Comparing Big Event Datasets*

Patrick T. Brandt, Vito J. D'Orazio, Jared R. Looper
pbrandt@utdallas.edu, dorazio@utdallas.edu, jrl140030@utdallas.edu
School of Economic, Political, and Policy Sciences
University of Texas, Dallas

July 17, 2018

Abstract

Here we discuss different ways to access and utilize some of the new large event datasets that have come on-line in the last several years. We present an API that allows users to access the ICEWS, Cline Center, and our own Real Time event datasets coded in the CAMEO framework. We then explore some basic comparisons of these event datasets using both correlations and measures of mutual information across the datasets. We find that the changes in the coders used across these datasets produce different results, contrary to our expectations.

1 Introduction: the access problem

Political event data are machine coded from news reports for the quantitative study international interactions. Most of the automated event datasets in use are built on the Conflict Analysis and Mediation Event Ontology or CAMEO (Schrodt and Yilmaz, 2007). At present there are several large event datasets available. Past work done in this arena includes the Cline Center Event Data[1], which seeks to catalog post-World War II events by combing through news reports over a time range from 1945 to 2015 from sources such as the BBC Summary of World Broadcasts (SWB), the New York Times, and the Foreign Broadcast Information Service (FBIS). Additionally there is the the World-Wide Integrated Crisis Warning System Dataset (ICEWS) (Boschee et al., 2017), which seeks to provide an automated system to predict negative political conflict events for government actors.

Over time, political event data has incorporated changes in methodology to respond to the changing world of international interactions. One major change is the incorporation of non-state actors into analysis, as these actors, including terrorist groups and multinational corporations, have risen greatly in prominence and geopolitical importance in the past few decades. Additionally, political event data systems have incorporated

*Presented at the XXXVth Society for Political Methodology Annual Meeting at Brigham Young University, Provo, Utah. This material is based upon work supported by the National Science Foundation under Grant No. SBE-SMA-1539302. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Thanks to Kate Kim and Sayeed Salam for their assistance with the API and the R code discussed here. This paper builds on Looper's UT Dallas honors' thesis.

[1]This dataset is described at http://www.clinecenter.illinois.edu/data/event/.

technological developments in web-based access and expert systems to create online access to the datasets and automated or machine-assisted coding mechanisms. One elusive goal that past event data systems have not met, however, is the generation of near-real-time political event data (Schrodt, 2012).

In addition to these event data efforts, we have implemented many of the suggestions of Beieler et al. (2016) as part of a Spark-based Political Event Coding (SPEC) system (Solaimani et al., 2016). This system uses a distributed computing environment running on the Texas Advanced Computing Center's Jetstream cloud cluster (Stewart et al., 2015; Towns et al., 2014). This allows us to also create an automated political event data system that codes new events from over 384 different English newspaper sources on a daily basis in the CAMEO event ontology with geolocation of the events as well.

To access all of these data sources — the Cline datasets, ICEWS, and our real-time data — we have also created a server that hosts all of the datasets and provide an Application Programming Interface (API) to access the data at `evendata.utdallas.edu`. We did this for several reasons. First, it obviates the need for people to store the data locally, since these datasets are large and growing. For example, the ICEWS data is over 3.5GB of data. Second, most of the time users do not want the entirety of any of these datasets. The API then allows users to construct queries of the data (`https://github.com/Sayeedsalam/spec-event-data-server`).

These event datasets are aggregated into dyadic time series and compared using mutual information and Pearson correlations. Comparisons indicate that there are few similarities across event datasets, suggesting they are not interchangeable. We identify three factors that may contribute to these observed differences: (1) the use of the PETRARCH 2 event coder, which codes fewer events than the PETRARCH 1 or TABARI event coders; (2) the use of single source archives to code events from; and (3) short periods of overlap for some of the datasets. These findings reinforce the need for infrastructure and software development in event data research. There is clearly the need for improved event detection methods and access to large archives from which to extract data.

2 API and data details

The SPEC interface allows the data to be accessed by generating Mongo DB[2] queries and retrieving the data through a web-based interface described above. The REST API[3], a web-based interface that allows access of the data hosted on the SPEC server, has several benefits. It allows efficient, large-scale access of the data. The API also allows remote access and does not require the end-user to host any data on their

[2]The MongoDB query language is documented at `https://docs.mongodb.com/`.

[3]The description for the REST API can be found at `http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm`.

own servers. This creates the opportunity to perform analysis of large amounts of data without the need for large amounts of storage on the part of potential end-users. Another benefit of the REST API is that it does not need constant maintenance and can run without significant oversight. A further benefit is that the setup of the REST API requires an individualized key to access the data, allowing for controlled use. Prospective users of this real-time data can apply for an API key, combining ease of access with the need to ensure there are some checks to prevent malicious use of the data or direct denial of service attacks.

