Improving Forecasts of International Events of Interest. *

Bryan Arva, John Beieler, Ben Fisher, Gustavo Lara,

Philip A. Schrodt, Wonjun Song, Marsha Sowell & Sam Stehle

The Pennsylvania State University

Version 1.0 : July 3, 2013

*Paper presented at the European Political Studies Association meetings, Barcelona, June 2013. The results and findings in no way represent the views of Lockheed-Martin, the Department of Defense, DARPA, NSF, or any person or institution associated with the ICEWS project. The ICEWS data analyzed in this paper is used in accordance with the terms of the Penn State subcontract 4100038729 and are from the research phase of ICEWS, not the current operational implementation. Contact author: Philip Schrodt, Department of Political Science, Pennsylvania State University, University Park, PA 16802 USA. Authors are listed in alphabetical order. The current version of the GDELT data set can be found at `http://GDELT.utdallas.edu`. Supplementary online materials for the paper can be found at `http://eventdata.psu.edu/papers.dir/EPSA13.SOM.dir/EPSA13.SOM.html`

Abstract

The paper compares the forecasting utility of the new GDELT—Global Data on Events, Location and Tone—dataset with some developmental versions of the ICEWS—Integrated Conflict Early Warning System—data using several alternative methods, including random forests, ADABoost, and Bayesian model averaging. Generally we find that the GDELT data performs as well or better than the data in the original ICEWS—quite possibly due to excessive attention in ICEWS to the eliminate of false positives, Kahneman's "what you see is all there is" pathology—and that these newer methods are quite promising as forecasting methods.

1 Overview

Prediction of political events has become more of interest in the present day as a result of the transition from interstate conflict to intrastate conflict since World War II. Given this rise in interest, the U.S. government has invested in two large-scale projects, the Political Instability Task Force (PITF) and the Integrated Conflict Early Warning System (ICEWS), which make use of quantitative data and statistical methods in order to forecast events of political instability.

The key difference between the ICEWS event data coding efforts and those of earlier NSF-funded event data efforts was the scale. As O'Brien—the ICEWS project director—notes,

> ...the ICEWS performers used input data from a variety of sources. Notably, they collected 6.5 million news stories about countries in the Pacific Command (PACOM) AOR [area of responsibility] for the period 1998-2006. This resulted in a dataset about two orders of magnitude greater than any other with which we are aware. These stories comprise 253 million lines of text and came from over 75 international sources (AP, UPI, and BBC Monitor) as well as regional sources (*India Today, Jakarta Post, Pakistan Newswire,* and *Saigon Times*).

The ICEWS data are composed of five indicators of political instability for twenty-nine countries in Asia:

1. Domestic Crisis
2. Ethnic Violence
3. Insurgency
4. International Crisis
5. Rebellion

Models developed in the project were able to predict conflict in Asia six months in advance with approximately 80% accuracy.

As ICEWS was expanded to a global scale, it became classified and limited to the U.S. government's use. However, with the availability of news texts on the internet as well as open-source software to gather and process these texts, a new global data set, the Global Data on Events, Location and Tone (GDELT) has become available (Leetaru and Schrodt, 2013). This event data set is coded using the TABARI coding engine (Schrodt, 2011) into the same CAMEO typology (Schrodt, Gerner and Yilmaz, 2009; Schrodt, 2012) as the ICEWS data,

but is many times larger, with more than 200 million events with global coverage from 1979 to the present. The GDELT data, which are updated daily, as well as assorted visualizations and scripts for working with the data, are available at `http://GDELT.utdallas.edu`.

Our goal in this study is to use GDELT to replicate and compare results to those produced using the version of ICEWS available during the research phase of that project. We seek to determine whether the additional scope and number of events in GDELT relative to ICEWS allows us to make more accurate and precise predictions. Following a few general comparisons of ICEWS and GDELT, we will use GDELT to predict the same response variables as used in the ICEWS models.

We will note at the outset that the major limitation of our analysis is the absence of a canonical unclassified version of the ICEWS data. Prior to implementation within the U.S. government, ICEWS utilized entirely unclassified source materials and methodologies, and the researchers on the project were assured repeatedly that the unclassified components would not only be available for academic research, but were actively encouraged to do pursue refereed publications using the data. This resulted in a number of papers using the data to be presented in the first several years of the project.

However, despite the recent "Open Data Policy" executive order from the Obama administration that unclassified data paid for with public money—which most certainly includes the development versions of ICEWS—mandating that "that, going forward, data generated by the government be made available in open, machine-readable formats, while appropriately safeguarding privacy, confidentiality, and security"[1] the ICEWS data have yet to be forthcoming. Consequently, we will be using the best data we have available at this time that has already been used in open presentations, with the expectation of re-doing the analysis when the data are finally released.

The organization of this paper is as follows: first we will show some basic visual and statistical comparisons between the GDELT and ICEWS data, particularly on the variables of interest. We will then cover the pre-processing of the GDELT data. This will be followed by a discussion of the geographic distribution of events and spatial forecasting, specifically conflict spillover detection. Our next section will cover model selection using Bayesian model averaging. Lastly, our modeling section will cover the application of different predictive techniques to both the ICEWS data and the GDELT data, with more emphasis on the latter.

[1] http://www.whitehouse.gov/the-press-office/2013/05/09/obama-administration-releases-historic-open-data-rules-enhance-governmen

2 Data Pre-processing

The GDELT dataset is simply a record of events in the international system over a span of years. Prior developing models, a subset of the data comparable to that in ICEWS needs to be created, then we aggregate this to a specific unit of analysis. Towards this end, the relevant countries were identified using the ICEWS dataset, found in Table 1, and a subset of of the GDELT data was created by pulling only events that had one of these country codes in the first three characters of either the source or target actors. This subset was then further reduced by obtaining only the relevant years, specifically 1997 to 2010. Following this, we identified the relevant variables the ICEWS dataset—Table 2—that are derived from the event data itself.

Table 1: ICEWS Countries

ISO Code	Country
AUS	Australia
BGD	Banglades
BTN	Bhutan
MMR	Myanmar
KHM	Cambodia
CHN	China
COM	Comoros
FJI	Fiji
IND	India
IDN	Indonesia
JPN	Japan
LAO	Laos
MDG	Madagascar
MYS	Malaysia
MUS	Mauritius
MNG	Mongolia
NPL	Nepal
NZL	New Zealand
PRK	Democratic People's Republic of Korea
PNG	Papua New Guinea
PHL	Philippines
RUS	Russia
SGP	Singapore
SLB	Solomon Islands
KOR	Republic of Korea
LKA	Sri Lanka
TWN	Taiwan
THA	Thailand
VNM	Vietnam

Table 2: ICEWS Variables

Variables			
gov_gov_vercp	gov_gov_matcp	gov_gov_vercf	gov_gov_matcf
gov_gov_gold	gov_par_matcp	gov_par_vercf	gov_opp_vercp
gov_opp_matcp	gov_opp_vercf	gov_opp_matcf	gov_opp_gold
gov_soc_matcp	gov_soc_vercf	gov_ios_vercp	gov_ios_matcp
gov_ios_vercf	gov_ios_matcf	gov_ios_gold	gov_sta_matcp
gov_sta_vercf	gov_usa_vercp	gov_usa_matcp	gov_usa_vercf
gov_usa_matcf	gov_usa_gold	par_par_gold	par_opp_gold
par_sta_gold	par_usa_gold	opp_gov_vercp	opp_gov_matcp
opp_gov_vercf	opp_gov_gold	opp_par_vercp	opp_par_matcf
opp_par_gold	opp_opp_vercp	opp_opp_matcp	opp_opp_vercf
opp_opp_gold	opp_soc_vercp	opp_soc_matcf	opp_soc_gold
opp_ios_vercp	opp_ios_matcp	opp_ios_vercf	opp_ios_gold
opp_sta_vercp	opp_sta_matcf	opp_sta_gold	opp_usa_vercp
opp_usa_matcp	opp_usa_vercf	opp_usa_gold	soc_gov_gold
soc_soc_gold	soc_ios_gold	soc_sta_gold	soc_usa_gold
ios_gov_gold	ios_opp_gold	ios_soc_gold	ios_usa_gold
sta_gov_gold	sta_par_gold	sta_opp_gold	sta_soc_gold
usa_gov_gold			

Consistent with the typical ICEWS modeling efforts, these variables are created by first identifying which type of actor is present in both the source and the target actors of an event. The possible categories for the actor types are:

- GOV - Government Actors; GOV, MIL, JUD, country code
- PAR - Political Opposition; OPP
- OPP - Militant Opposition; REB, INS
- SOC - Civil Society; EDU, BUS, MED
- IOS - International Organizations; NGO, IGO
- STA - International State System; country code
- USA - United States; USA

Using this information, each actor is coded into one of the above actor types. The variables are then derived from the combinations of these actor types, along with the type of event using the standard CAMEO "quad categories":

- *Verbal Cooperation* [VERCP]: The occurrence of dialogue-based meetings (i.e. negotiations, peace talks), statements that express a desire to cooperate or appeal for assistance (other than material aid) from other actors. CAMEO categories 01 to 05.
- *Material Cooperation* [MATCP]: Physical acts of collaboration or assistance, including receiving or sending aid, reducing bans and sentencing, etc. CAMEO categories 06 to 09.
- *Verbal Conflict* [VERCF]: A spoken criticism, threat, or accusation, often related to past or future potential acts of material conflict. CAMEO categories 10 to 14.
- *Material Conflict* [MATCF]: Physical acts of a conflictual nature, including armed attacks, destruction of property, assassination, etc. CAMEO categories 15 to 20.

As an example, if there was a material conflict event between two actors, the variable "gov_gov_matcf" would be coded as a one. Additional variables are drawn from the Goldstein values of the event (Goldstein, 1992). In other words, for an event between two government entities, the "gov_gov_goldstein" variable would be coded as the Goldstein-scale value of the event. Once these variables are coded for each individual event, it is necessary to aggregate the data into a usable form. Towards this end, the data is partitioned by country; if an event has either source or target actor that matches a country code, for example "AUS" for Australia, that event is considered to pertain to that country. For each country subset, the

variables are aggregated into monthly sums. Each of the country subsets is then recombined into a final dataset, a monthly time-series cross section.

The final step adds the appropriate dependent variables: these are the events of interest (EOIs) discussed above: International Crisis, Ethnic/Religious Conflict, Domestic Crisis, Rebellion, and Insurgency. For each of these dependent variables, both a three-month and six-month lag is created. This is to ensure that data from time $t-n$ is being used to predict events at time t. Once these lagged variables are created, they are merged into the existing time-series cross section data.

3 Direct Comparison of the Event Series

GDELT, like ICEWS—which is coded from a number of sources from Factiva—is based on multiple news sources, including all international news coverage from AfricaNews, Agence France Presse, Associated Press Online, Associated Press Worldstream, BBC Monitoring, Christian Science Monitor, Facts on File, Foreign Broadcast Information Service, United Press International, and the Washington Post. Additional sources examined include all national and international news coverage from the *New York Times*, all international and major US national stories from the Associated Press, and after 2003, all national and international news from Google News with the exception of sports, entertainment, and strictly economic news.

The approximate distribution of the events over time is shown in Figure 1, which shows the total size of the files by year. Unsurprisingly, given the very substantial changes over the past two decades in both the international news environment and the availability of news on the web, the density of the data is anything but constant, and shows a dramatic increase since the beginning of the twenty-first century. This increase is particularly dramatic after about 2003, which is when Google News begins to be developed, which triggers a more general proliferation of web-based news sources.

3.1 Descriptive Statistics

Our predictive models include a large number of variables, so in an effort keep this paper a reasonable length, we limit our descriptive focus to three variables that were highlighted as important predictors via Bayesian model averaging (Section 4.1). Those variables are the

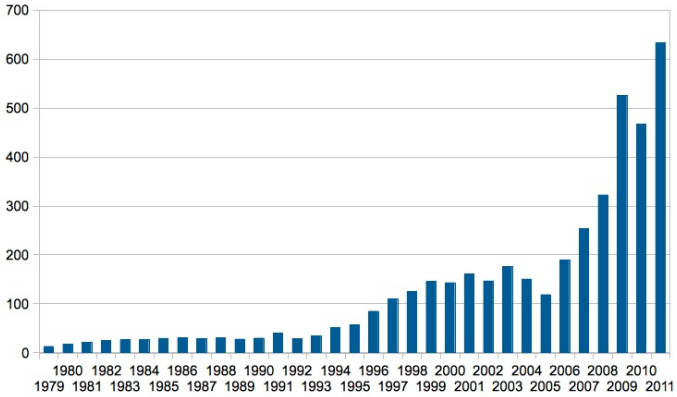


Figure 1: Distribution of GDELT events over time, Mb per year

following: government – opposition verbal conflict counts, government – opposition verbal cooperation counts, and government – opposition material cooperation counts.

The spatial distribution across datasets for these variables is shown in Tables 3 to 5. These consistently show roughly a 20:1 ratio of total event counts, with GDELT reporting many more events than ICEWS, though this differs considerably across country. For example, ICEWS seems to have very good coverage in Fiji and Cambodia but very poor coverage in Indonesia, China, Russia and Sri Lanka. In Cambodia, the ratio between ICEWS and GDELT is less than 1:2 in all three of these tables while the ratio for Fiji falls somewhere between 1:2 and 1:4. The trends across the three tables are fairly consistent so we discuss them together.

When we look at the ratios between the two datasets with regard to Indonesia, China, Russia and Sri Lanka, the numbers are even more worrisome. In Table 3 the number of events reported by ICEWS compared to GDELT for those countries are as follows: 4/503, 0/302, 0/1371, and 2/2737 respectively. This means that for those four countries, where we know there is a great deal of verbal conflict going on, ICEWS only has 6 events over a 12-year period. The enormous disparity across these datasets, for these three variables, seems to indicate that ICEWS has spatial coverage problems. Furthermore, since these four countries make up roughly half of the total events in the three tables , it is possible that they play a key role in our results. In future iterations of this paper, we would like to drop out some of these countries from our analyses and assess differences in predictive accuracy.

Table 3: Government – Opposition Verbal Conflict Counts per country

Country	ICEWS	GDELT
Australia	0	223
Bangladesh	28	159
Bhutan	0	46
China	0	302
Comoros	1	8
Fiji	42	97
Indonesia	4	503
India	47	510
Japan	14	99
Cambodia	256	363
South Korea	0	79
Laos	7	38
Sri Lanka	0	1371
Madagascar	2	32
Burma	15	237
Mongolia	0	3
Mauritius	0	1
Malaysia	0	109
Nepal	9	460
New Zealand	0	28
Philippines	21	1486
Papua New Guinea	1	26
North Korea	0	15
Russia	2	2737
Singapore	0	5
Solomon Islands	5	19
Thailand	14	392
Taiwan	2	29
Vietnam	0	33
Total	470	9410

Table 4: Government – Opposition Verbal Cooperation Counts per country

Country	ICEWS	GDELT
Australia	4	599
Bangladesh	22	445
Bhutan	0	60
China	0	710
Comoros	0	16
Fiji	96	329
Indonesia	7	1090
India	102	1185
Japan	20	760
Cambodia	608	867
South Korea	1	452
Laos	5	63
Sri Lanka	2	2595
Madagascar	0	15
Burma	34	528
Mongolia	0	4
Mauritius	2	10
Malaysia	2	462
Nepal	32	1411
New Zealand	0	97
Philippines	46	4059
Papua New Guinea	0	124
North Korea	0	88
Russia	5	4834
Singapore	0	29
Solomon Islands	41	69
Thailand	16	863
Taiwan	1	62
Vietnam	1	105
Total	1047	21931

Table 5: Government – Opposition Material Cooperation Counts per country

Country	ICEWS	GDELT
Australia	0	195
Bangladesh	14	120
Bhutan	0	25
China	0	408
Comoros	0	14
Fiji	20	75
Indonesia	3	287
India	26	274
Japan	14	103
Cambodia	110	139
South Korea	0	93
Laos	0	15
Sri Lanka	0	569
Madagascar	2	6
Burma	17	143
Mongolia	0	10
Mauritius	0	0
Malaysia	1	80
Nepal	5	388
New Zealand	0	18
Philippines	9	974
Papua New Guinea	0	12
North Korea	0	10
Russia	0	1257
Singapore	0	10
Solomon Islands	13	6
Thailand	8	212
Taiwan	2	20
Vietnam	0	17
Total	244	5480

Table 6: Government – Opposition Verbal Conflict Counts per year

Year	ICEWS	GDELT
1997	58	291
1998	42	253
1999	62	586
2000	66	993
2001	11	724
2002	15	747
2003	16	808
2004	17	678
2005	25	527
2006	26	697
2007	23	593
2008	14	819
2009	30	1161
2010	65	533
Total	470	9410

3.2 Times Series Comparison

The second thing that we examine is the distribution of event counts across time for the two datasets for the period 1997-2010. The same variables and time periods are used for the following three variables as were used for the tables looking at spatial distribution of the event counts. Thus, the totals are the same and have a ratio of about 20:1, with GDELT having the much larger number of events. As evident in Figure 1, the density of GDELT is not uniform and the news environment has changed a great deal over the period we are studying. Additionally, we also want to see if the two datasets have relatively consistent ratios of events counted from year to year. Once again, because we know that different collection techniques were used for these two datasets, we expect there to be some differences between them in terms of their coverage from year to year.

By looking at Tables 6 to 8, we can see that our instincts were correct. The differences are not as noticeable across time as they are across space, but they are present nonetheless. The variables used also do not exhibit the same trends across all of the tables as they did with regards to country coverage.

