GDELT and ICEWS Comparison

Michael D. Ward & Andreas Beger & Josh Cutler & Matthew Dickenson & Cassy Dorff

September 23, 2013

This research was undertaken at `mdwardlab.com` at Duke University.

A very brief history of event data

Studying Conflict Fractally has been a long-term goal, going back decades. Many scholars are using large databases to examine localized conflicts. In August, a graphic of reports of protest activity around the world, based on the GDELT database, created by John Beiler receives over 150,000 views on the web. This is because large amounts of data that are structured to reveal "who does what to whom" are now available. These data are called event data, and have been around since the mid-1960s when they were invented by Charles McClelland who aimed at creating a way to study diplomatic history in a systematic way.[1]

From WEIS, through COPDAB, CREON, and many others, event data collections have long served the policy and academic community as a working sensor revealing details about political interactions among countries.[2]In reality, Philip A. Schrodt has led the way, initially developing the KEDS project. This was the first project to rely on an automated content analysis of textual information to create event data, as prior efforts had largely relied on human coding of compiled chronologies. KEDS led to TABARI and an ontology for coding called CAMEO that have been used widely.

CAMEO serves as the coding basis for GDELT, a "global database of events, language, and tone," that has been introduced in the past year and has generated a large amount of excitement in the policy and academic community. GDELT is well described elsewhere, and has the great benefit of being open source, and continuously updated, permitting its widespread use in academic as well as policy studies. The active site contains links to many articles covering GDELT, the complete documentation, computer programs that have been used to analyze the data, and the actual data: `http://gdelt.utdallas.edu/`. According to recent reports, GDELT now includes more than 250 million events, ranging from 1979 to the present, and has many planned enhancements that Schrodt and Letaru have announced.[3]

ICEWS is an early warning system designed to help US policy analysts predict a variety of international crises to which the US might have to respond. These include international and domestic

[1] McClelland's efforts eventually were institutionalized into the World Event Interaction Survey (aka WEIS). The start of this thread of research can be found in McClelland's early article, "The Acute International Crisis," *World Politics*, Volume 14, Special Issue 01, October 1961, pages 182-204. Recently, there is a wonderful summary from two younger scholars on the idea of using text as data: Justin Grimmer and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* 21.3 (2013):267-297. Stewart worked on an early version of the ICEWS project at the University of Georgia. See also Robert C. North, Ole R. Holsti, George Zaninovich, and Dina A. Zinnes. *Content analysis: A handbook with applications for the study of international crisis*. Vol. 184. Evanston, IL: Northwestern University Press, 1963. A great summary of the early work is found in Deborah J. Gerner, et al. "The analysis of political events using machine coded data." *International Studies Quarterly* 38.1 (1994): 91-119.

[3] Phil Schrodt, "GDELT: Global Data on Events, Location, and Tone," a presentation for the Conflict Research Society, Essex University, 17 September 2013

crises, ethnic and religious violence, as well as rebellion and insurgency. This project was created at the Defense Advanced Projects Research Agency under the guidance of Sean O'Brien, but has since been funded (through 2013) by the Office of Naval Research.[4] While it started with a test bed of twenty-five countries in the US Pacific Command, currently, ICEWS has coverage of 167 countries, excluding the U.S., and collates textual data on them.

The ICEWS system began as a 4-year DARPA program in 2007, and was established to demonstrate the potential of using social science models and theory to forecast and understand nation-state instability across a range of countries. The program proved highly successful and spawned 3 component tools iTRACE (news analytics), iCAST (instability forecasting), and iSENT (sentiment analysis and opinion propagation in social media).[5]

Four aspects of the project are noteworthy: (1) it produces and consumes a very rich corpus of text which is analyzed with powerful techniques of automated event-data production.[6] Indeed, Schrodt was involved in the first phases of the project; (2) it uses a variety of systematic (mostly statistical) models to generate predictions for the basic five variables mentioned above–these predictions are for six months in advance and are graded for accuracy[7]; the various predictions are averaged using ensemble methods to create an ensemble prediction that is more accurate, with fewer false positives and false negatives, than any of the individual models[8]; and importantly (4) a version of this decision aid has been deployed for several years, and has a large number of government users. We have been a participant in this research since the beginning, and have several recent papers related to our efforts at the models and the statistics behind them.[9]

But, given the nascent existence of the GDELT and ICEWS data sets, people want to know 1) how are they different and 2) which is better?