The API does come with drawbacks, however. The first is that the REST API itself has no way to generate the Mongo DB queries to access the data. Thus, they must be user-generated, meaning that without an external interface, any user wishing to access data from the API must have intimate knowledge of the Mongo DB query language. Another drawback is the way the data is generated and stored. After a web-based query is sent to the API to access the data, the data is presented to the user in a web page in a JSON format. This data is difficult to read or analyze without the presence of other tools to change the format of the data.

There are two main audiences that would need to access this data. The first audience is composed of users who are not high-throughput users and who do not seek to perform quantitative analysis. These users are typically high-level government officials, public office holders, and other users of interest who typically manage large departments or create policy rather than perform analyses. The second audience is composed of users who are high-througput users who are technically sophisticated and who want to perform quantitative analysis with the data. Typical members of this group are political scientists studying international interactions and government analysts working in defense or intelligence agencies who are creating predictive models for political conflict.

Each audience demands different solutions to the two aforementioned problems. For the first audience, the tool that interacts with the REST API needs to hide as much of the technical underpinnings from the user as possible. A solution to this problem has already been provided through the TwoRavens tool at `http://eventdata.2ravens.org/` (D'Orazio, Deng and Shoemate, 2018). This tool provides a graphical user interface that allows drag-and-drop selection of specific subsets of the data. This drag-and-drop method provides an easy way to generate the Mongo DB queries and does not require any technical sophistication on the part of the user. The data representation problem is solved by allowing users to download the data generated by the queries. Additionally, the TwoRavens interface already provides basic summary statistics so that the user does not necessarily have to download the data on their own computer.

The second audience, the technically sophisticated one, requires an access interface that allows its members to incorporate the data on the SPEC server into their already existing analysis. Additionally, with the availability of near-real-time data on the SPEC server, the access modality should be one that can take

advantage of these real-time updates in a way that is not cumbersome. For the two aforementioned problems, this means the interface should allow a dynamic and user-driven generation of queries and an easily integrated format of the data to allow easy adoption of this new technology.

We have created an R package UTDEventData (`https://github.com/KateHyoung/UTDEventData`) that will allow political scientists and analysts to easily analyze and compare any of the datasets in the SPEC server setup via `eventdata.utdallas.edu`. R was chosen as the programming language for this interface because of its common use among the political science community. Since one of the main purposes of the SPEC server is to provide an integrated and real-time event data repository, using the *lingua franca* of the scientific community allows for the maximum possible integration into already existing tools.

This R package solves the problem of automatic query generation for the REST API by creating a group of functions that allows the user to generate the queries on an ad hoc basis in the way that best suits the present task. The functions in this R package first generate queries for basic analysis blocks, including queries for retrieving events involving certain countries, certain time ranges, certain dyads of actors, and certain latitude-longitude locations. The R package also offers a function that allows the user to match any variable in the API table to any regular expression, allowing for unlimited functionality in the base package. This package further allows the end user to combine the aforementioned queries in any way using intersection and union, allowing for the access of any possible subset of the data.

3 Comparisons

For each of the data sources we can compare their contents using basic summary statistics. This gives a first-order impression of the similarity between the datasets and the various characteristics thereof. The first comparison we can do is a basic count of the number of events in each dataset, detailed in Table 1 below.

Table 1: Count of Events in Each Dataset, July 11, 2018

Dataset	Count	Time Coverage
Phoenix RT	725427	Mar 2017 - Present
FBIS	817955	Jan 1995 - Dec 2006
SWB	2906715	Jan 1979 - Dec 2015
NYT	1092211	Jan 1945 - Dec 2005
ICEWS	15220347	Apr 1995 - Dec 2015

Another way to compare the datasets is to look at the most common countries present in each dataset. Table 2 shows the ten most common countries in each dataset and the frequency that they appear as locations for events in the dataset. These calculations show the same countries appearing in each dataset among the most common countries. The country names are presented as ISO-3 codes. Three countries, the United