Table 7: Government – Opposition Verbal Cooperation Counts per year

Year	ICEWS	GDELT
1997	115	1120
1998	124	744
1999	120	1188
2000	164	2263
2001	60	1706
2002	54	1544
2003	69	1759
2004	59	1415
2005	62	1104
2006	46	2018
2007	43	1557
2008	37	1475
2009	47	2507
2010	47	1531
Total	104	21931

Tables 6 and 7 indicate that ICEWS has much higher event counts for verbal conflict and cooperation in the first four years of the dataset: 1997-2000. In Table 6 those four years have 4 of the 5 highest event counts while in Table 7 they make up the four highest event counts. This would not be a problem if the same trend was visible in the GDELT data but that is not the case. In fact, in the GDELT data, those years have some of the lowest event counts. We are unsure of why the disparity in event counts was so drastic for those four years, but it is definitely something that should be looked into further because uncovering the reason could help us better understand why the outputs from the two datasets are different. The rest of the years for the two aforementioned tables seem to be skewed, in terms of ratio, to be much higher for GDELT. The only other year that really stands out is 2009. In the GDELT data, 2009 has the highest event counts for verbal conflict and cooperation while, in the ICEWS dataset, the event count for 2009 is the lowest in Table 7. One would think there is a reason for this but, we are unsure of the explanation at this time. By looking at Table 8, we can see that the differences in coverage are not exactly the same across variables. While ICEWS does seem to have higher event counts for 1997-2000, the differences are not as great here as in other tables. However, the one thing that does stay consistent is that 2009 remains the year with the highest event count for GDELT while it is one of the lower ones for ICEWS.

Table 8: Government – Opposition Material Cooperation Counts per year

Year	ICEWS	GDELT
1997	18	204
1998	22	150
1999	13	314
2000	28	562
2001	21	420
2002	13	481
2003	11	396
2004	7	352
2005	16	254
2006	12	363
2007	6	392
2008	8	471
2009	10	736
2010	59	385
Total	244	5480

To further look at the issue of temporal variation, we compare the two data sets on four international dyads: China→Taiwan, India→Pakistan, South Korea→North Korea and USA→Japan. Our reference here is one of the later ICEWS data sets, labeled "Release 28" and coded either with JABARI or possibly some version of JABARI-NLP (Schrodt and Van Brackle, 2013); this would have been one of the last versions of the Asian ICEWS data before the project switched to development of the global W-ICEWS set. The data nominally go to Mar-2011 but the last couple of months have very low counts and may have been incomplete, so the series, which begins in Jan-1998, was truncated at Dec-2010.

Figure 2 shows the China→Taiwan comparison. The correlations—[0.024 , -0.16, 0.620, 0.275] in the order [VERCP, MATCP, VERCF, MATCF]—are relatively low except for the VERCF counts. GDELT has a higher density of events,about two to three times higher. VERCF is again the exception to this, roughly equal densities in that category.

These are not particularly good correlations. Three factors may be contributing to the divergence. First, most of the GDELT sequence being compared here is in the post-2000 period when GDELT is experiencing an exponential increase in density. Second, GDELT includes Xinhua, which ICEWS does not include, and Xinhua may be throwing off the totals when compared to the international sources. This in particular might explain why the VERCF (verbal conflict) indicator has the highest correlation: that would be consistent with Xinhua being used as a tool of the Chinese government's generally belligerent foreign policy towards Taiwan, and those policy pronouncements, in turn, would be monitored by

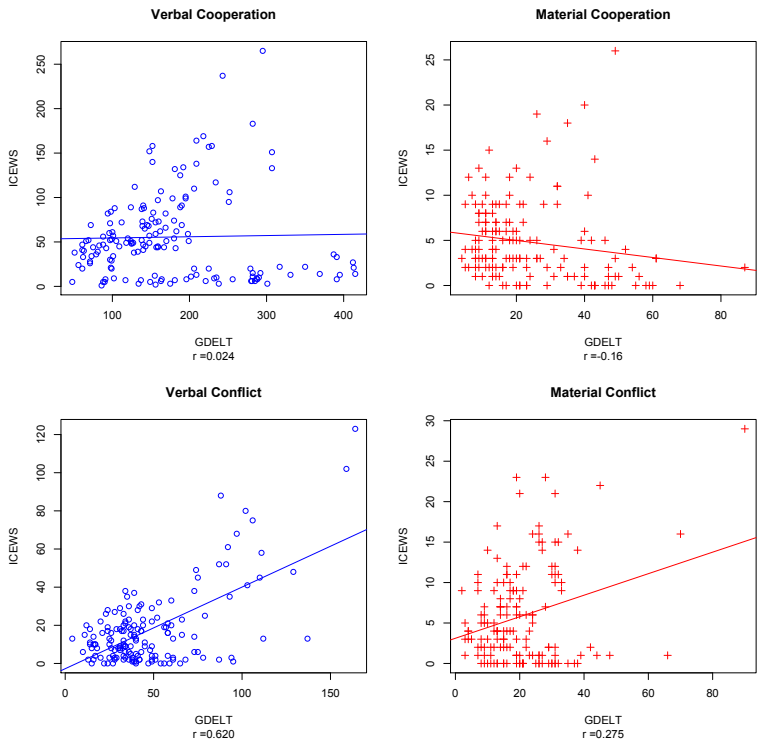


Figure 2: Comparison of ICEWS and GDELT Quadcounts: China→Taiwan 1998-2010

the international media.

Finally, there appears to be serious discontinuity in the latter part of the ICEWS sequence, which drops from reporting tens of events per month to ones of events. Using $VERCP >= 30$—that is, an average of at least one event per day—as a threshold, the correlations improve considerably, as shown in Figure 3. The vector of correlations here is [0.525, 0.350., 0.804, 0.583].

Figures 4, 5 and 6 show scattergrams for the remaining dyads: these generally show patterns similar to those seen in the China→Taiwan case. Figure 4 for India→Pakistan shows the same pattern of a high correlation [0.623] for VERCP—though this is clearly inflated by an outlying point which is similar in both data sets—and relatively low correlations [0.16 to 0.25] for the remaining counts; the ratio of the GDELT to ICEWS counts is again in the range of two to three.[2] South Korea→North Korea, Figure 5, has higher correlations, in the range [0.55-0.75] except for MATCP, though again these are inflated by outliers. The ratio of GDELT to ICEWS counts is substantially higher here, in the range of five to ten; again it is possible that Xinhua accounts for the difference. Finally, the USA→Japan dyad, Figure 6, has very low correlations, once again strongly influenced by the very low counts on all of

[2]We experimented with eliminating low frequency VERCP cases here and it did not make much difference.

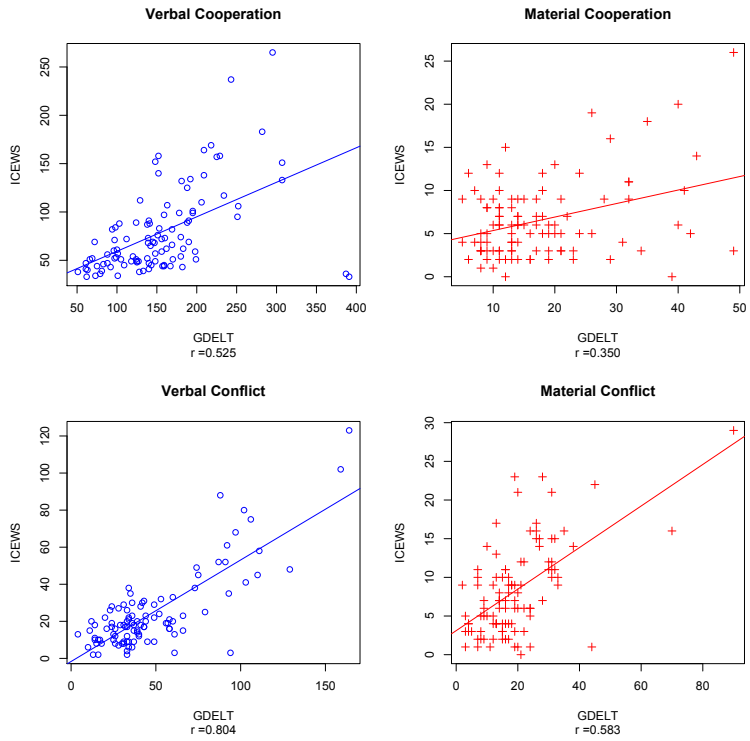


Figure 3: Comparison of ICEWS and GDELT Quadcounts: China→Taiwan 1998-2010 with $VERCP < 30$ cases removed

the ICEWS indicators except VERCP.

These comparisons clearly need to be explored in further detail, the most critical issue being further analysis to ascertain whether the difference in the counts is due to ICEWS being more selective—at various points, the project was working on making the coding very sensitive to false positives, particularly on conflict events—or whether GDELT is capturing more detail, particularly in the post-2000 period, because it is using a wider variety of web-based sources.[3] A systemic drop in reports in ICEWS at the same time GDELT is experiencing an exponential increase in reports would, of course, be a perfect storm for the sequences not correlating.

To further explore this possibility, we ran two-group t-tests on the 32 sequences (quad categories x dyad x GDELT/ICEWS), splitting the series at Jan-2005. Twelve of the sixteen tests showed significant differences ($p < 0.05$) for ICEWS, and the same ratio occurred for GDELT. However, the *direction* of these were quite different between the two data sets: of the twelve significant differences in ICEWS, ten were positive (the counts in the pre-2005

[3]An additional factor that might accounts for some of the differences is that fact that GDELT is using *location-based* duplicate filtering—events are considered duplicates only if they are the same event and occur in the same city—whereas as far as we know, ICEWS was using only *dyad-based* duplicate filtering.

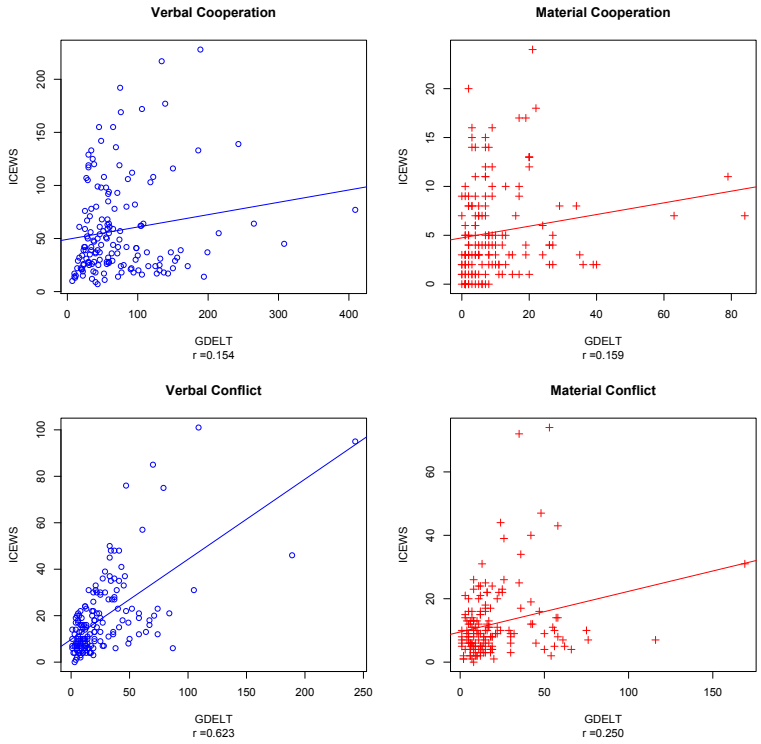


Figure 4: Comparison of ICEWS and GDELT Quadcounts: India→Pakistan 1998-2010

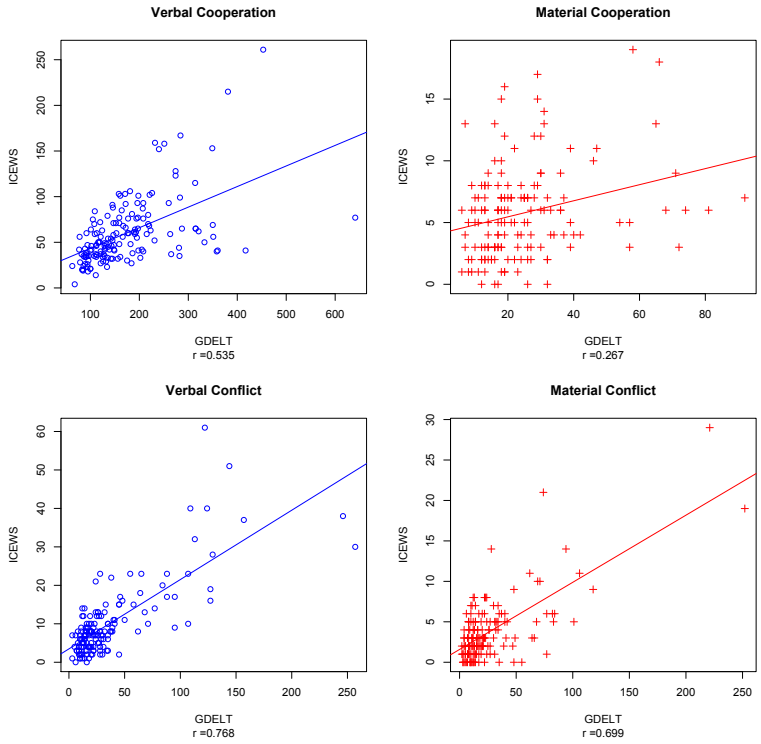


Figure 5: Comparison of ICEWS and GDELT Quadcounts: South Korea→North Korea 1998-2010

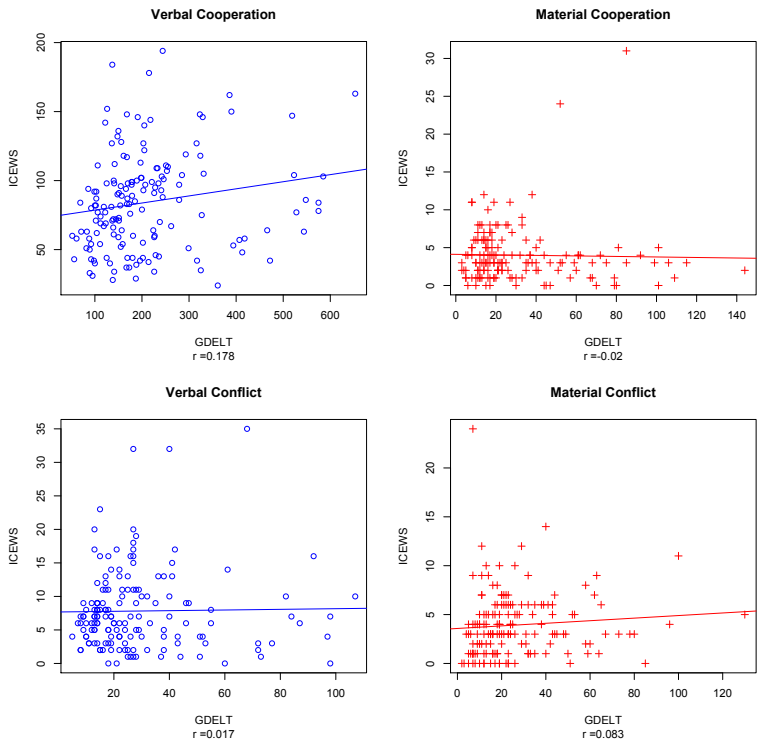


Figure 6: Comparison of ICEWS and GDELT Quadcounts: USA→Japan 1998-2010

were higher), whereas for GDELT only four were positive. Finally, except in the South Korea→North Korea (where all directions were the same) in all of the cases where both t-tests were significant, the changes were in opposite directions. All of this would suggest that much of the difference between the sets is due to changes in the baseline frequencies rather than the coding of specific events.

All factors being equal (though in this case, they clearly aren't equal due to differences in the source texts), the ICEWS data should be more accurate than the GDELT data because Jabari-NLP has some clear advantages over Tabari (Schrodt and Van Brackle, 2013). Our sense, however, is that there is more going on here than just the difference in coding engines, since those accuracy improvements are unlikely to have exclusively resulted in the elimination of events. ICEWS focus on the elimination of false positives would do exactly this, however, and that approach—rather than simply allowing the data to contain a "natural" mix of false positives and false negatives—may have significantly degraded the quality of the dataset, a possibility we will see reflected in further detail when we look at the performance of the two data sets in forecasting models.

3.3 Geographic Representation

The GDELT data presents many possibilities for understanding the geographic distribution of events. While events must take place in a defined location, the scale of that location varies widely depending on the type of event. Robberies take place at specific street addresses, demonstrations typically occur in public spaces on the order of hectares in size, new policy procedures may take place at the community level, while economic fluctuations may only be traceable to level of the state. Nonetheless, the ICEWS analysis aggregated everything to the nation-state level, even in cases of huge countries like Russia, China or Indonesia where violent conflicts tend to be isolated into the peripheries of the states, far from the population centers, and in the case Russia, usually occurs almost 10,000 kilometers from the Pacific region. Given these constraints in ICEWS, we will look at spatial trends by aggregating events to the country level despite the much higher level of geographical detail available in GDELT.

The most important variable, as described by our Bayesian Modal Averaging procedure (Section 4.1), was the one representing intergovernmental material conflict (gov_gov_matcf). As raw numbers of events in this category for each country indicate, a small number of countries consistently dominate the rest of the data in each recorded year, particularly Russia and China. To display the phenomena of interest, which is not raw values but proportion of conflict events, we normalized the total intergovernmental material conflict by the total verbal and material conflict and cooperation over the same time frame. Additionally, spatial trends at the monthly time scale would be difficult to find without 12 maps times 13 years of information to compare. Instead, we aggregated our monthly data to the quarter and compared the first quarter of each year to one another.

