unimportant to international affairs.

Second, the contention that fishing disputes are qualitatively different from other forms of MIDs because they are less likely to escalate to war is dubious from a research design perspective. Excluding militarized disputes because they are less likely to escalate to war is a form of sampling on the dependent variable. One should not, therefore, systematically exclude fishing disputes on the basis of their probability of escalating to higher levels of hostility.

Disputes vs incidents

A key decision faced by researchers planning to employ the data presented here is whether to utilize the MID or MII data. Each dataset has relative advantages and disadvantages. One advantage of the MID data is its vast temporal domain, spanning from 1816 to 2010 after including this most recent update. The MID data also includes information about whether and how disputes are ultimately resolved. By construction, however, MIDs are composed of one or more discrete incidents. As a result, MIDs are often highly aggregated, and a dispute comprising hundreds of incidents over the course of several years and a single-day, single-incident dispute will each be recorded as "1" when a researcher is utilizing MID onset or initiation as the dependent variable. Furthermore, the 6 month cutoff rule described above means that a set of two incidents between the same actors over the same issue will be recorded as a single MID if they occur within 6 months of one another, but two MIDs if they occur 6 months and 1 day from one another.

MIIs, on the other hand, allow for a more fine-grained analysis of conflict onset. Unlike MIDs, a 3 day cutoff rule is employed for MIIs, meaning that similar actions taken by the same actors in a similar geographic region are considered to be the same MII only if they occur within 3 days of one another. This greatly ameliorates the problem of weighting multiple-incident MIDs and single-incident MIDs equally in statistical analyses. MII data also allows researchers to more closely track when states took actions to escalate international disputes, and periods when disputes became less intense. In many ways, then, the MII data is considerably more flexible than the MID data for researchers interested in examining conflict onset and escalation.

Directionality

Another choice facing researchers utilizing the MID and MII data is whether to employ the dyad or directed dyad. As Bennett and Stam (2000) note, the research question should be the guiding force behind which unit of analysis is chosen. If a scholar is interested in which factors are associated with the decision to *initiate* a militarized conflict, then the directed dyad is likely to be preferable, while if a scholar is interested in factors that determine whether a militarized conflict is likely, then the dyad is the appropriate choice.

Several issues should be considered when conducting directional analysis, prominent among them the meaning of MID "initiation". In particular, MID/MII *initiation* should not be conflated with MID/MII instigation, or with insitgating the disagreement within a MID/MII (Bennett and Stam, 2000; Ghosn et al., 2003; Palmer et al., 2004). The Side A variable that is sometimes used to identify one state as the initiator pertains only to the first state to take a military action.[17] This may or may not be the state that was largely "responsible" for the beginning of the political disagreement that underlies the MID; it may or may not be the state attempting to change the status quo via the militarized action; it may or may not

identify the state responsible for the beginning of militarized hostilities, or the state responsible for escalation at later stages of the dispute. In short, it is erroneous to ascribe properties to Side A outside its identification as the state that took the first codable militarized action in a dispute. Researchers should, therefore, carefully approach the theoretical construct in which they are interested, and only then determine whether MID/MII initiation adequately captures that construct.

Ongoing dispute years

Users of the data must decide how to deal with ongoing dispute years in their analysis.[18] Often, scholars are interested in conflict onset or initiation. When this is the case, scholars typically take one of three actions: coding ongoing disputes as "1"; coding ongoing disputes as "0"; or dropping years of ongoing disputes entirely.[19] Coding ongoing disputes as "1" in analyses is problematic to the extent one believes that dispute (or incident) continuation is distinct from dispute onset (Bennett and Stam, 2000). Coding ongoing dispute years as "0", on the other hand, might be considered justifiable because it is feasible, if uncommon, for states to become engaged in two simultaneous disputes. Nevertheless, this choice might be considered dubious, because in taking this action, a researcher is deciding to code a period of ongoing conflict between two or more states identically to a period of peace.[20] Furthermore, this rests on the perhaps tenuous assumption that the initiation of the second dispute is independent of the onset of the first dispute. Finally, scholars may choose to drop ongoing observations, under the belief that dispute onset, continuation and simultaneous dispute initiation are separate processes. Like Bennett and Stam (2000) and Beck et al. (1998), we believe that this is perhaps the most appropriate choice for the majority of analyses using the MID and MII data. Regardless of which option a researcher elects to take, it is critical that this choice be explicitly stated, along with the assumptions behind, and implications of, the choice.