Table 2: Frequency of Most Common Countries

Phoenix RT	Percent	FBIS	Percent	SWB	Percent	NYT	Percent	ICEWS	Percent
USA	27.0%	USA	9.46%	RUS	7.36%	USA	37.8%	IND	6.57%
IND	5.6%	RUS	8.90%	USA	6.77%	GBR	5.7%	RUS	5.68%
PAK	4.5%	CHN	5.48%	CHN	5.61%	RUS	3.8%	USA	5.57%
GBR	4.3%	FRA	3.08%	AFG	4.70%	FRA	3.7%	CHN	4.34%
NGA	3.6%	IRQ	2.53%	IRN	4.55%	DEU	3.0%	GBR	2.60%
CHN	3.0%	IRN	2.37%	PAK	2.65%	ISR	2.4%	JAP	2.51%
RUS	3.0%	ISR	2.13%	IRQ	2.38%	CAN	2.1%	AUS	2.49%
SYR	2.4%	TUR	1.90%	UKR	2.24%	JAP	2.0%	IRN	2.43%
KOR	2.1%	PAK	1.73%	JAP	2.06%	CHN	1.6%	IRQ	2.31%
ISR	1.7%	JAP	1.73%	FRA	1.90%	VNM	1.5%	PSE	2.21%

States, Russia, and China, appear in each dataset among the most common countries. Seven more countries, Pakistan, the United Kingdom, Israel, France, Iraq, Iran, and Japan, appear in three datasets among the most common countries. There are several countries that appear only once in each dataset, but overall there is a great degree of similarity between the most common countries in each dataset even though the time coverage of the datasets sometimes differs substantially.

Additional variables of interest for comparing the datasets are the CAMEO code values and the derived pentaclass values. The pentaclasses code neutral events, verbal conflict, verbal cooperation, material conflict, and material cooperation from the CAMEO codes. These comparisons are shown in Figure 1, which compares the distribution across the different datasets of CAMEO Codes and pentaclass values. Both graphs tell a similar story. Overall, the distributions between the datasets are remarkably similar, with the CAMEO Codes of 01 (Making a Public Statement), 03 (Expressing an Intent to Cooperate), and 04 (Consulation), dominating each dataset and the pentaclass values of 0 (Neutral Interactions) and 1 (Verbal Cooperation) dominating each dataset. The fact that these values dominate across each dataset indicates that the nature of international cooperation of conflict has not notably changed over the course of the time periods the datasets catalog (1945–Present). Additionally, this shows a lack of bias with respect to the types of events each dataset records. These similarities indicate that the datasets are similar enough to present a sort of continuity across time, even in spite of the different source texts and different coding mechanisms.

The Real Time (RT) dataset is unique in its breadth of sources. Currently, the Real Time dataset draws from a list of 384 different news sources daily to automatically generate events. Table 3 below presents the most common sources present in the dataset. News aggregators dominate the most common sources, with seven of the 10 most common being aggregators (the various World News Network websites, denoted by the prefix wn, and Google News).

Another way to analyze the sources is to examine the distributions of the pentaclass value present for the most common sources. The distributions of these classes are detailed in Figure 2. The graph on the

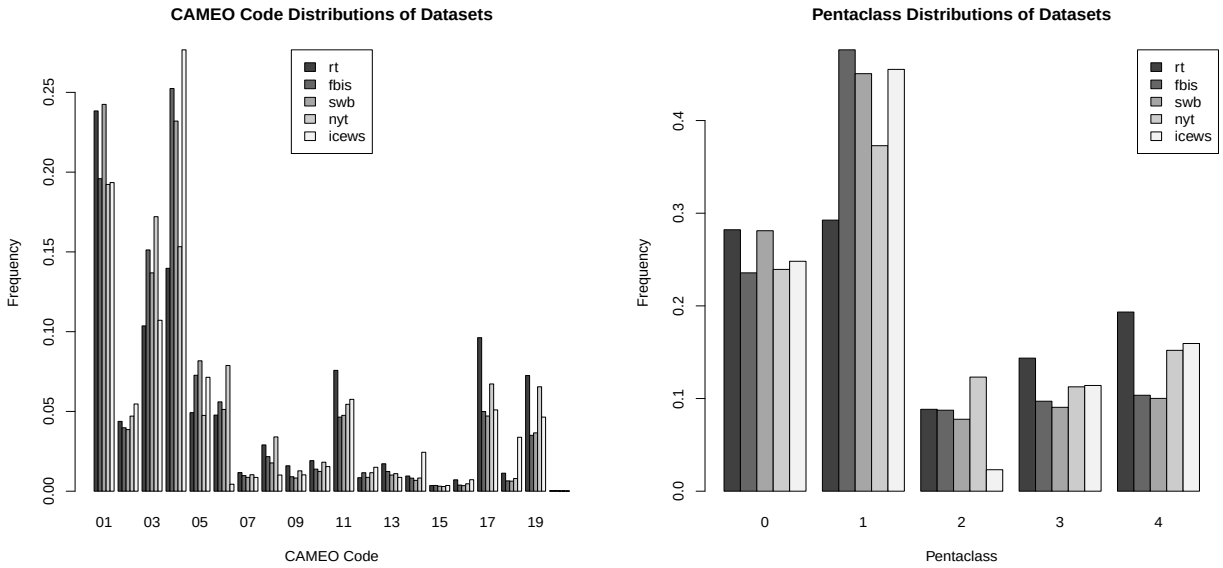

Figure 1: The graph on the left compares the distributions of CAMEO Code values for the five aforementioned datasets. The graph on the right compares the distributions of the pentaclass values.