We performed two elementary analyses on the resulting quarterly data by country. The first was to generate a weighted mean center based on the gov_gov_matcf variable. A spatial mean center finds the most central point (which may not be a point in the original data) taking only spatial extent into account. By weighting the mean center with gov_gov_matcf, the mean center point shifts location based on the locations of features with high weights. We would expect to find trends in the movement of the mean center over time as conflict becomes spatially concentrated and/or shifts regions. While the mean center could be helpful in many cases, it would be unable to represent an instance where conflict concentrates in two geographically separate regions because those values would offset and the mean center would be weighted evenly in both directions. For that reason, we also generated a one standard deviational ellipse from the mean center location. The ellipse would indicate the

tendency of the gov_gov_matcf values to cluster around the mean center (collapsed ellipse) or spread more evenly through the spatial extent of the data (ellipse with a larger radius). The result is several maps of our Asia study area showing spatial concentration of the ratio of intergovernmental material conflict to total intergovernmental conflict and cooperation (see Figures 7 to 9 for a sample, additional maps available at request).

A visual examination of the yearly information shown on the map reveals very little useful information toward the goal of forecasting the spatial location or intensity of material conflict. No single countries exhibit any identifiable trends, nor do any regions appear as significant centers of conflict. The standard deviational ellipse shows no significant trends in size, except for indicating the relatively clustered appearance of conflict in the southern Asian island countries in 2004 by getting smaller. One significant conclusion can be made by observing that the mean center moves east and west, but stays fairly close to the same parallel. There is one likely explanation for this phenomenon: the conflict ratios in the extreme north and south countries (dominated by the presence of Russia and Australia) remain very constant across the years we have included in our analysis, while those countries within 20 degrees of the equator fluctuate at a much greater pace.

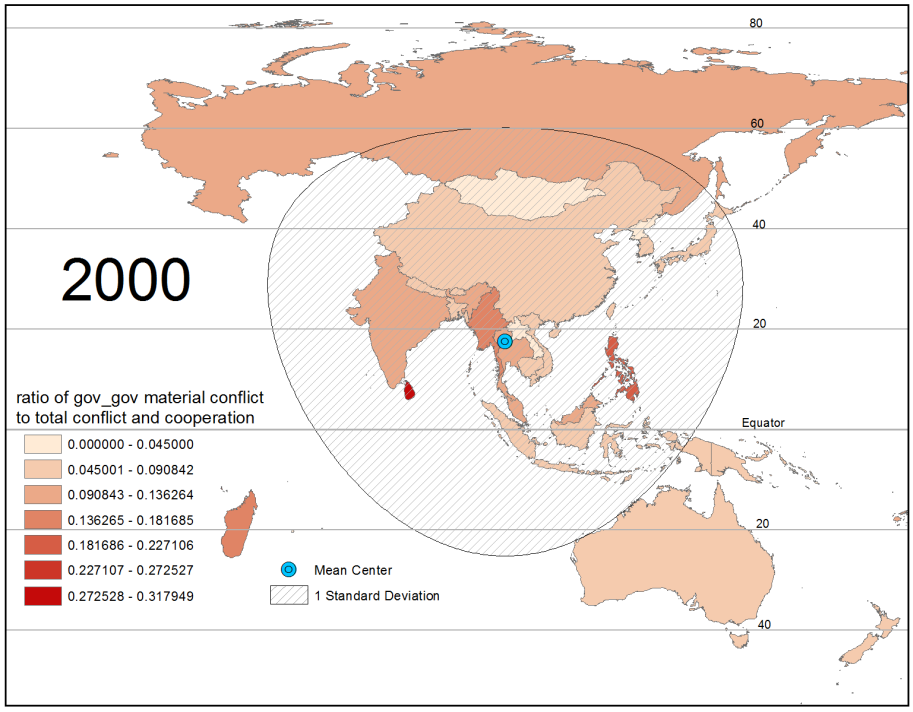


Figure 7: gov_gov_matcf as a ratio to total material and verbal cooperation and conflict.

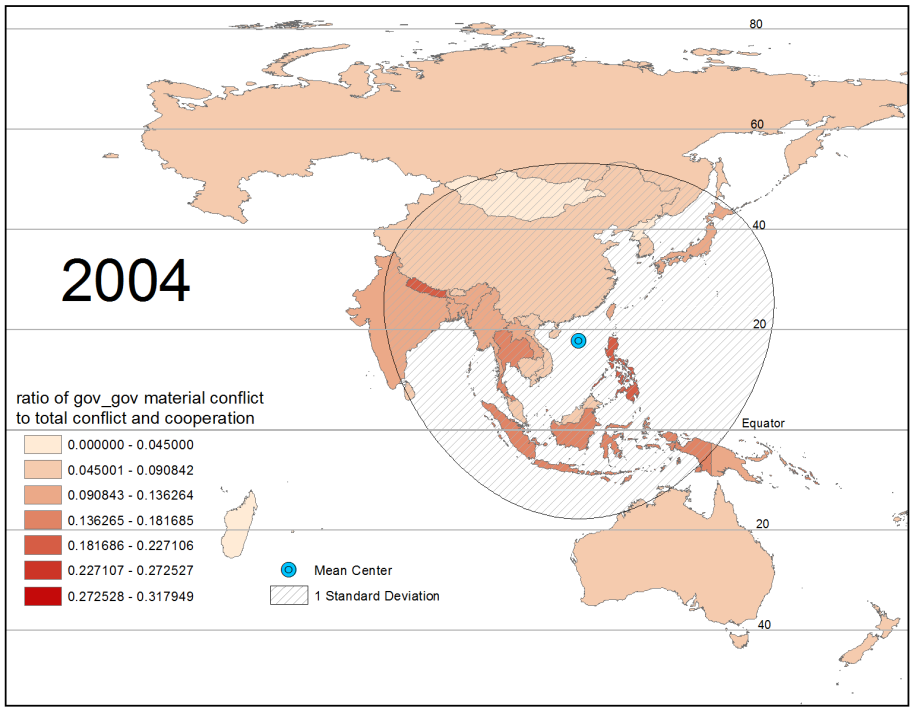


Figure 8: gov_gov_matcf as a ratio to total material and verbal cooperation and conflict.

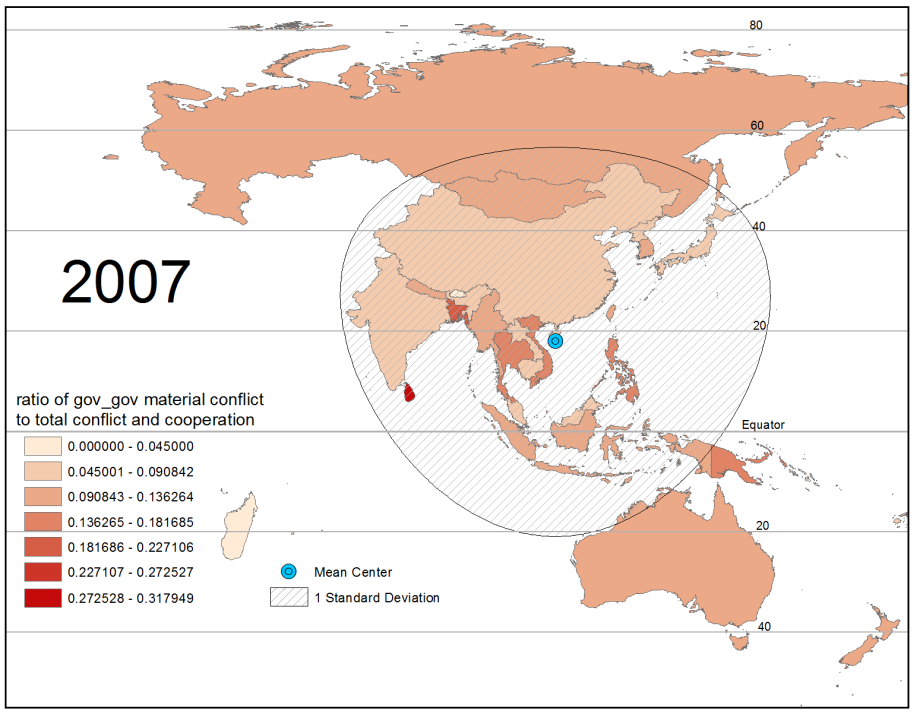


Figure 9: gov_gov_matcf as a ratio to total material and verbal cooperation and conflict.

3.4 Conflict Spillover Detection

The concept of "conflict spillover" could provide a better understanding of spatial trends of conflict: a working definition of "spillover" is that over time, the countries neighboring a country experiencing intense conflict will also experience similar conflict as it transcends the artificial borders of nations. Of course, geographical boundaries and island nations create a wrinkle to this theory, but for simplicity, we have ignored them here. We used two methods for detecting spatial trends applied to intergovernmental material conflict available in the ArcGIS software program: Anselin Local Moran's I for cluster/outlier detection, and Getis-Ord Gi* for hot-spot detection.

Anselin Local Moran's I (Anselin, 1995) identifies clusters of similar high or low values for a given weight as well as outliers that are significantly different from surrounding features. A cluster is defined by similar high or low values found within a critical distance interval, which is the defining factor of a neighborhood. To run the analysis, the minimum critical distance must be equal to or greater than the minimum distance such that each feature has at least one neighbor. This creates some problems in the GDELT data because island nations in Asia are significantly further from their closest neighbor than countries in the southeast region. Some samples maps resulting from the Moran's I technique are shown in Figures 10 to 12 (more are available upon request).

The first thing to notice about the output from Moran's I in this case is that clusters and outliers are somewhat rare. On the one hand, only the most significant clusters and outliers appear, but on the other hand, it is unclear whether they are a result of the data being quite regular or if it is a function of the defined critical distance. What does seem to be certain is that a lot of fluctuation is present in the material conflict variable of this data. Southeast Asia is consistently high, but on unpredictable time frames the level of conflict is high enough relative to surrounding countries to be considered a cluster. Also contrary to the assumption of visible conflict spillover, with the exception of the 2007 results, countries that are identified as clusters of high conflict are isolated—their neighbors do not share the same cluster designation.

Next, we attempted a process of identifying hot spots of conflict using the Getis-Ord Gi* statistic (Getis and Ord, 1996). Similarly to Moran's I, Geti-Ord requires a neighborhood to be defined by a critical distance, and we used the minimum such that every country has at least one neighbor. A country is designated as a hot spot by this process when the weight associated with it is high or low and the weights of its neighbors are similarly high or low. This makes the definition similar to the cluster designation of Morans I, but removes the

binary of 'cluster or not' by assigning a score to the country as a function of the standard deviation from the mean rate of conflict in the data. Sample maps are displayed in Figures 13 to 15 (and more are available by request).

Clusters that span multiple countries are much more visible in this analysis, as are some trends in the concentration of conflict hot spots. For the images above, a recurring cluster ranging from India to southeast Asia is present. This is not true in every year, but consistently this region appears together as a hot spot of conflict. This may, or course, be a product of the defined critical distance, but does seem to indicate similar values among these countries as opposed to the rest of Asia. One other thing that is obvious following the Geti-Ord procedure is that hot spot of conflict are much more present than cold spots (which would indicate centers of a lack of conflict). This would seem to be important in indicating that countries experiencing a lack of conflict are much more isolated, and as a result, have less tendency to spill over into neighboring countries. This hypothesis warrants further examination.

Unfortunately, at the nation-state level of aggregation used in ICEWS, the trends in concentration of intergovernmental material conflict appear to be unpredictable and therefore of little use to the goal of forecasting the locations of future conflict centers. Some form of more disaggregated spatial time series analysis may be necessary to quantify what appears visually in these generated maps.