Conclusion

The aims of the MID4 project have been straightforward. Militarized interstate conflict persists as one of the notable characteristics of the international system. To advance our understanding of the mechanisms driving such conflict, the MID4 project updated previous iterations of the Militarized Interstate Dispute datasets with data on militarized conflicts through 2010. While conflicts in the latest coding period conform to many of the overarching characteristics of MIDs in the historical record, several new and interesting trends in interstate conflict emerge. The MID4 project, like its immediate predecessor, performs increasingly well at recording shorter and lower-hostility disputes than did previous iterations. Substantively, threats to use force and occurrences of interstate war are much less common in recent periods as states increasingly favor displays and uses of forces short of war. Furthermore, participants in militarized interstate disputes appear to be becoming less revisionist, as MIDs appear increasingly to be directed at non-state actors occupying territory in other states.

In addition to the substantive impacts of the MID4 data, the MID4 project has sought to act as a mechanism of methodological innovation in the collection of observational data. To be sure, the automated document retrieval and classification of the MID4 project was a significant step forward in increasing the overall efficiency of the MID data collection process. However, while this hurdle was largely overcome, there remains significant room for

improvement in future iterations of the MID project. Following automated document classification, the relatively small number of "expert" manual coders trained in the nuances of the MID coding rules caused a notable log-jam in project efficiency in terms of time to completion. Even working at maximum productivity, the sheer number of events requiring coding vastly extends the time to coding completion. To circumvent this issue, future iterations of the MID project will attempt to utilize crowdsourced data collection to leverage large numbers of manual coders for increased coding efficiency.

Acknowledgement

At the time of this manuscript's submission, one of the authors was the journal's editor. The review process was therefore atypical and Ray Dacey agreed to serve as *de facto* editor: he recruited reviewers, protected their identities, passed the reviewers' comments to the authors, made the editorial decisions, and offered well taken editorial suggestions. The authors and the journal are grateful to Ray for his most professional assistance.

The authors are grateful to Scott Bennett, Alex Braithwaite, Faten Ghosn, Douglas Gibler, Steven Landis, Phil Schrodt, Ric Stoll, Zeev Maoz and three anonymous reviewers for helpful comments and contributing work on this paper, and to the many graduate and undergraduate students who helped with data collection during the course of this project. The authors are also grateful for the funding from the National Science Foundation, without which this work would not have been possible. All of the datasets described in this paper are available at the Correlates of War website.

Funding

The research reported in this paper was supported through grants from the National Science Foundation, specifically, SES-0719634 and SES-0924240.

Notes

1. The selected incidents and the spatial coverage of the news reports overlapped, either through actors or locations. On average, using the inclusive search string defined above, *global* sources retrieved 2365 stories per day, *regional* sources contributed an additional 266 stories per day, and *national* sources added a daily average of 320.
2. The full classification procedure is detailed in D'Orazio et al. (2014b), and therefore only briefly described here.
3. Phil Schrodt's XML file, CountryInfo.txt, is used to identify this set of relevant proper nouns. It is available at https://github.com/philip-schrodt/.
4. Replication materials for each step, as well as the related software, may be found in D'Orazio et al. (2013).
5. SVMs are, in some respects, not unlike linear regression. One way of thinking about the difference is that SVMs and regression have different objective functions: where regression minimizes the sum of the squared errors, SVMs maximize the distance between the hyperplane and the closest points in an effort to make the cleanest cut through the data.
6. A labeled document is a document that MID coders have read and found to be either "MID-relevant", a label of 1, or "irrelevant", a label of −1.
7. The primary difference between inductive and transductive SVMs is that the classification rule for inductive SVMs is learned using only the *training*, or labeled, data. The classification rule for transductive SVMs is learned using both the *training* and the *test*, or unlabeled, data. See Joachims (2002) for additional details.