Table 3: Frequency of Sources in RT Dataset

News Source	Frequency
wn_world	17.8%
wn_politics	13.0%
ap	5.2%
google	4.9%
wn_africa	4.2%
wn_asia	3.5%
wn_europe	2.6%
sfgate_national	2.2%
wn_mideast	2.1%
nzherald_world	1.6%

left of Figure 2 demonstrates that the counts of each pentaclass value are dominated by the events from the most common source, World News. The graph on the right demonstrates that even though this domination in the raw count of events occurs it does not bias the data toward the characteristics of one source. This is because the distributions of pentaclass values over each news source is remarkably similar. This means that the Real Time dataset, at least with respect to the pentaclass values, is not biased with respect to certain news sources among the 384.

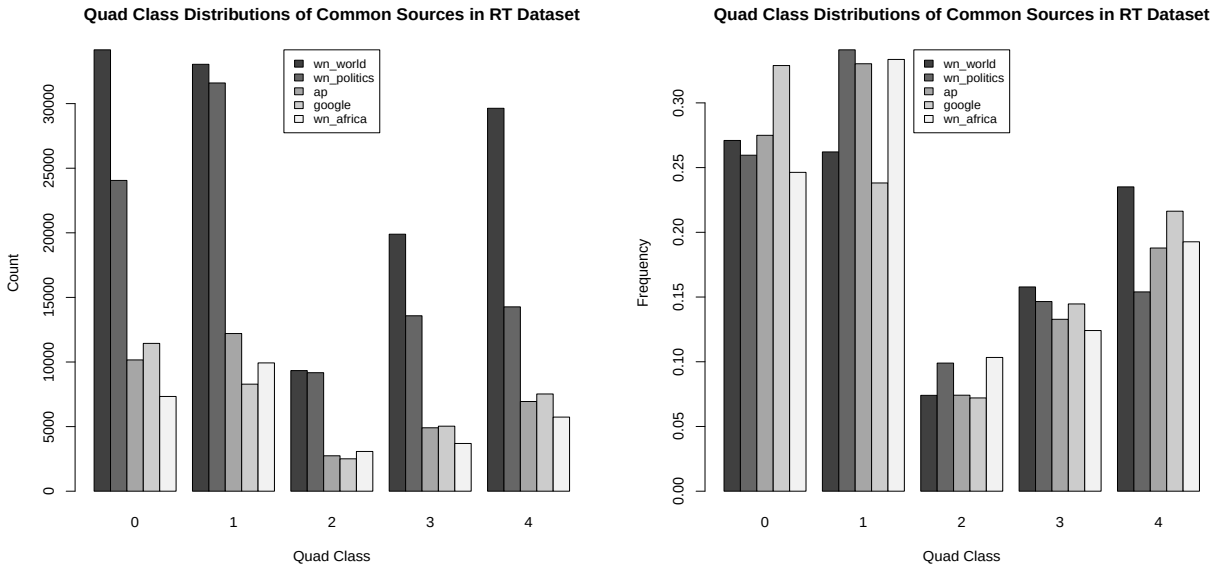

Figure 2: The graph on the left compares the raw counts of the pentaclass values of events from the five most common news sources in the real time dataset. The graph on the right compares the normalized distributions of pentaclass values for the same news sources.

4 Aggregate Time Series Comparisons

For each of our data sources we can generate aggregate time series of the levels of comments, verbal cooperation events, material cooperation events, verbal conflict events, and material conflict events. This collapse of the CAMEO categories captures much of the dynamics and variation in the event data and is a common approach when working with event data (Beger, Dorff and Ward, 2016).

The ICEWS data is the most established, so we begin with a time-series plot of the USA-RUS dyad for purposes of validity. Here we have scaled the counts of the cooperation measures positively, and those for the conflict measures negatively. Figure 3 shows the time series. The majority of events are verbal cooperation, and appear to occur regularly. Verbal conflict events are less regular and tend to spike sharply. The material cooperation/conflict categories contain fewer events. These patterns follow what one might reasonably expect for the USA-RUS dyad.