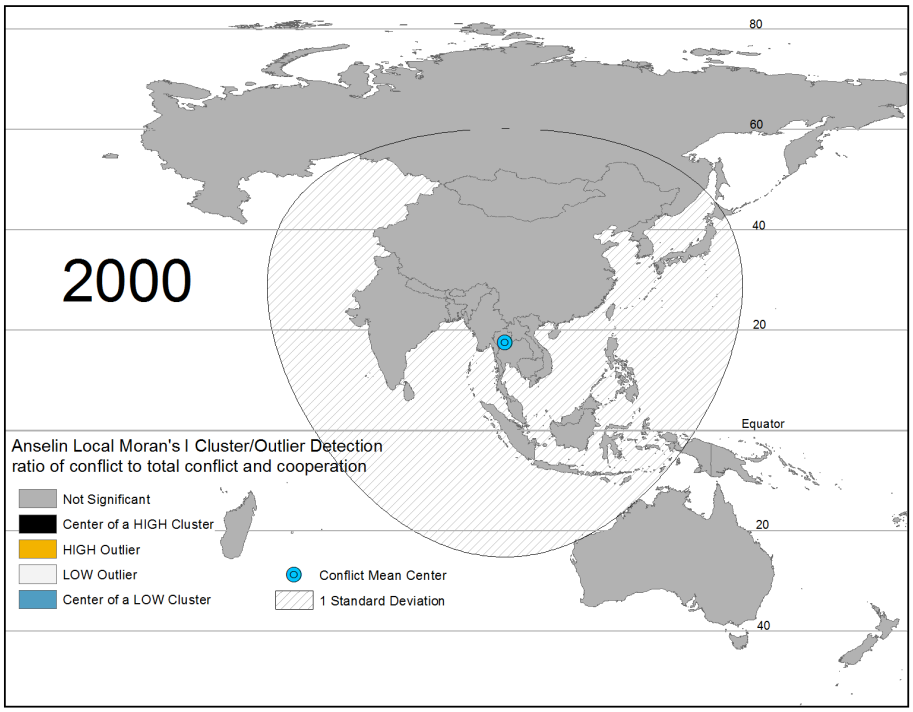


Figure 10: Clusters/outliers of gov_gov_matcf variable ratio

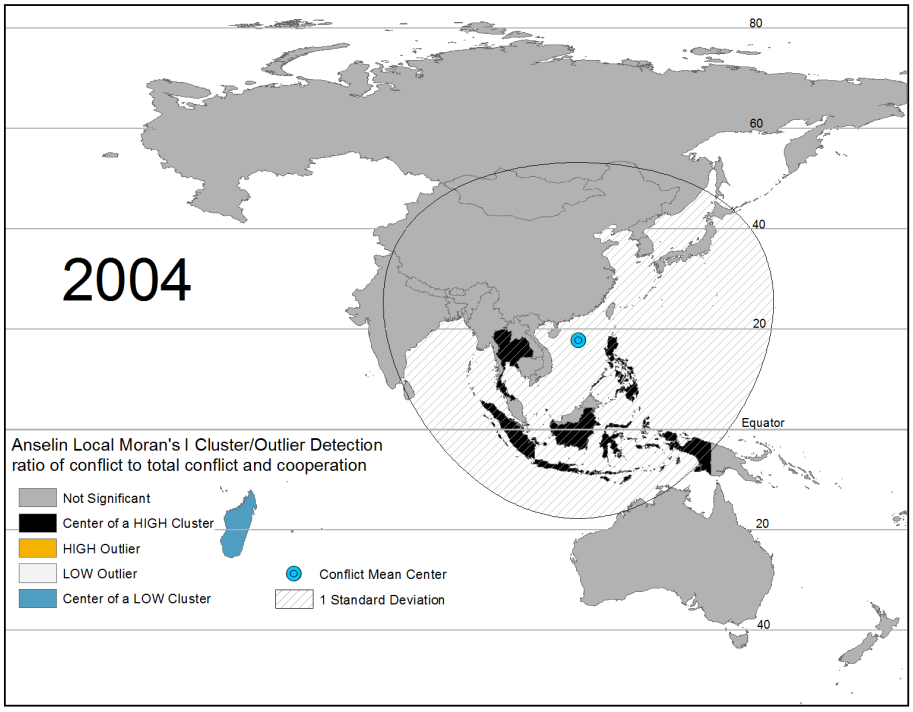


Figure 11: Clusters/outliers of gov_gov_matcf variable ratio

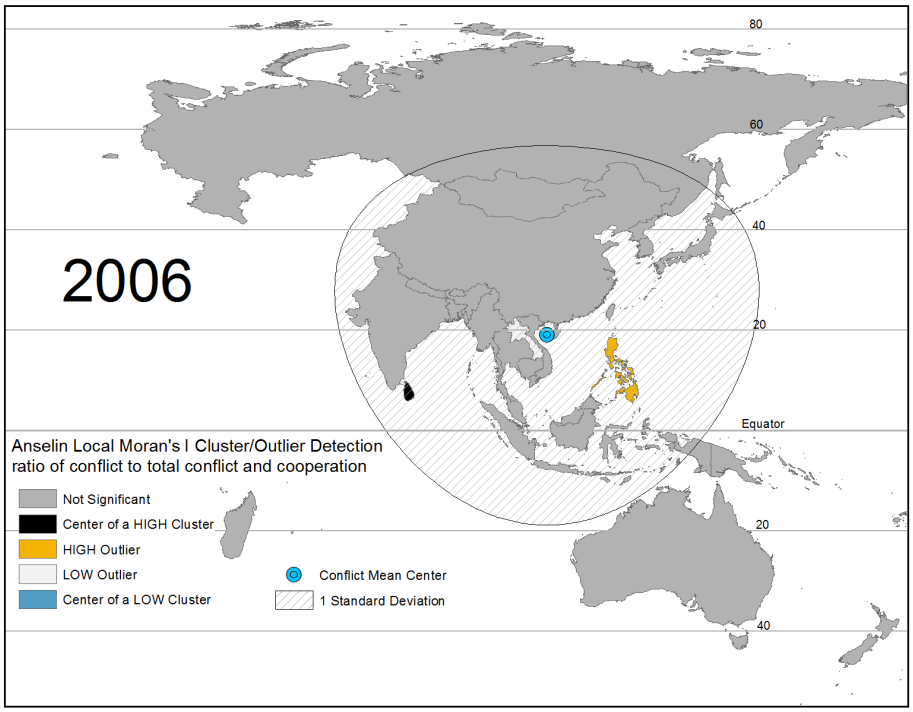


Figure 12: Clusters/outliers of gov_gov_matcf variable ratio

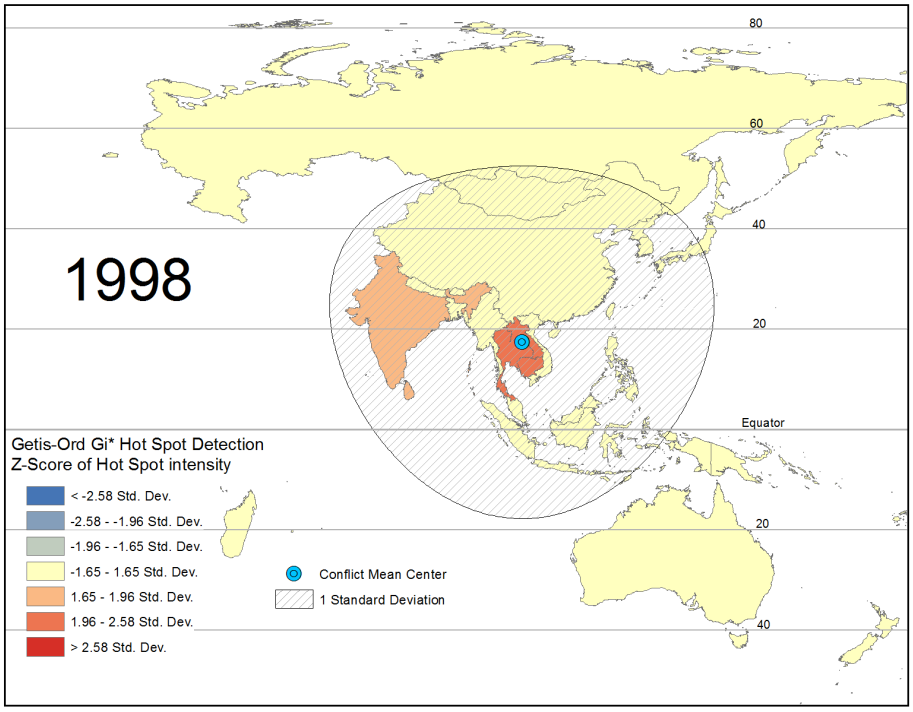


Figure 13: Hot/cold spots of gov_gov_matcf variable ratio

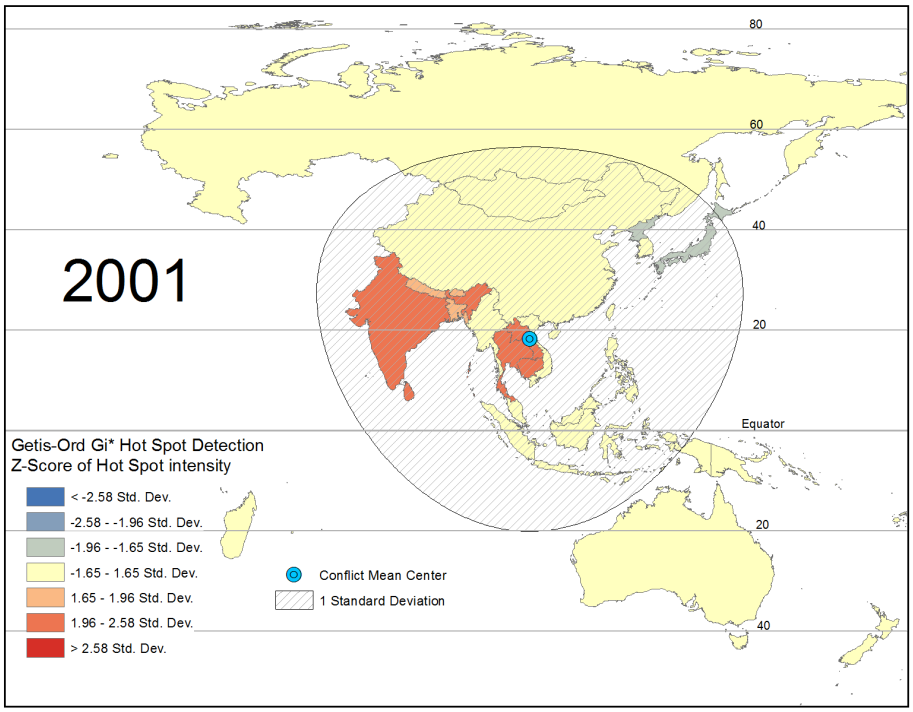


Figure 14: Hot/cold spots of gov_gov_matcf variable ratio

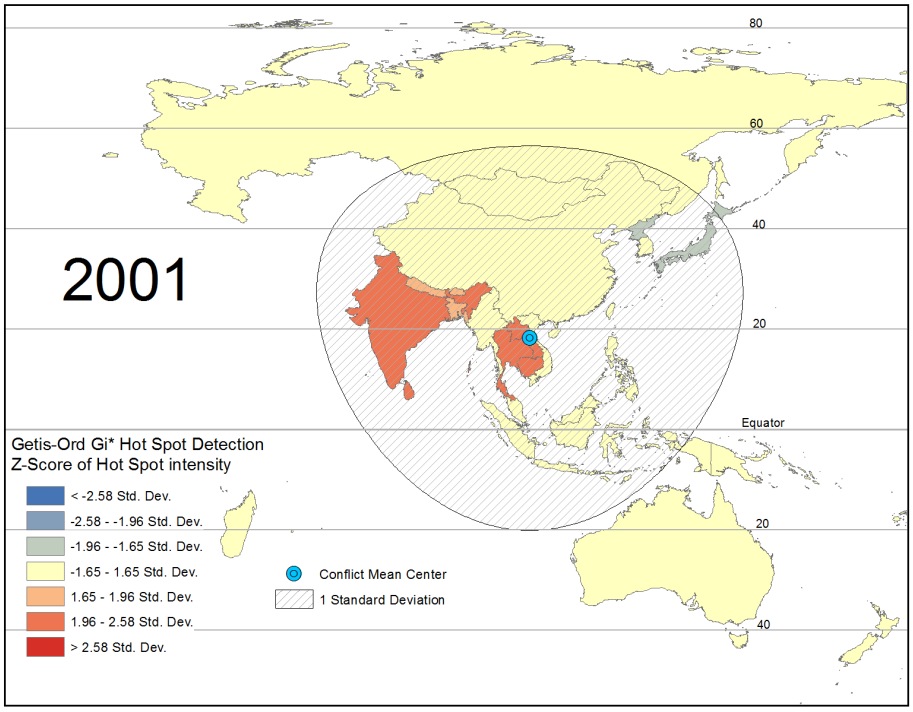


Figure 15: Hot/cold spots of gov_gov_matcf variable ratio

4 Model Selection

4.1 Bayesian Model Averaging

The ICEWS models contain several dozen variables, but which ones actually contribute as predictors? In order to assess this, we use a statistical method known as Bayesian model averaging (BMA; Bartels (1997); Montgomery and Nyhan (2010); Montgomery, Hollenbach and Ward (2012)). BMA assesses the relative importance of variables by running 2^k model combinations using different independent variables and then assigns a posterior probability to those variables' coefficient estimates.

For the sake of space in this paper, we only present the results for rebellion and international crisis; results for the remaining EOIs are comparable and are available from the authors. We estimated the models using the R package 'BMS' on its default settings (Feldkircher and Zeugner, 2009; Zeugner, N.d.). We assigned a uniform model prior to both as well. In addition to estimating the event variables from the GDELT data, we include an additional model where we estimate these variables along with the non-event ICEWS variables (GDP, population, etc.). In both cases, several of the non-event variables become as important as the top event variables from the previous model. The event-only model and the full model contain 68 and 78 variables respectively, which makes including a full table of results impractical. Instead, we include visualizations, such as Figure 16, which following the standard conventions for the BMA results, are ranked in order of the variable importance as whether the effect is positive or negative (red = negative, blue = positive).[4]

Unsurprisingly, gov_opp conflict variables are in models of rebellion and have a positive effect, meaning they make rebellion more likely. Gov_opp cooperative events are important as well and have a negative effect. This is likely due to negotiations between government and rebel groups to bring an end to rebellions. When the non-event variables from the original ICEWS data are included, all except size and trade levels have a high level of importance. Of these, only per capita GDP and being a primarily commodities exporter have negative coefficient estimates.

Many of these non-event variables make theoretical sense. A country with noncontiguous territory (i.e. the Philippines) would be ripe for rebellion due to the fractured nature of the state and the fact that islands would provide natural bases for rebel groups. Additionally, a country with a high level of ethnic fractionalization might be prone to rebellion by ethnic

[4]In order for the variable labels to be readable, these figures only include the variables with the highest probabilities; figures with all of the variables can be found on the supplemental online materials web page.

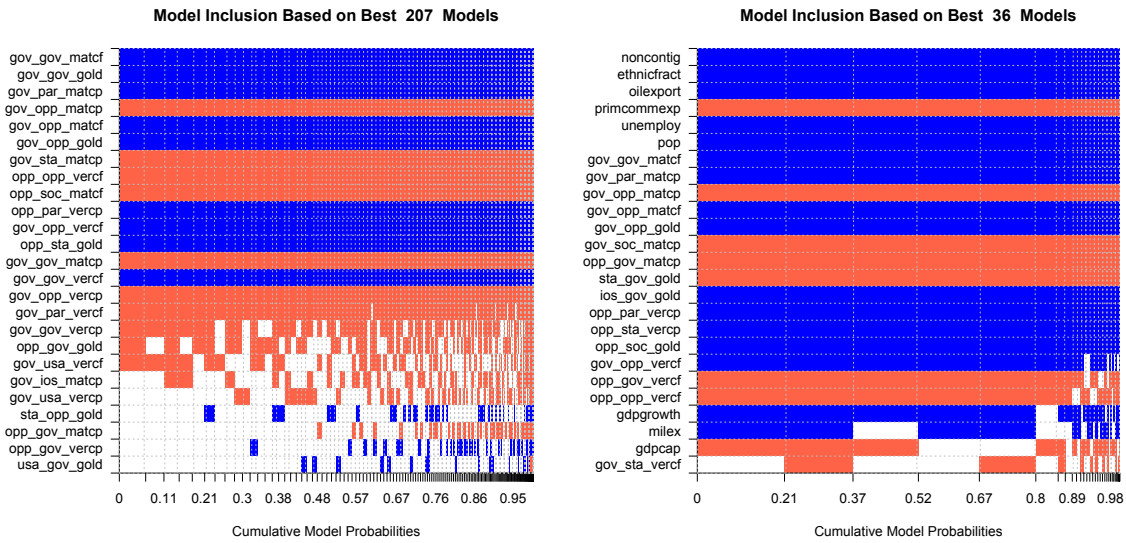


Figure 16: Variable Inclusion for Events Model and Full Model (Rebellion)

groups who may feel marginalized or wish to secede and form their own state. The fact that BMA considers ethnic fractionalization more important than the GDP variables and that it has a larger coefficient estimate runs contrary to the findings of the well known result in Fearon and Laitin (97). Figure 17 gives the posterior model distribution sizes for the events-only and the full models. The majority of the events models contain between 16 and 25 variables, while the full models contain about 6 more on average. This suggests that the non-event variables do not displace the important event variables, but are best included along with the event variables in models of rebellion.

The next event of interest we examine is international crisis, which is conflict between two states.[5] Before proceeding to the results, we must note that due to an error while subletting GDELT, we have neither the gov_sta_matcf variable nor the gov_sta_vercp variable. Despite this, we are more or less certain that they would both be in the majority of the models, considering that gov_sta_matcp and gov_sta_vercf are in the models. The process here is autoregressive. Conflict is likely to be followed by more conflict, so verbal conflict and (probably) material conflict are going to be good predictors. Two other interesting variables that have importance are gov_gov_vercf and gov_gov_vercp. The latter has a negative coefficient estimate and the former a positive one.

The likely story here is that states where government actors are bickering amongst themselves either do not have the inclination or the the capacity to initiate a conflict with another

[5]In the absence of a codebook, it is unclear these ICEWS EOIs include verbal threats or not

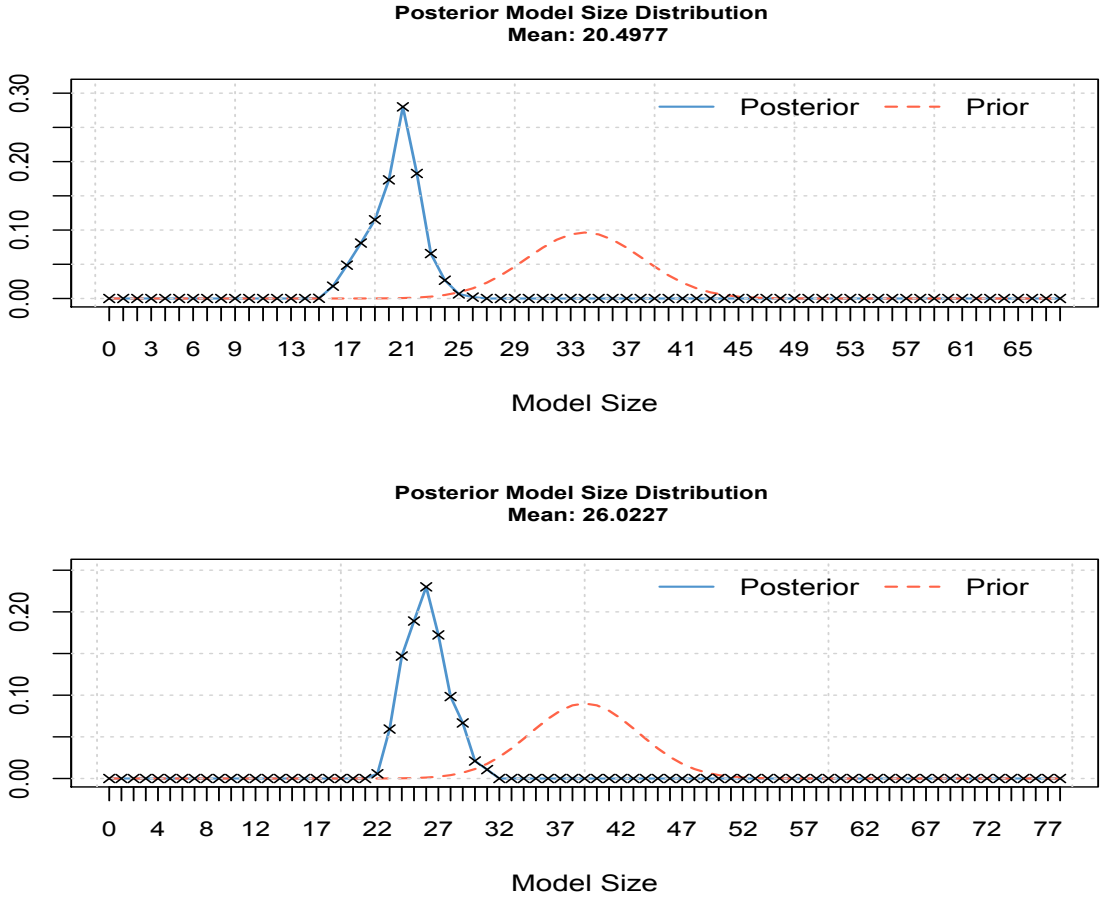


Figure 17: Posterior Model Size Distributions for Events Model and Full Model (Rebellion)

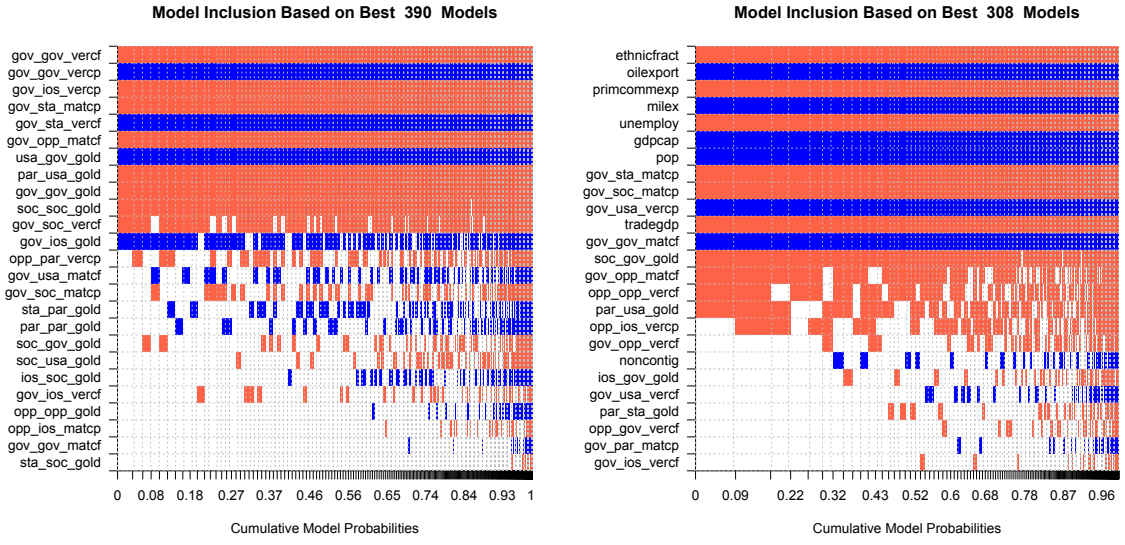


Figure 18: Variable Inclusion for Events Model and Full Model (International Crisis)