8. We are very grateful to Ric Stoll and Doug Gibler for their assistance in mobilizing undergraduate students at Rice and Alabama, respectively, to participate in the MID Project. Additionally, selected undergraduate students at Penn State University coded some of the cases.
9. The MID4 staff was able to quickly come to agreement on the correct coding for the overwhelming majority of cases. We do note, however, that identifying the correct issue type typically took a greater degree of discussion than the other variables contained within data. This is because identifying what the substantive source of a particular incident or dispute is often requires a reasonable degree of interpretation, particularly when compared with the more objective characteristics of MIDs and MIIs (i.e. action type, actors, start and end dates, etc.).
10. An exception to this rule are disputes that begin prior to 2002, but continue through the MID4 coding period. These disputes notably include, among others, the ongoing dispute between Israel and Lebanon (MID #4182), the dispute that would eventually become the Iraq War of 2003 (MID #4273) and the conflict between India and Pakistan over control of Kashmir (MID #4277). All of these continuing MIDs have been retroactively coded with new identifiers.
11. Although we attempt to distinguish between US military drone strikes and strikes conducted by the CIA, it is often difficult to infer the source of drone strikes based on information in news reports. It is therefore highly possible that the MID data reflect several CIA drone strikes.
12. Although such actions were most prevalent between Ethiopia and Somalia, similar cases occurred between other dyads in the international system. The same rules were used to include or exclude these cases as well.
13. It is also possible that a dramatic increase in Israel–Lebanon incidents in 2006 caused a problem with the SVM document classification, as the SVM classification began to give greater weight to the presence of the words "Israel" and "Lebanon" in a story at the expense of other classifiers. After extensive testing and several re-classifications, this hypothesis was rejected.
14. These include MID #4182, #4137, #4273, and #4343
15. During this time, we also uncovered a small number of errors in the MID data, which were subsequently corrected. These included corrections of the maximum and minimum duration variables and the removal of a few duplicate observations from the incident level data. These changes are detailed in the "Summary of Changes as of MID 4.0" file made available at http://correlatesofwar.org/.
16. All figures reported here from the 1993–2001 period reflect characteristics of the data after having removed these incidents.
17. We stress that, for incidents that are coded as "clashes", the assignment of one or another state to "Side A" is often arbitrary.
18. Similar research design issues arise when using the MII data. Since the issue pertaining to MIIs has been explored less extensively, however, we leave the issue of ongoing MII years to future research.
19. When Bennett and Stam (2000) originally addressed this issue, they only considered the first and third options.
20. This issue is likely to be compounded when one only examines dispute initiation between originators. If one chose to only examine originator onsets, and coded all ongoing disputes as "0", for example, this would lead to coding the Russia–Germany dyad as being at peace (or equivalent to peace) throughout the First World War.

References

Abe S (2010) *Support Vector Machines for Pattern Classification*. Berlin: Springer.

Beck N, Katz J and Tucker R (1998) Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Journal of Political Science* 42(4): 1260–1288.

Bennett DS and Stam A (2000) Research design and estimator choices in the analysis of interstate dyads: When decisions matter. *Journal of Conflict Resolution* 44(5): 653–685.

D'Orazio V, Landis ST, Palmer G and Schrodt P (2013) Replication data for: Separating the wheat from the chaff: Applications of automated document classification using support vector machines. IQSS Dataverse Network. V1.

D'Orazio V, Kenwick M, Lane M, Palmer G and Reitter D (2014a) Crowdsourcing the collection of observational data. Paper presented at the *International Studies Association Annual Meeting*, Toronto, March 2014.

D'Orazio V, Landis ST, Palmer G and Schrodt P (2014b) Separating the wheat from the chaff: Applications of automated document classification using support vector machines. *Political Analysis* 22(2): 224–242.

Dumais S, Platt J, Heckerman D and Sahami M (1998) Inductive learning algorithms and representations for text categorization. In: *Proceedings of the Seventh International Conference on Information and Knowledge Management*, Bethesda, MD.

Ghosn F, Palmer G and Bremer S (2003) The MID3 data set, 1993–2001: Procedures, coding rules, and description. *Conflict Management and Peace Science* 21: 133–154.

Joachims T (1998) Text categorization with support vector machines: Learning with many relevant features. In: *Tenth European Conference on Machine Learning*.

Joachims T (2002) *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Norwell, MA: Kluwer Academic.

Jones D, Bremer S and Singer JD (1996) Militarized interstate disputes, 1816–1992: Rationale, coding rules, and empirical patterns. *Conflict Management and Peace Science* 15(2): 163–213.

Palmer G, London T and Regan P (2004) What's stopping you?: The sources of political constraints on international conflict behavior in parliamentary democracies. *International Interactions* 30: 1–24.

Schrodt PA, Palmer G and Haptipoglu ME (2008) Automated detection of reports of militarized interstate disputes: The SVM document classification algorithm. *Presented at the Annual Meeting of the American Political Science Association*, Toronto.

Vapnik VN (1995) *The Nature of Statistical Learning Theory*. New York: Springer.

Vapnik VN (1998) *Statistical Learning Theory*. New York: John Wiley and Sons.

Zhang T and Oles FJ (2001) Text categorization based on regularized linear classification methods. *Information Retrieval* 4(1): 5–31.