A focused comparison of GDELT and ICEWS data

How to compare different databases?

An important dimension is availability. GDELT has since the summer of 2013 been open and freely available. That is a big win for the policy and academic community. Anyone, including research from Walmart, JPMorgan Chase, Goldman Sachs, Barclays, Expedia, the Central Intelligence Agency, the Human Rights Data Analysis Group, and even mdwardlab.com can freely use the data. This is tremendous and merits acknowledgment and thanks. ICEWS data are not. The full story of why can't be told here, but suffice it to say that ICEWS's

[4] Sean P. O'Brien, "Crisis early warning and decision support: Contemporary approaches and thoughts on future research." *International Studies Review* 12.1 (2010): 87-104. But especially see: Sean P. O'Brien, "A multi-method approach for near real time conflict and crisis early warning." *Handbook of Computational Approaches to Counterterrorism*. Springer New York, 2013. 401-418.

[5] In July 2012, a Technology Transition Agreement was established between OASD (R&E), ISPAN, and ONR for the transition of an extended (worldwide coverage) and hardened version of all three tools to the ISPAN program beginning in October 2011. In January of 2012, ONR established a contract with Lockheed Martin Advanced Technology Laboratories (ATL) for the worldwide extension, hardening, and transition of iTRACE and iCAST. In April 2012, SPAWAR established a similar contract with Lockheed Martin IS&GS for the refinement, hardening, and transition of the iSENT component. These efforts were sponsored by OSD (ASD R&E Ð HSCB Program) and have a period of performance of 3 years.

[6] Boschee, Elizabeth, Premkumar Natarajan, and Ralph Weischedel. "Automatic extraction of events from open source text for predictive forecasting." *Handbook of Computational Approaches to Counterterrorism*. Springer New York, 2013. 51-67

[7] Michael D. Ward, Nils W. Metternich, Christopher Carrington, Cassy Dorff, Max Gallop, Florian M. Hollenbach, Anna Schultz, & Simon Weschle. "Geographical Models of Crises: Evidence from ICEWS," *Advances in Design for Cross-Cultural Activities, Part I*, CRC Press, edited by Dylan D. Schmorrow and Denise M. Nicholson, 2012, pp. 429-438

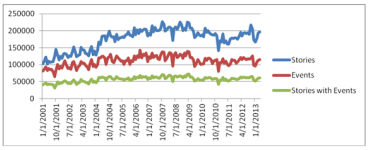
[8] Jacob M. Montgomery, Florian Hollenbach, Michael D. Ward. "Improving Predictions Using Ensemble Bayesian Model Averaging," *Political Analysis* 20.3 (2012): 271-291

[9] See Michael D. Ward, Nils W. Metternich, Cassy Dorff, Max Gallop, Florian M. Hollenbach, and Simon Weschle. "Learning from the past and stepping into the future: toward a new generation of conflict prediction." *International Studies Review* 15.4 (2013): in press.

success within the operational community of the US government led to a reversal of policy and the contravention of extensive plans, operational as recently as 2010, to make all the ICEWS data freely available to all users.

A second approach is to look at the goals of each database. W-ICEWS event data collection (hereafter W-ICEWS) has a traditional approach, but modern mechanisms. The collection tries to accurately reflect the activities among and within nations and their actors. Thus, a fair amount of effort goes into filtering the raw stream of reported stories into a unique stream of filtered stories. That filtering attempts to eliminate multiple reports of the same event, even if they appear on adjacent days. Stories about the history of violence between, for example, Japan and Korea, during the 1930s are eliminated from the stream of events that apply to the current era, even if they appear in the contemporary press. So too are stories about the war being waged by the Bank of Japan on the Indian currency filtered out– as are the many business and sports stories that use the language of politics to describe contests that fall largely outside the realm of politics. Indeed, an early experiment of the W-ICEWS data team involved a comparison of coding of stories by intelligence analysts and the automated coding ontologies of W-ICEWS.