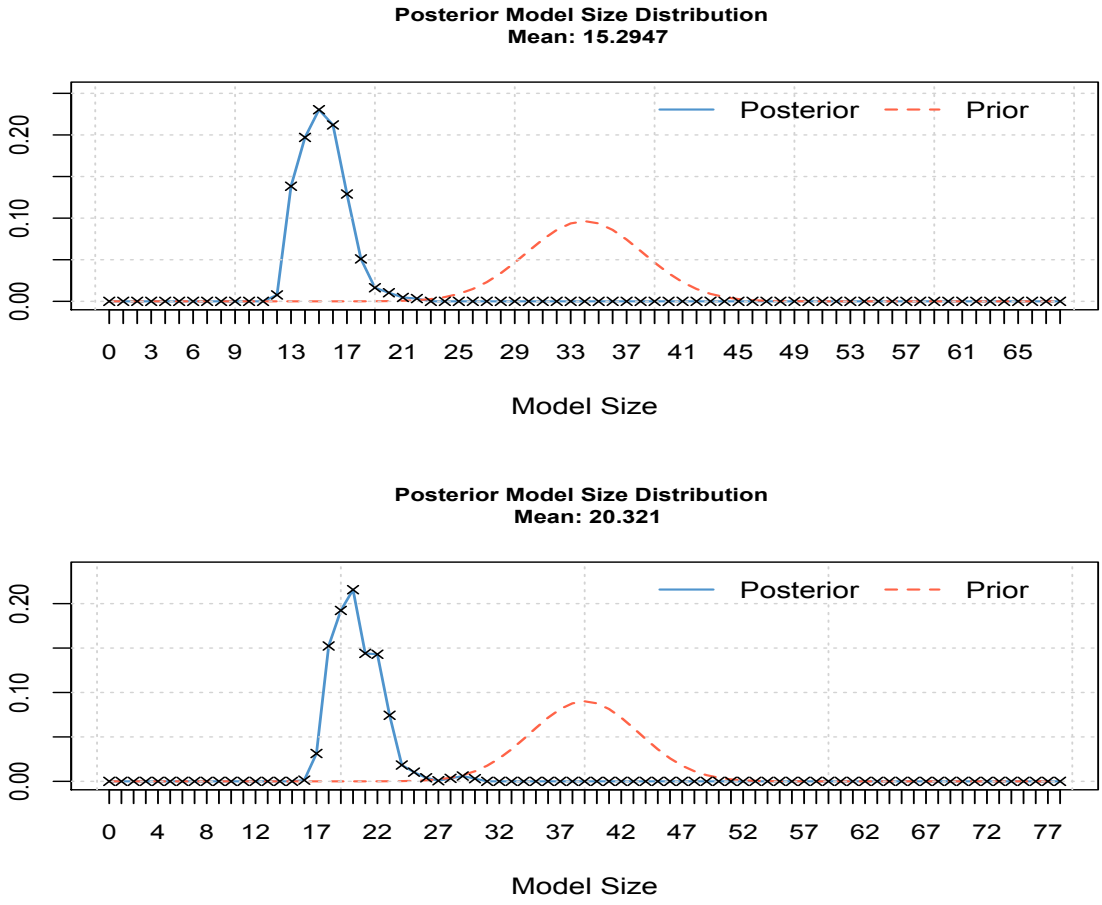


Figure 19: Posterior Model Size Distributions for Events Model and Full Model (International Crisis)

state. Similar to rebellion, most of the non-event variables are present across the models. Unlike rebellion though, some of the event variables actually become less important when the non-event variables are added. This is particularly true for gov_sta_vercf which drops from a PIP score of 1 to a PIP score of about .04. Figure 19 contains the posterior model distribution sizes for the events-only and full models of international crisis. The events models contain 12 to 21 variables, while the full models contain about 4 more variables on average. Once again, non-event variables are important when developing forecasting models for international events.

In summary, the Bayesian model averaging supports two conclusions. First, despite the very large number of event interaction variables which are available in the ICEWS set, a small and fairly consistent number are sufficient for most of the forecasting. Second, it is critical to include structural variables like GDP and ethnic fractionalization in our forecasting models.

The benefits of including the structural variables can be seen in our forecasting results in the next part of the paper.

5 Modeling

In this section we will compare the existing ICEWS data and GDELT using two standard machine learning approaches—random forests and adaptive boosting (ADABoost)—to develop the forecasting models. We first fit these new predictive models to the existing ICEWS data in order to determine if any gains in predictive accuracy are possible. Second, we create predictive models using the GDELT data as a basis, and compare the accuracy on the various EOIs with the accuracy when using the ICEWS data.

Briefly, a random forest is a ensemble method that combines numerous decision trees in order to create predictions for a given set of data. Each decision tree is passed both a bootstrapped sample of the data, along with a random selection of the variables. Each decision tree then chooses splits based on this data to minimize some error metric, often the classification error rate or Gini coefficient (Hastie, Tibshirani and Friedman, 2009). Each tree then generates predictions, or predicted probabilities, for each observation. These predictions are then combined using some approach, often the average, to generate one final prediction.[6]

Adaptive boosting makes use of an iterative fitting process as opposed to an ensemble of many different models. In each iteration, the algorithm creates predictions for each observation. These predictions are compared to the true values of the observation, creating an error rate. The error rates are compared and those observations with greater error rates are up-weighted in the next iteration. The algorithm can make better predictions after each iteration as weights are chosen that can better predict the events.[7]

5.1 ICEWS

In this section, various algorithms are run on the original ICEWS dataset to see whether we can obtain results better than the "official" ICEWS models. The Events of Interest (EOIs) used here are same as the GDELT models (*Rebellion*, *Insurgency*, *Domestic Crisis*, *Ethnic or Religious Violence*, and *International Crisis*).

[6]For additional detail: `http://en.wikipedia.org/wiki/Random_forest`. Strictly speaking, "random forest" is trademarked and should read "random forest^{TM}". Just like "Pet Rock.^{TM}"

[7]For additional detail: `http://en.wikipedia.org/wiki/Adaboost`. "ADABoost" is not trademarked.

Each EOI has three sets of results: (1) the original ICEWS logistic regression model, (2) the Random Forest model with 200 trees, and (3) the Adaptive Boosting model. All results are conducted as out-of-sample tests. The models are trained using data from 1998 to 2004 and we are predicting on 2005 and 2006. The ICEWS model uses a cutoff point of *Predicted Probability* = .50 to predict an event as occurring (*Pr(EOI = 1)*). We will report the classification tables for each EOI along with the classification error rates, in addition to some figures that compare the performance of the different models. The classification error rates aim to measure misclassification rates of the models. The equation for the classification error rates is:

$$ClassificationErrorRate = \frac{(FP + FN)}{(TP + TN + FP + FN)} \tag{1}$$

where TP and FP stand for true and false positives (the model predicted event but it did not happen), while TN and FN stand for true and false negatives (the model did not predict the event but it happened). Note that this is $1 - classification\_accuracy$, so lower values of CER indicate a better fit.

These are the three sets of results for the *Rebellion* EOI. The first table reports results for the ICEWS model, the second for the random forest model, and the third reports results for the Adaptive Boosting model. The classification error rates are reported at the bottom of each table. The original ICEWS model has predicted 24 rebellion events and 184 non-events correctly. There are 15 false positives and 9 false negatives. In terms of the error rates, the random forest does worse than the original ICEWS model, while using Adaptive Boost gives about the same predictive power as the original model.

As Figure 21 shows, applying random forests and adaptive boosting to the *Insurgency* EOI lowers the classification error rate by large margins when compared to the ICEWS model. The classification error rates for these two algorithms drop by more than 50%. While the original ICEWS model has 24 false positives, the random forest has 0 and Adaptive Boost only has 2. There is not much difference in terms of number of false negatives.

The results for the *Domestic Crisis* EOI, shown in Figure 22, demonstrate that the random forest and adaptive boosting do decrease the classification error rates, but not as great as when applied to *Insurgency* events. The models do a better job in reducing the number of false positives than the number of false negatives, however. In terms of the overall error rate, adaptive boosting performs better than the random forest.

For *Ethnic or Religious Conflict*, the results demonstrate that the random forest and adaptive

boosting do not perform much better when compared to the original ICEWS model. While the new models have decreased the number of false positives, the number of true positives has decreased and the number of false negatives has increased. Taken together, this causes the classification error rate for the various models to differ only slightly.

The new models perform better when applied to *International Crisis*, as shown in 24. The random forest decreases the error rate by a greater margin than adaptive boosting, but not as much as in the *Insurgency* models. Both random forest and adaptive boosting decrease the number of false positives but have mixed results for false negatives and true positives.

While the results presented above are interesting, it is useful to see how the models compare across EOIs and on various metrics. Towards this end, the graph in Figure 25 compares the number of true positives for the three models across all EOIs. On the X axis, five EOIs are shown. The Y axis displays the number of true positives. The red bars indicate the number of true positives for the original ICEWS model, the green shows the true positives for a random forest, and blue shows the results of the adaptive boosting algorithm. We can argue that we have a mixed results: adaptive boosting fares well in *Rebellion* and *Insurgency*, but the original ICEWS does better in *Domestic Crisis* and *Ethnic or Religious Violence*, while the random forest is the best model for predicting *International Crisis* events.

In contrast to the results for the true positives, the new models perform much better than the original ICEWS model when predicting true negatives as shown in 26. In all five EOIs, the random forest and adaptive boosting predict higher number of true negatives than the original ICEWS predicts. They do much better in predicting *Insurgency* and *Domestic Crisis*, but not much better in *Rebellion*, *Ethnic or Religious Violence*, and *International Crisis*.

In sum, random forests and adaptive boosting are good at capturing true negatives in the dataset, but have mixed results for predicting the actual events. Also, how well those models perform depends on what the EOI is. For example, the new models fared much better at predicting *Insurgency* but not in *Rebellion* or *Ethnic or Religious Violence*. This is puzzling because as shown below, random forests and adaptive boosting perform worse in predicting *Insurgency* in GDLET. Overall, the takeaway from these models is that while it is possible to do marginally better at predicting EOIs using the ICEWS data, just making use of new models does not produce a significant gain.

Figure 20: Classification Table for Rebellion

Table 1: Original ICEWS

	Predicted	
Actual	0	1
0	184	15
1	9	24

CER = 0.103448275862069

Table 2: Random Forest

	Predicted	
Actual	0	1
0	184	15
1	10	23

CER = 0.107758620689655

Table 3: ADA Boost

	Predicted	
Actual	0	1
0	181	18
1	6	27

CER = 0.103448275862069

Figure 21: Classification Table for Insurgency

Table 1: Original ICEWS

	Predicted	
Actual	0	1
0	194	24
1	8	6

CER = 0.137931034482759

Table 2: Random Forest

	Predicted	
Actual	0	1
0	218	0
1	9	5

CER = 0.0387931034482759

Table 3: ADA Boost

	Predicted	
Actual	0	1
0	216	2
1	8	6

CER = 0.0431034482758621

Figure 22: Classification Table for Domestic Crisis

Table 1: Original ICEWS

	Predicted	
Actual	0	1
0	211	10
1	9	2

CER = 0.0818965517241379

Table 2: Random Forest

	Predicted	
Actual	0	1
0	218	3
1	11	0

CER = 0.0603448275862069

Table 3: ADA Boost

	Predicted	
Actual	0	1
0	220	1
1	10	1

CER = 0.0474137931034483

Figure 23: Classification Table for Ethnic or Religious Conflict

Table 1: Original ICEWS

	Predicted	
Actual	0	1
0	207	8
1	8	9

CER = 0.0689655172413793

Table 2: Random Forest

	Predicted	
Actual	0	1
0	215	0
1	16	1

CER = 0.0689655172413793

Table 3: ADA Boost

	Predicted	
Actual	0	1
0	211	4
1	11	6

CER = 0.0646551724137931

Figure 24: Classification Table for International Crisis

Table 1: Original ICEWS

	Predicted	
Actual	0	1
0	163	19
1	24	26

CER = 0.185344827586207

Table 2: Random Forest

	Predicted	
Actual	0	1
0	172	10
1	21	29

CER = 0.133620689655172

Table 3: ADA Boost

	Predicted	
Actual	0	1
0	170	12
1	27	23

CER = 0.168103448275862

Figure 25: True Positive Comparison

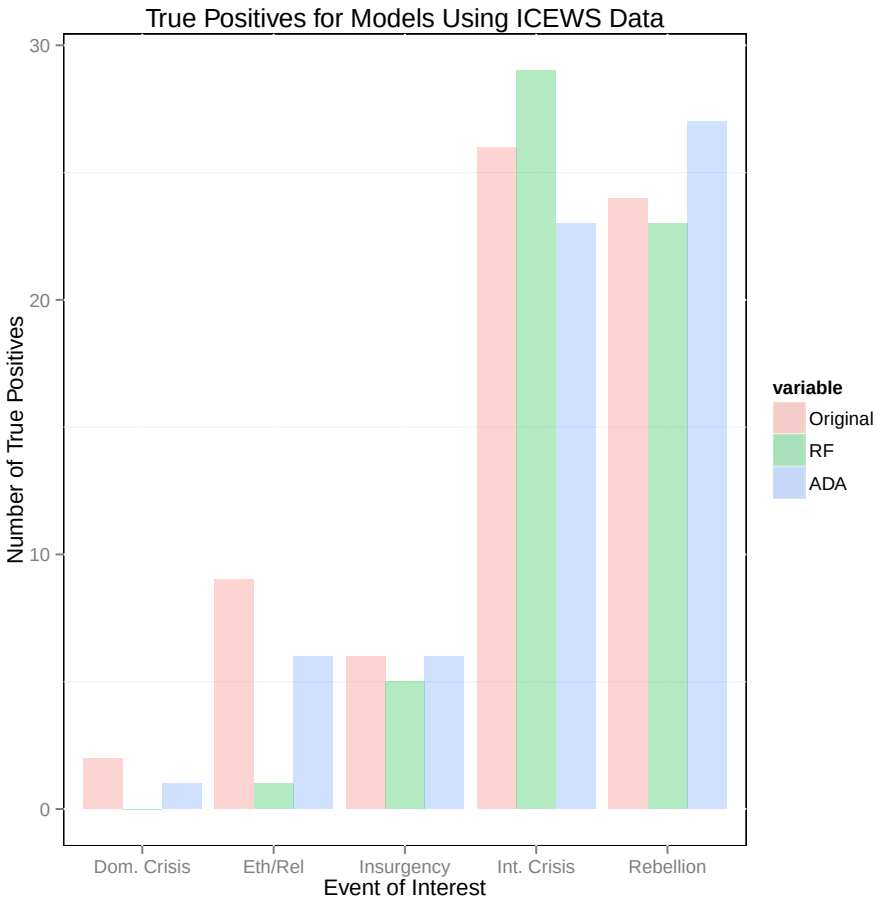


Figure 26: True Negative Comparison

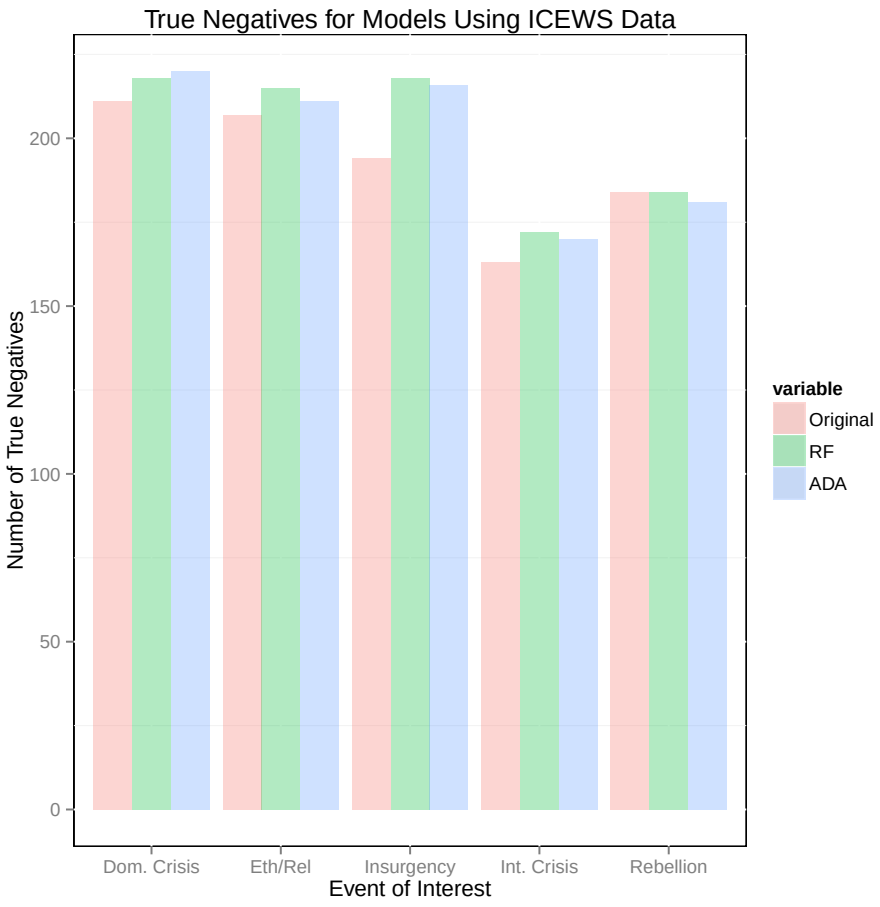


Table 9: Classification Table - Domestic Crisis

	Predicted	
Actual	0	1
0	619	2
1	51	24

5.2 Initial GDELT Models

The first-cut analysis run on the GDELT data makes use of only the event-data variables discussed in Section 2. For each EOI in the dataset, a random forest made up of 1000 trees was created. The random forests in this section were created using the `scikit-learn` package in Python (Pedregosa et al., 2011). The other possible settings for the random forest, such as depth of growth and minimum splits, were set to the defaults for the package. The results of the analysis are presented below. In general, the results obtained using the GDELT data are as good, or in many cases better, than those obtained using the ICEWS data. The model for each variable is assessed using three methods: classification tables that show the counts for each prediction type, e.g., true positive, false positive, etc., common metrics such as precision and recall, and receiver operator characteristic (ROC) plots, along with the area under the curve (AUC). The ROC plots plot the true positive rate and false positive rate at varying cutoffs for the probability of an observation being classified as a one on the dependent variable. The AUC, then, can be interpreted as how well a model is doing at predicting the EOI, while at the same time not over fitting the data. As a final note, the models presented below are all for the 6-month lagged variables.

For the first EOI under examination, *Domestic Crisis*, the results show that our model performs only moderately well. Figure 27 shows that we obtain an AUC of 93%, which is a high score, but a look at the classification table in Table 9 tells more of the story. We predict very few false positives, but fail to fully capture many of the positive observations of *Domestic Crisis*; we incorrectly predict no event when there is one in 51 cases.The is result is also illustrated by the recall score for the ones, .32, which indicates further that our model is not doing only an "okay" job at predicting the occurrence of domestic crises.

The next model predicts the occurrence of ethnic or religious violence within a given country, with results in Figure 28 and Table 10. The results for this model are the best of any

Table 10: Classification Table - Ethnic/Religious

	Predicted	
Actual	0	1
0	655	9
1	5	27

Table 11: Classification Table - Insurgency

	Predicted	
Actual	0	1
0	648	0
1	48	0

presented in this paper; we obtain an AUC of 99%, which is a remarkably high level of accuracy. A look at the classification table shows that this result is not a fluke of the data. We correctly identify 27 of the occurrences of ethnic or religious violence within a country, which, as shown by the recall score for the ones, is about 84% of the total. In addition, the model only falsely predicts ethnic or religious violence nine times. In short, our model does a remarkable job of predicting the occurrence of ethnic or religious violence, especially when considering the relative scarcity of such events in the dataset.

The third model has *Insurgency* as the response variable, with results presented in Figure 29 and Table 11. In contrast to the previous model presented for *Ethnic/Religious Violence*, this model performs the worst out of all models presented in this paper. We obtain an AUC of 81%, which at first glance seems rather good, but a further look at the data indicates that this is an artifact of the predictions. Our model predicts no positive observations of *Insurgency*; the model reduces to a naive model. This result is obviously problematic, but the following sections will address this issue in greater detail.

Table 12: Classification Table - International Crisis

	Predicted	
Actual	0	1
0	615	4
1	28	49

Table 13: Classification Table - Rebellion

	Predicted	
Actual	0	1
0	566	38
1	12	80

The penultimate model examines the *International Crisis* variable. This model performs fairly well, with an AUC in Figure 30 of 95% and a recall for the ones of .64. Table 12 indicates that the model correctly classifies 49 out of 77 events, while only falsely predicting an international crisis four times. The model does, however, incorrectlypredict 28 observations of international crisis as negative observations.

The final EOI, *Rebellion*, has the second-best model presented in this section. We correctly classify 80 of the 92 rebellion events, and the model obtains an AUC of 96%. What is interesting about this model, however, is the high number of false positives, i.e., incorrect predictions of a rebellion.This finding is interesting, since these false positives may serve as a "watch list" for states that have the potential of experiencing a rebellion, but for one reason or another fail to actually experience the event.

The main takeaway from these models is that the GDELT data, on average, does a better job of predicting than the ICEWS data. The models are *very* good at capturing the zeros, or non events, in the dataset, but have varying success in predicting the actual events of interest. This may be the case due to the extremely rare nature of some events within the dataset, such as *Insurgency*. It is possible that in these cases random forests are not the most appropriate model to use, since the bootstrapped subset of data used to construct the decision trees may cause a low number of positive observations to occur in the subset. In addition, the variables derived from event data may not be the best at predicting someEOIs; adding in structural variables, such as GDP, may raise the predictive accuracy of the models.The following section examines these questions in greater detail.

Figure 27: Domestic Crisis Results

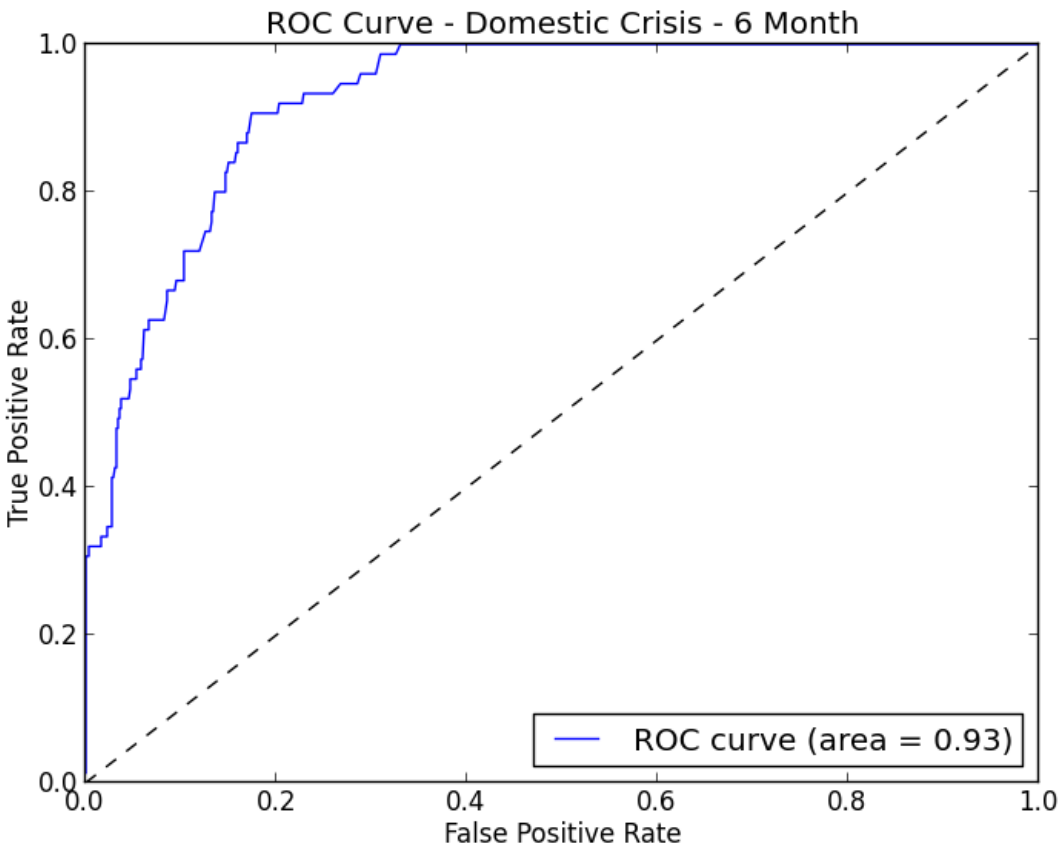


Figure 28: Ethnic/Religious Results

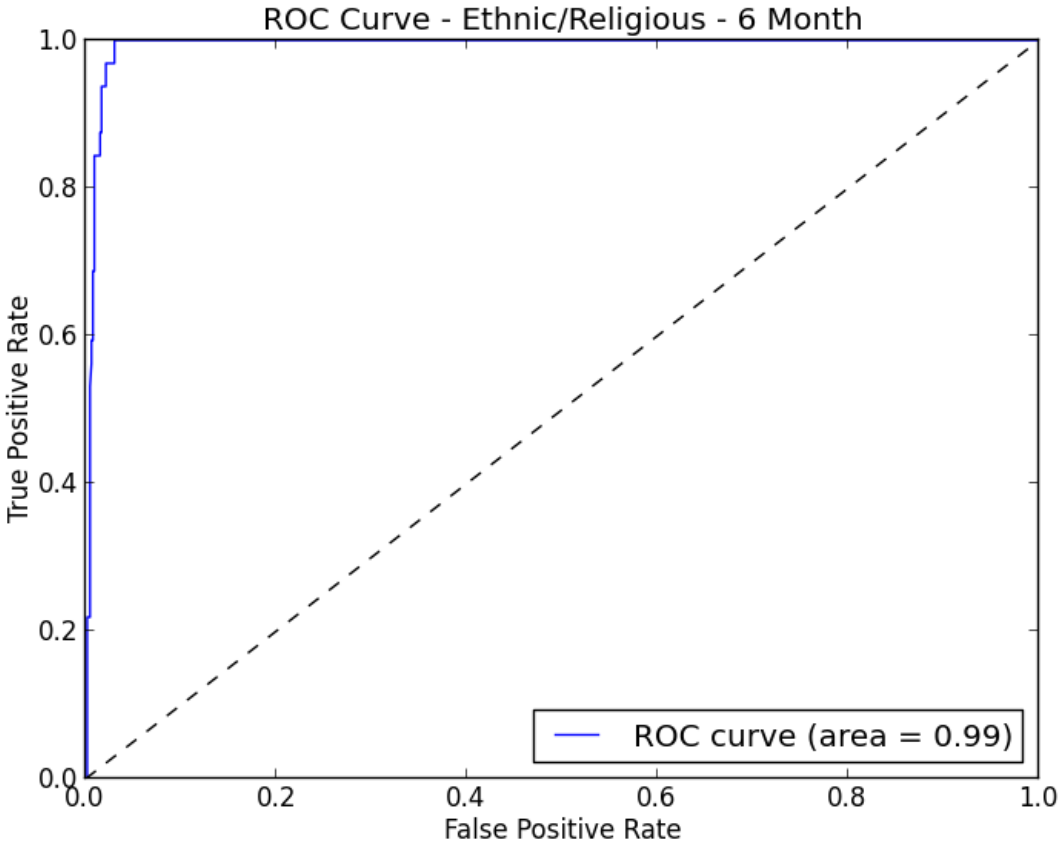


Figure 29: Insurgency Results

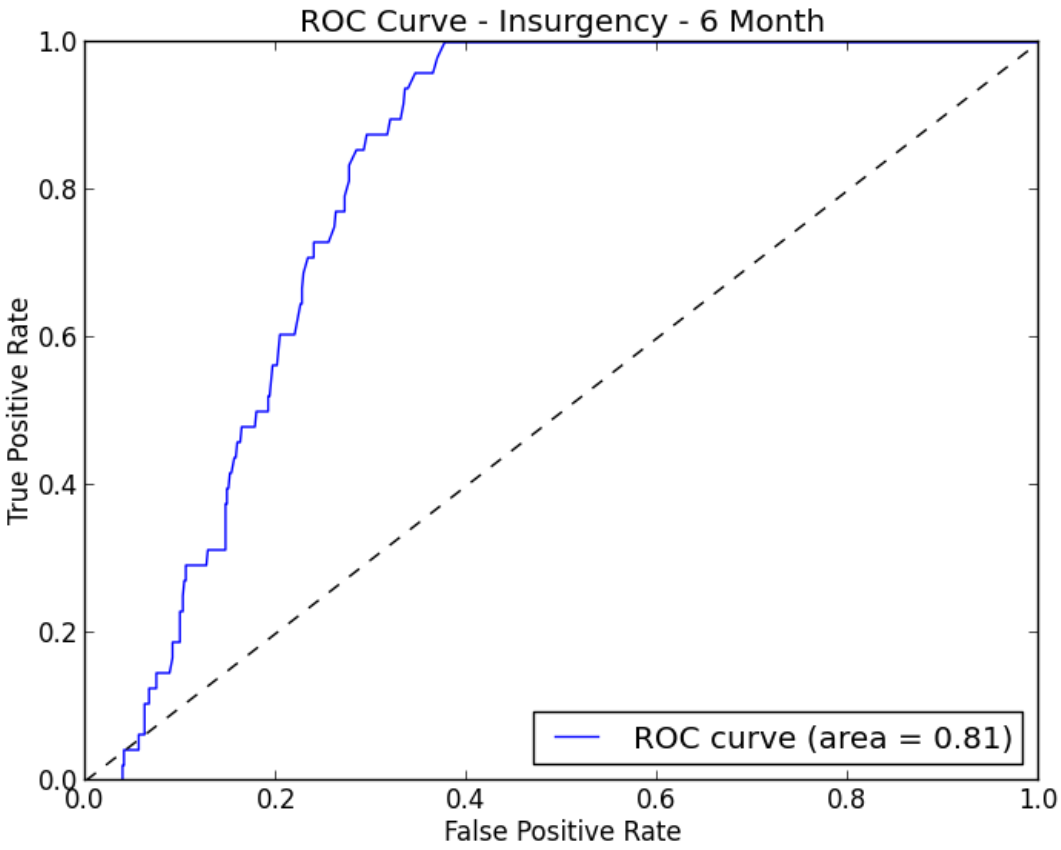


Figure 30: International Crisis Results

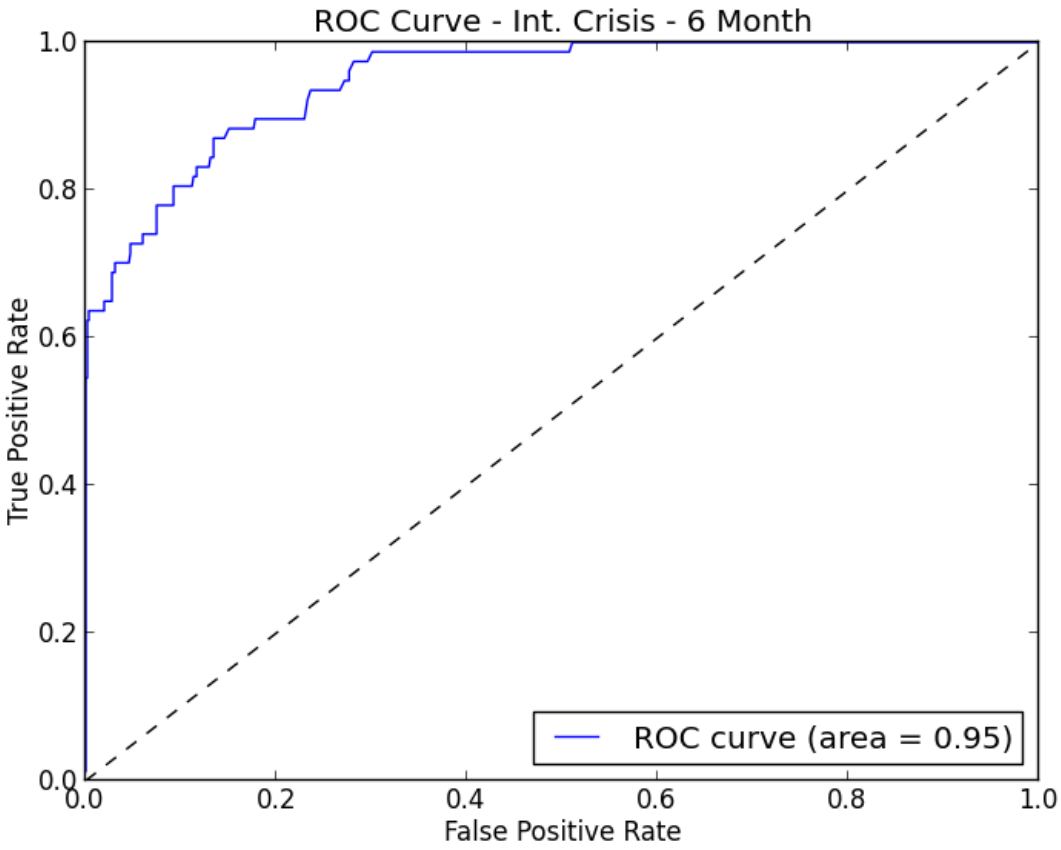
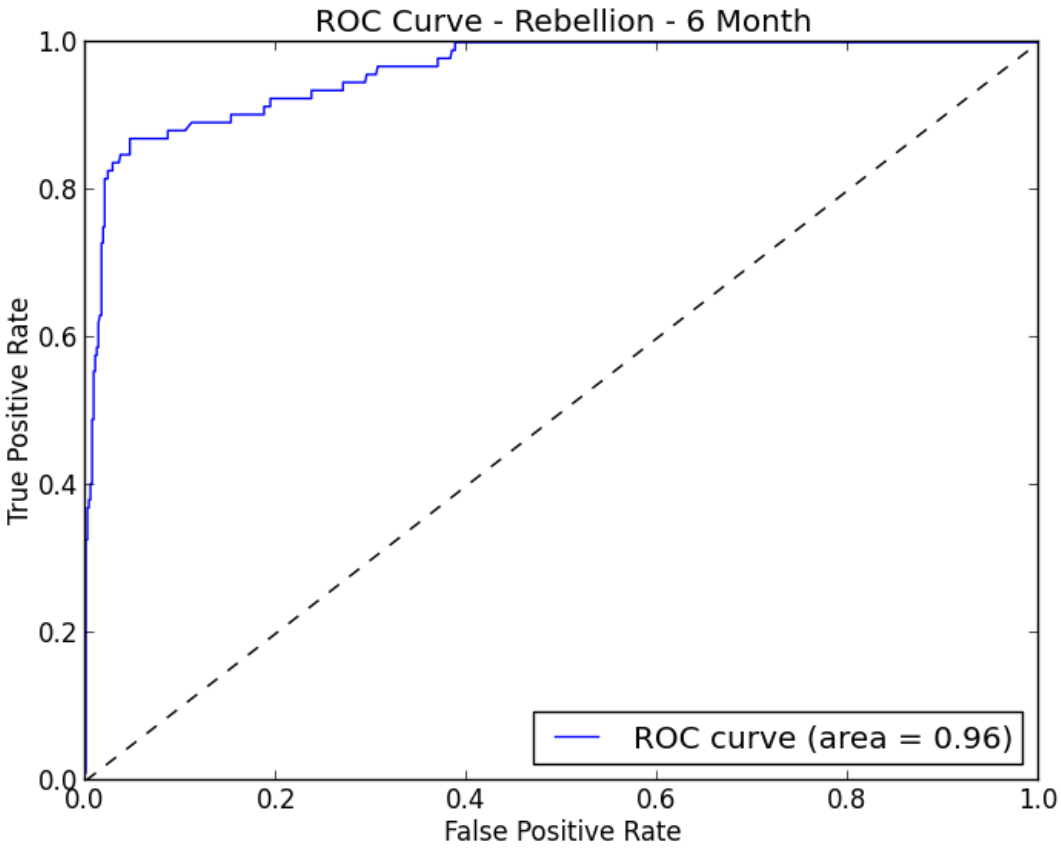


Figure 31: Rebellion Results



5.3 Further Testing with GDELT

In order to determine if different models are able to better predict the EOIs, this section applies adaptive boosting algorithms to the GDELT data, along with adding in external data to determine if new data increasesthe predictive accuracy. The models are first assessed using the receiver-operator characteristic (ROC) plots, with false positive rates on the X axis and true positive rates on the Y axis. The ROC plots have the area under the curve (AUC), showing how well the models are predicting the EOIs. Larger area under the curve indicates greater predictability. Second, we compare true positive and true negative rates between random forest and Adaptive Boosting for each event of interest. We find that adaptive boosting does a better job at identifying true positives but random forest does a better job at identifying true negatives. Last, we examine the models with addition of structural, stationary variables. We find that adding those variables increases the model predictability compared to when we used only event variables. All models presented in this section are based on the 3-month lagged variables.