At present, the ICEWS event data go back to 2001 and contain about 30 million "stories" that are parsed and coded using NLP techniques based on word graphs using a specially developed ontology based on CAMEO. These are gleaned from about 6000 sources, but many of these are aggregators of hundreds of other sources. So the number of sources is not really informative. What is useful to know is that these sources span international, regional, national, and local sources. Importantly, these are all filtered and subjected to the developed ontology using the NLP techniques developed by BBN. The rough equivalence to the rate of collected stories, events, and stories with events is quite remarkable.



Unfortunately, there is no ground truth to use to gauge the accuracy of these data. Each data point needs to be assessed by drilling down to the story, reading it, and figuring out if the coding is correct. To do so, obviates the goal of automated event coding, but can be useful in identifying errors in that coding. While individual mileage may vary, our experience has been reasonably reassuring to us that generally ICEWS is getting at something real. Of course, we don't know what stories were not written, and like the well-known bias in SIGACTS, these events only happen when they get reported.

The GDELT data collection starts from an entirely different philosophy. Rather than trying to get to the "truth" it tries to capture an extensive picture of what is reported, both in its details (who, what,

were, when) and its extensiveness (how many reports are there). And so what we might expect is that GDELT will have many more events per country per unit time. It certainly does in the aggregate. GDELT has about 68,000 country months (34 years by 167 countries) compared to about 24,000 country months in ICEWS. Yet, GDELT has an order of magnitude more stories on which its events are based, so we should expect more events, in total. Though this doesn't mean more data in every specific situation. Moreover, the volume of data being harvested by GDELT is growing exponentially, as are the base level of events therein. As a result, GDELT has–at present–by design a collection mechanism that tries to actually maximize reports, but no mechanism for pruning those events to eliminate the false positive reports. In part because of the winnowing process in ICEWS data collection, there is no corresponding exponential increase seen, though there is a much smaller time frame involved at present. Indeed, the number of events is relatively stable since 2001 to the present.

We can compare the overall correlations for all countries in all time points. If that were really high, it would give us some confidence that they were both measuring the same thing identically. But we know they are not, so this kind of comparison is less than informative. Scholars at Penn State have shown that in total, and for most countries in the Pacific Rim, there are more GDELT events than ICEWS events. These comparison are interesting, yet use an early version of the ICEWS data that does not take advantages of several years of improvements. Why? Because the newer data are not released. Without a doubt, the PSU team would have used that if it had been made available to them. However, the earlier data have been completely replaced by better data. That newer ICEWS data no longer uses the bag-of-words approach, but is based on more advanced NLP techniques developed by BBN. Importantly the CAMEO-based ontology has been further developed and amplified to provide better guidance for the identification and coding of events. As part of that process, subject matter experts also coded events and compared their own codings to those provided by ICEWS. This process revealed substantial improvement of the newer ICEWS framework over the early stage efforts in the first years of ICEWS. We have no desire to redo the massive comparisons that were undertaken earlier. But we do herein show a more modest comparison of GDELT and ICEWS data.

see Bryan Arva, John Beieler, Bejamin Fisher, Gustavo Lara, Philip A. Schrodt, Wonjun Song, Marsha Sowell, and Sam Stehle. "Improving Forecasts of International Events of Interest." In EPSA 2013 Annual General Conference Paper, vol. 78. 2013

Suffice it to say that it is enormously frustrating that so much of this research has not been released, or published.

Protest and demonstrations in Egypt and Turkey, and fighting in Syria provide a specific set of interesting cases on which to compare the widely available GDELT data with the latest event data used by the ICEWS project.