The first EOI under examination, shown in Figure 32, is *Rebellion*. The ROC plot for this EOI shows an AUC of 97%, which is a very good result. The next figure shows the results for *Insurgency*. This is our worst adaptive boosting model. The AUC shows 87%. It is interesting that random forest has also done worst in predicting *Insurgency*, as shown in the previous section, with an AUC of 81%. The third EOI we examine is *Domestic Crisis*, shown in Figure 34. Here the model performs quite well, with an AUC of 94%, which is a high score. The next variable, *International Crisis*, is the second best adaptive boosting model along with *Rebellion* with a AUC of 97%.

The final model, shown in Figure 36, is the best model we obtained from adaptive boosting with an AUC of 99%. Interestingly, *Ethnic or Religious Conflict* is also the EOI that the random forest has highest predictive power on.

When we look at the AUCs of the random forest and adaptive boosting models, we find a pattern in the model predictability depending on the EOIs. In all cases, the AUC is highest for *Ethnic or Religious Conflict*, followed by *Rebellion* and *International Crisis*. The models for *Domestic Crisis* do a moderate job, with *Insurgency* being the most difficult EOI to predict for both random forest and adaptive boosting. We will compare the number of true positives and true negatives between random forest and adaptive boosting below. This allows us to compare the predictability of both models. This also allows us to see whether the pattern exists for different EOIs.

Figure 37 shows a comparison between the random forest and adaptive boosting in terms of

Figure 32: Rebellion

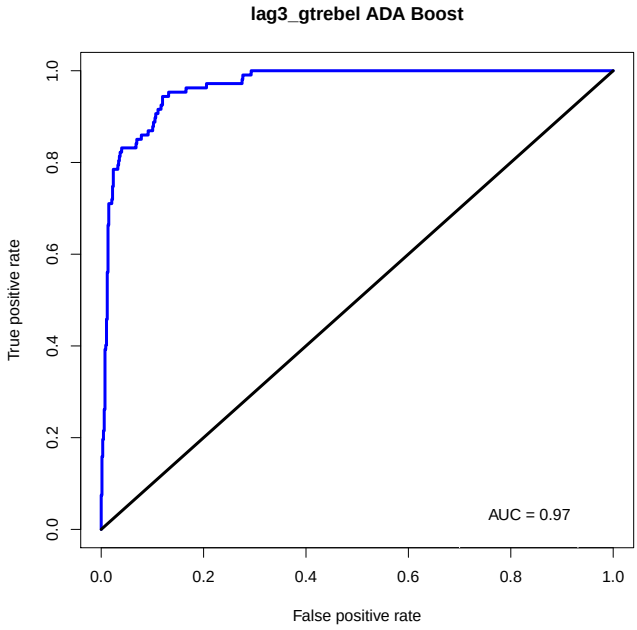


Figure 33: Insurgency

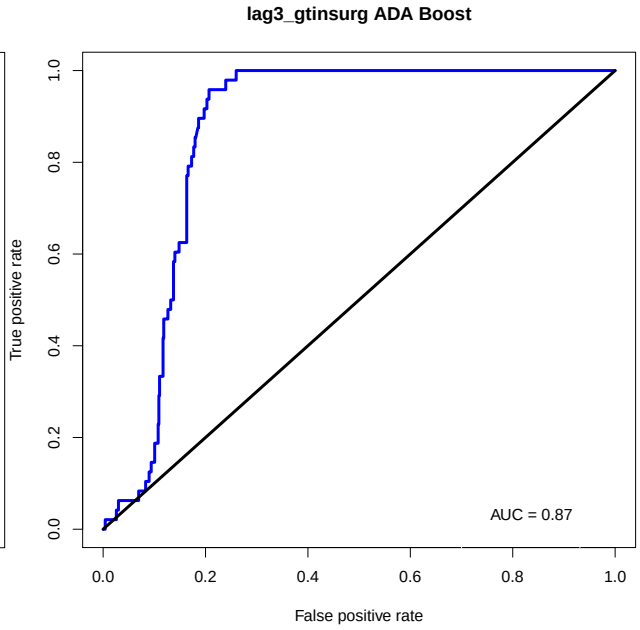


Figure 34: Domestic Crisis

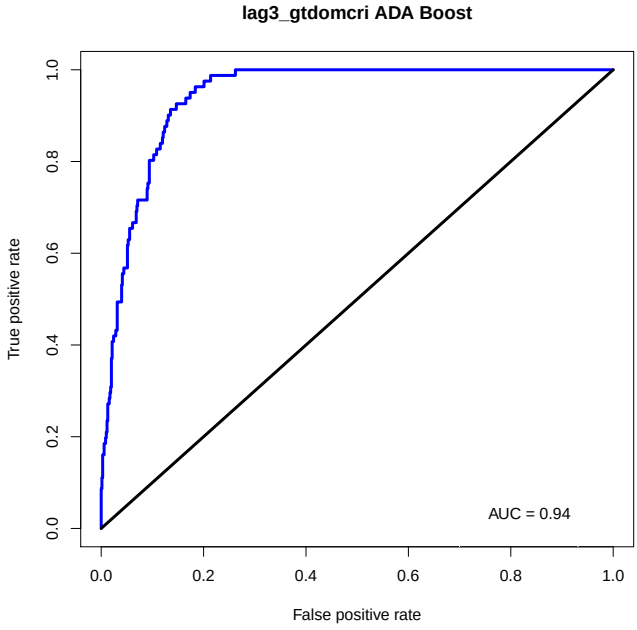
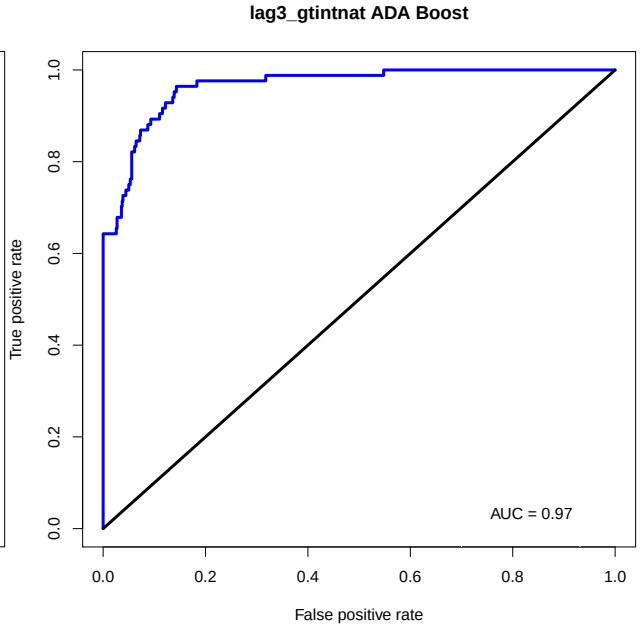


Figure 35: International Crisis



number of true positives (the number of events they have correctly predicted) for each EOI. The X axis shows each EOI and the Y axis shows the number of true positives. The chart shows that the adaptive boosting has done a better job in predicting true positives for all of five EOIs. This higher predictive accuracy on the true positives comes with a cost, however, as demonstrated by Figure 38.

As the figure shows, the random forest has done a better job in predicting true negatives (the events that the model has not predicted and have not occurred). We can also observe a consistent pattern in the different models. For both random forest and adaptive boosting, the numbers of true positives were greatest for *Rebellion* and lowest for *Insurgency*. The numbers start to rise moving along *Domestic Crisis*, *Ethnic or Religious Conflict*, and *International Crisis*. In terms of true negatives number, both models show M-shaped curves: lowest for *Rebellion* and *International Crisis*, followed by *Domestic Crisis*. The numbers are greatest for *Insurgency* and *Ethnic or Religious Conflict*.

Since our models do rather poorly on some variables, the question becomes are there any ways in which to improve the model predictability? One possible answer is the addition of structural variables to the model. The models we ran so far have only incorporated event variables. When including the more "stationary" variables, we find that the predictive accuracy of the model has increased by a great margin. The added structural variables include:

- Contiguity
- Oil exporter
- Ethnic fractionalization
- Primary commodity exports
- Trade GDP
- Military expenditures
- Unemployment
- GDP growth
- GDP per capita
- Population

Figure 39 shows on the left the ROC plot for the *Insurgency* EOI using adaptive boosting. The AUC has increased greatly compared to the AUC of the model without structural

variables, increasing from 73% to 97%. We find similar results when applying a random forest to the other EOIs; we find that the predictive accuracy improves in all instances when using additional data.

In the graph in Figure 40, a list of the EOIs are displayed on the X axis and the AUC values are on the Y-axis. The variables are lagged by 3 months. The results for each model are shown both with and without the use of extra data. In general, the models that make use of the external data do a better job at predicting than those that use only the event-data derived variables. The margin of change is greatest for *Insurgency* where the predictive accuracy of the models without extra variables are lowest. We can argue that adding structural variables is crucial for improving model predictability when using the GDELT data.

Figure 36: Ethnic/Religious Violence

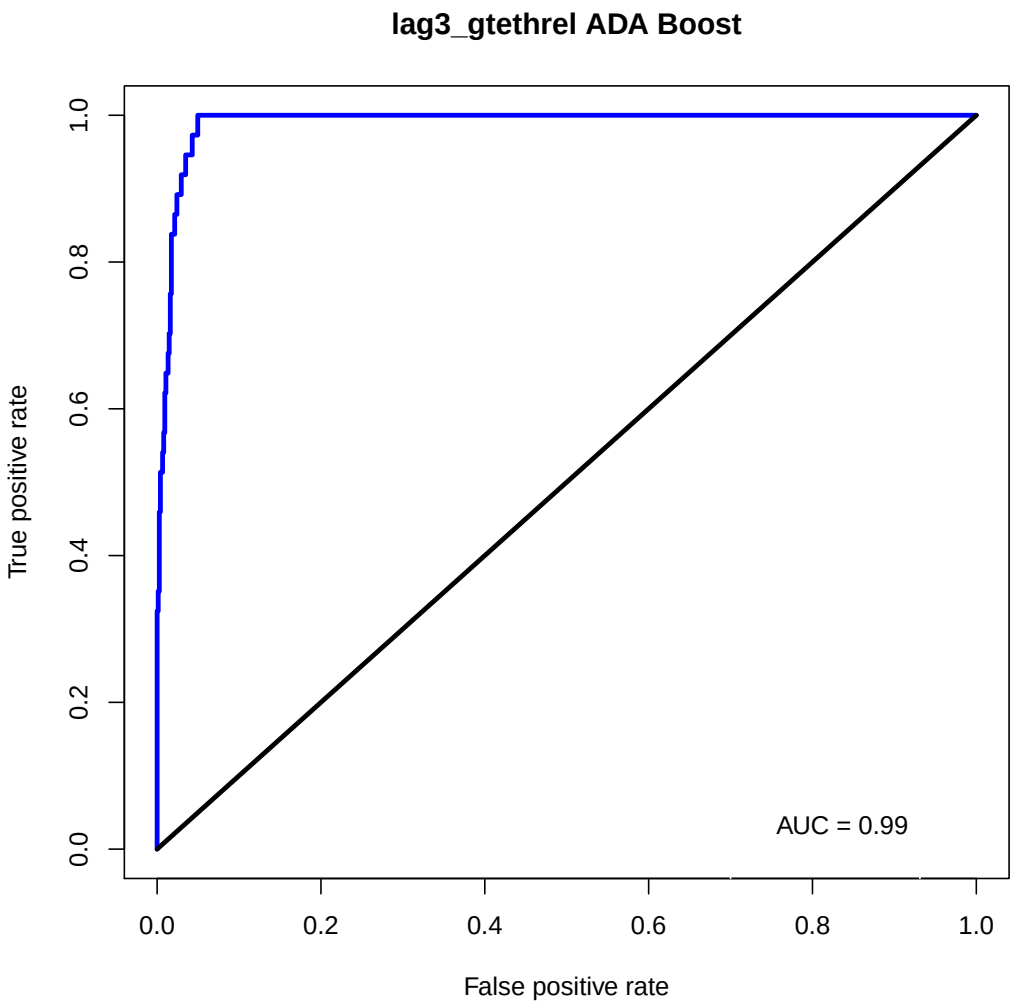


Figure 37: True Positives

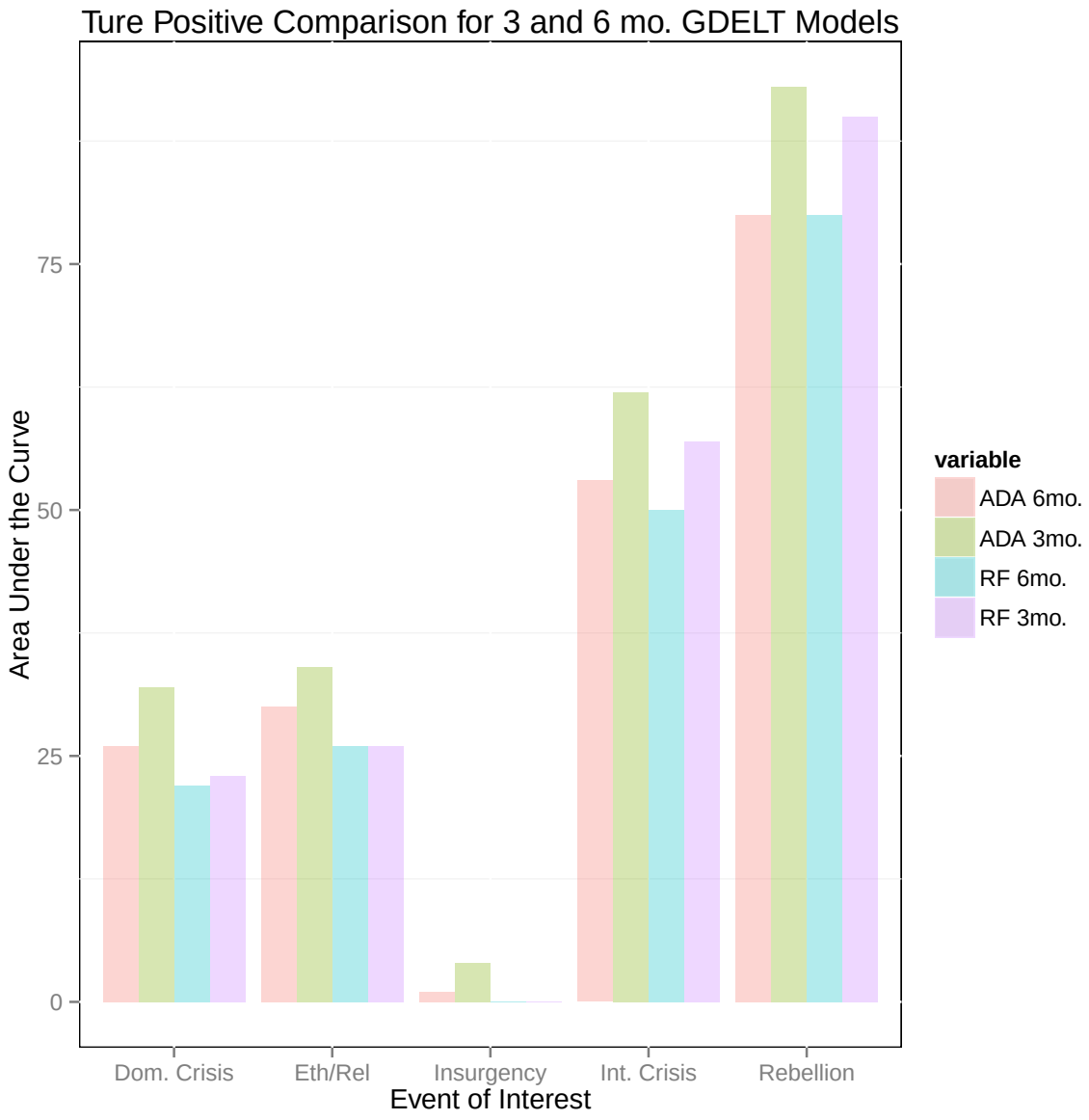


Figure 38: True Negatives

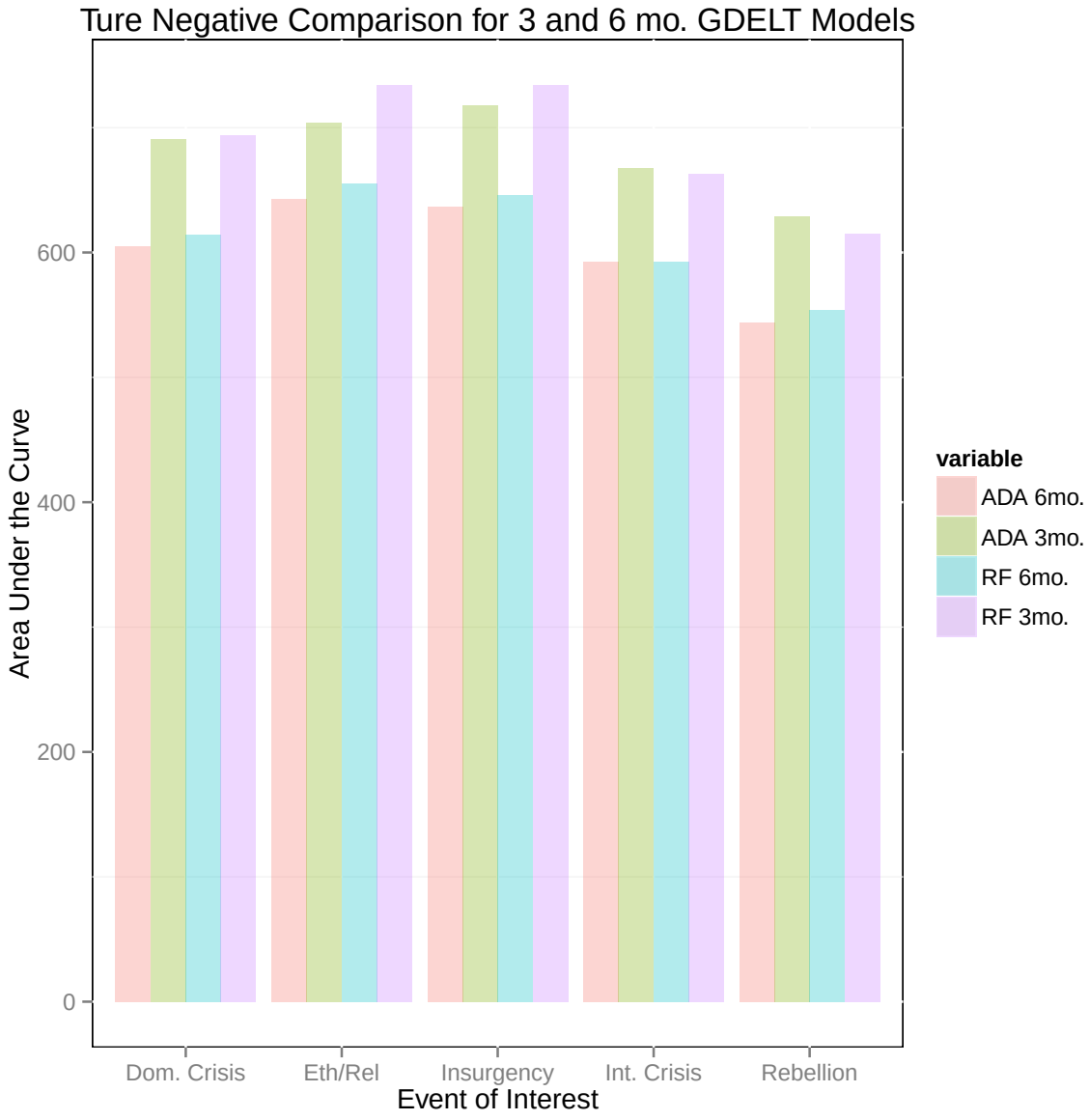


Figure 39: Insurgency EOI with Additional Variables

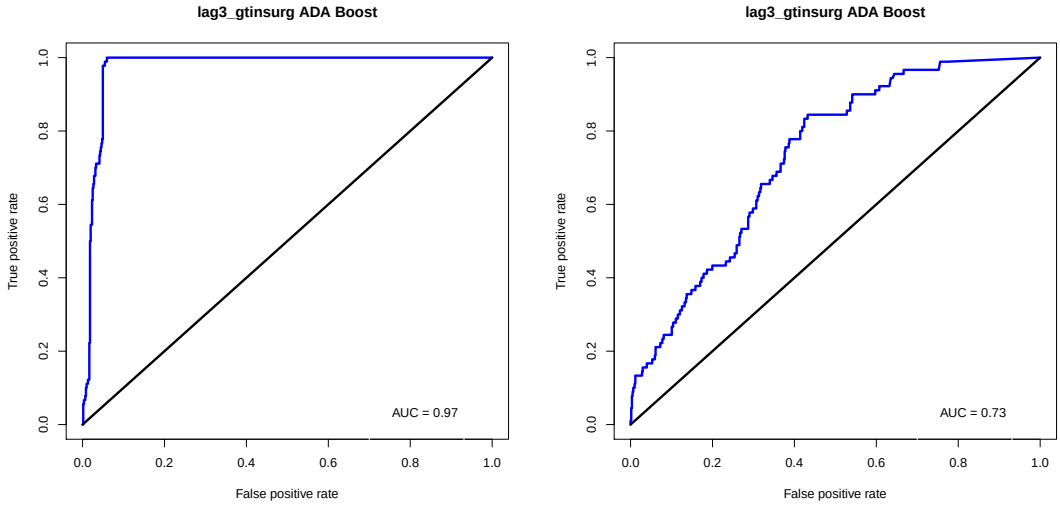
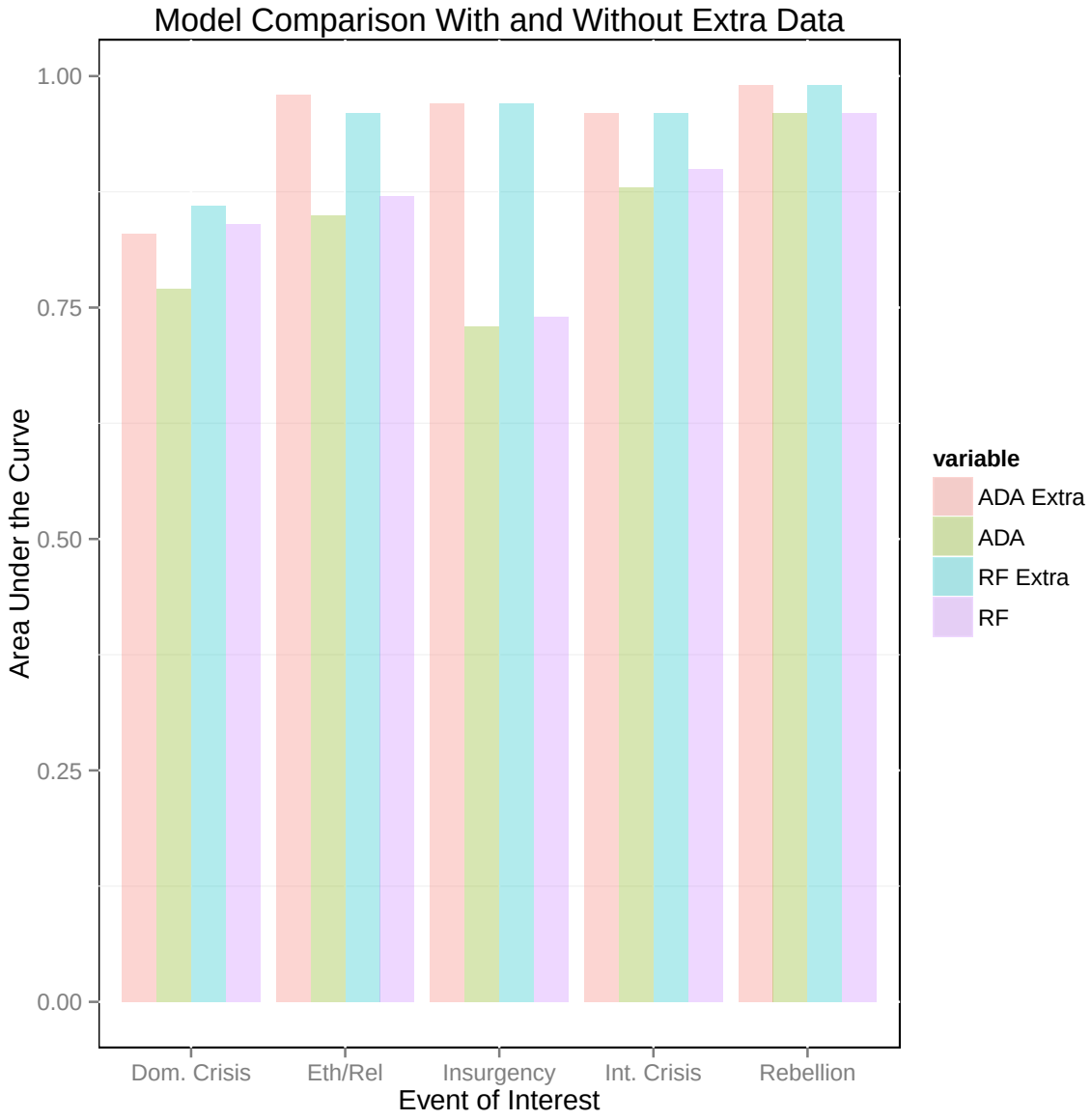


Figure 40: Impact of External Data



6 Conclusion

Our findings suggest that in general GDELT does a better job than the available ICEWS of predicting the five events of interest that were defined by the ICEWS project. From our geographical analysis, the main results are that a recurring cluster from India to Southeast Asia is present, though not consistently every year. In addition, hot spots of conflict are more likely to be present than cold spots, thus indicating that countries that do not have conflict are isolated and have a lower chance of conflict spillover.

From our BMA analysis of variable importance for rebellion and international crisis, we found that the inclusion of gov_opp conflict variables and gov_opp cooperative events is important in explaining rebellion. In addition, nonevent variables from the original ICEWS data are also important, such as ethnic fractionalization (more important than the GDP variables) and contiguity. For international crisis, the variables for verbal and material conflict are important, though this effect may be magnified by the autoregressive nature of conflict. Specifically, gov_gov_vercf and gov_gov_vercp are of particular interest as states who are arguing with each other are likely to either not have the inclination or the capacity to initiate a conflict.

From our application of new modeling techniques on ICEWS, we found that the random forest did worse in predicting Rebellion, while adaptive boost performed about the same as the original model. For Insurgency, the random forest and adaptive boost increased the predictability (the classification error rate decreased by more than 50%). In the case of Domestic Crisis, the random forest and adaptive boost performed better, but did not produce as dramatic of decrease of the classification error rate for insurgency. As for Ethnic/Religious Conflict, the random forest and adaptive boosting do not perform better predicting ethnic or religious conflict than the original ICEWS model. Lastly, both random forest and adaptive boosting performed better for International Crisis, though random forest does better than adaptive boost.

When we applied the random forest to GDELT, we found that we perform the best in predicting Ethnic/Religious Conflict, followed by Rebellion, International Crisis, and Domestic Crisis. We performed the worst in predicting Insurgency (in stark contrast to ICEWS where it was where we performed the best). Overall, GDELT does a better job than ICEWS at prediction, specifically at predicting nonevents. However, this may be due to the rarity of specific events such as Insurgency.

We found a similar pattern when we applied adaptive boosting to GDELT as the AUC is

the highest for Ethnic/Religious Conflict, followed by Rebellion, International Crisis, and Domestic Crisis. Once again, Insurgency is the most difficult to predict for adaptive boosting. Adaptive boosting did better in predicting true positives than random forest, however at the cost of performing worse at predicting true negatives. One way we improved our predictions was including structural variables. By doing so, we were able to generally increase our AUC values especially for Insurgency, from 73% without the structural variables to 97% with them included.

As we noted in the introduction, these results cannot be considered definitive until the long-promised official unclassified version of the ICEWS data is released, at which point we intend to re-do this analysis, though it seems unlikely that the results will be substantially. An even better situation would be for ICEWS to provide regular updates of the data they produce from the unclassified sources—GDELT, after all, is updated every 24 hours—so that an on-going comparison could be made of the two data streams, as the news environment continues to change. This initial results, however, certainly appear to indicate that at the very least the two sets are comparable for both descriptive and predictive purposes.

To the extent that one of the two sets is superior, however, it is clearly GDELT, at least based on the ICEWS measures we have available. This seems counter-intuitive, given the far greater efforts—millions of dollars invested in large teams, versus the one-person dissertation project which produced GDELT—invested in ICEWS. Ironically, however, the massive efforts invested in insuring that ICEWS was "clean" in the sense of having a very low false positive rate may have in fact significantly degraded the quality of the data.

The issue here is the general one that the Nobel Prize-winning economist Daniel Kahneman (2011) calls "what you see is all there is," and is one of the major pathologies identified in human decision-making: human experts are very bad at seeking out additional information. This problem may have affected the ICEWS data in two significant ways. First, the obsession with eliminating false positives was not counter-balanced by a comparable effort at eliminating false negatives, which are much harder to deal with since this requires a human analyst to know that something important occurred which is not present in the data, an extraordinarily difficult task to do consistently across an AOR covering more than half the world's population and a decade and a half. False positives, in contrast, are fairly simple: there is a report of a military clash between the US and Japan, the analyst knows these countries have been at peace the entire period, and further examination shows this to be a mis-coding of a commemoration of a WWII battle.

Second, the steps taken to eliminate false positives almost certainly eliminated a number

of true positives as well (again, this is just another facet of the failure to control for false negatives): this is certainly consistent with the very different densities of the two data sets which we saw in Section 3. This would be further complicated if the attention to false positives was not applied uniformly across all countries, for example if most of the attention had focused on a small number of countries with active conflicts at the time of the development of the data sets.

GDELT, in contrast, applies more of a "firehose" approach, which almost certainly has high false positive rates but may also have dramatically lower false negative rates. If one is using the data for *monitoring* purposes, this is usually a bad thing (or at least it seems like a bad thing since humans are generally insensitive to false negatives); for *statistical forecasting* purposes, however, GDELTs lower false positive rate may provide it with a significant edge, consistent with these initial results.

This has two implications for the further development of models based on event data. First, it reinforces the fact that no event data set, however expensive, or whether coded using human or automated methods, provides a "god's eye view": every data set has specific statistical characteristics and these need to be taken into consideration when used for modeling. Second, given that low-false-positive-rate datasets such as ICEWS may still have utility for monitoring, a next obvious step in the development of forecasting models would be ensemble approaches using both data sets. We look forward to the release of a canonical version of the ICEWS data so that such approaches could be developed further.

References

Anselin, L. 1995. "Local indicators of spatial association—LISA." *Geographical Analysis* 27(2):93–115.

Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41:641–674.

Fearon, James D. and David D. Laitin. 97. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* pp. 75–90.

Feldkircher, Martin and Stefan Zeugner. 2009. "Benchmark priors revisited: On adaptive shrinkage and the supermodel effect in bayesian model averaging." IMF Working Papers.

Getis, Arthur and J. Keith Ord. 1996. Local spatial statistics: an overview. In *Spatial analysis: modelling in a GIS environment*, ed. Paul A. Longley and Michael Batty. New York: Wiley pp. 261–275.

Goldstein, Joshua S. 1992. "A Conflict-Cooperation Scale for WEIS Events Data." *Journal of Conflict Resolution* 36:369–385.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2nd ed. Springer.

Kahneman, Daniel. 2011. *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.

Leetaru, Kalev and Philip A. Schrodt. 2013. "GDELT: Global Data on Events, Location and Tone, 1979-2012." Presented at the annual meeting of the International Studies Association.

Montgomery, Jacob M. and Brendan Nyhan. 2010. "Bayesian Model Averaging: Theoretical Developments and Practical Applications." *Political Analysis* 18(2):245–270.
URL: *http://pan.oxfordjournals.org/content/18/2/245.abstract*

Montgomery, Jacob M., Florian M. Hollenbach and Michael D. Ward. 2012. "Improving Predictions Using Ensemble Bayesian Model Averaging." *Political Analysis* .
URL: *http://pan.oxfordjournals.org/content/early/2012/03/22/pan.mps002.abstract*

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12:2825–2830.

Schrodt, Philip A. 2011. *TABARI: Textual Analysis By Augmented Replacement Instructions*. `http://eventdata.psu.edu/tabari.html`.

Schrodt, Philip A. 2012. *Conflict and Mediation Event Observations (CAMEO) Codebook*. `http://eventdata.psu.edu/data.dir/cameo.html`.

Schrodt, Philip A. and David Van Brackle. 2013. Automated Coding of Political Event Data. In *Handbook of Computational Approaches to Counterterrorism.*, ed. V.S. Subrahmanian. Springer pp. 23–50.

Schrodt, Philip A., Deborah J. Gerner and Ömür Yilmaz. 2009. Conflict and Mediation Event Observations (CAMEO): An Event Data Framework for a Post Cold War World. In *International Conflict Mediation: New Approaches and Findings*, ed. Jacob Bercovitch and Scott Gartner. New York: Routledge.

Zeugner, Stefan. N.d. *Bayesian model averaging with bms for bms version*. 0.3.0 ed. CRAN.