Time series are also shown for the Cline NYT and SWB collections in Figures 4 and 5. The NYT has very few events, making it difficult to see comparable patterns to that of ICEWS, which does contain known duplicat reports. However, it is clear that verbal cooperation and conflict are the dominant categories. SWB has more events than NYT, but considerably less than ICEWS. Here, the trend for verbal conflict appears similar to that of ICEWS, while verbal cooperation looks to be more serially correlated. The face validity of the NYT and SWB data, overall, is not as strong as that of ICEWS.

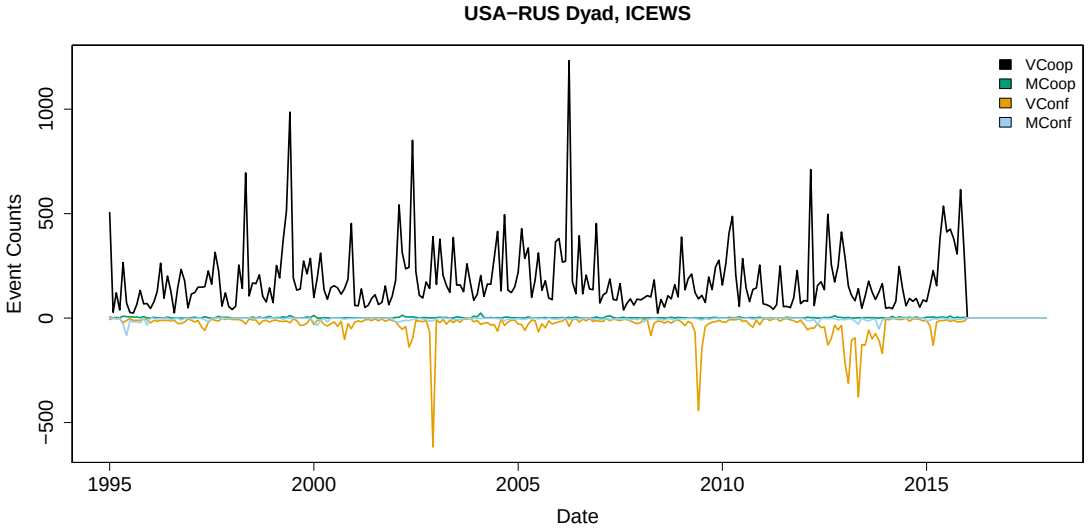

Figure 3: USA-RUS Time Series, ICEWS

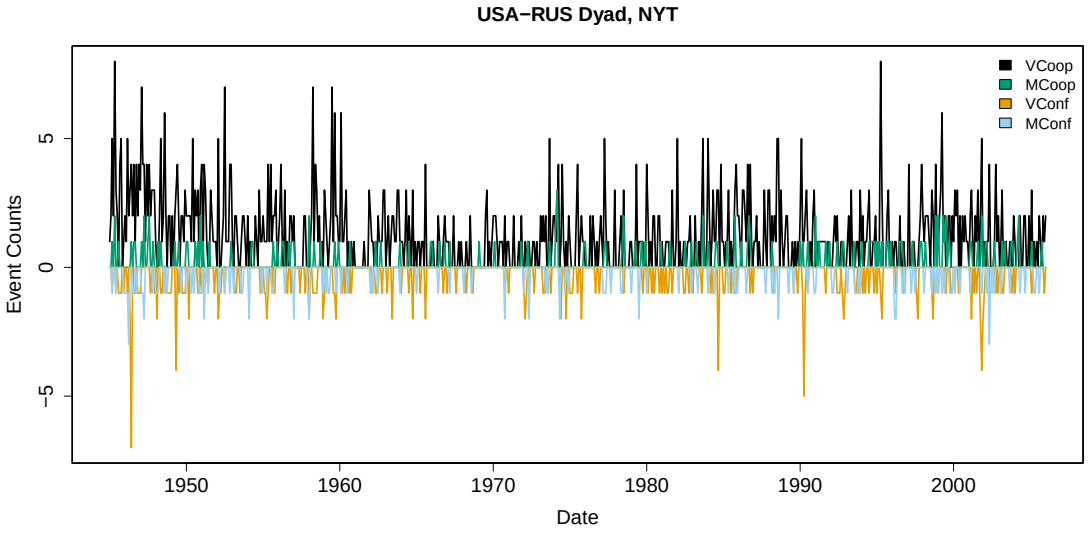

Figure 4: USA-RUS Time Series, NYT

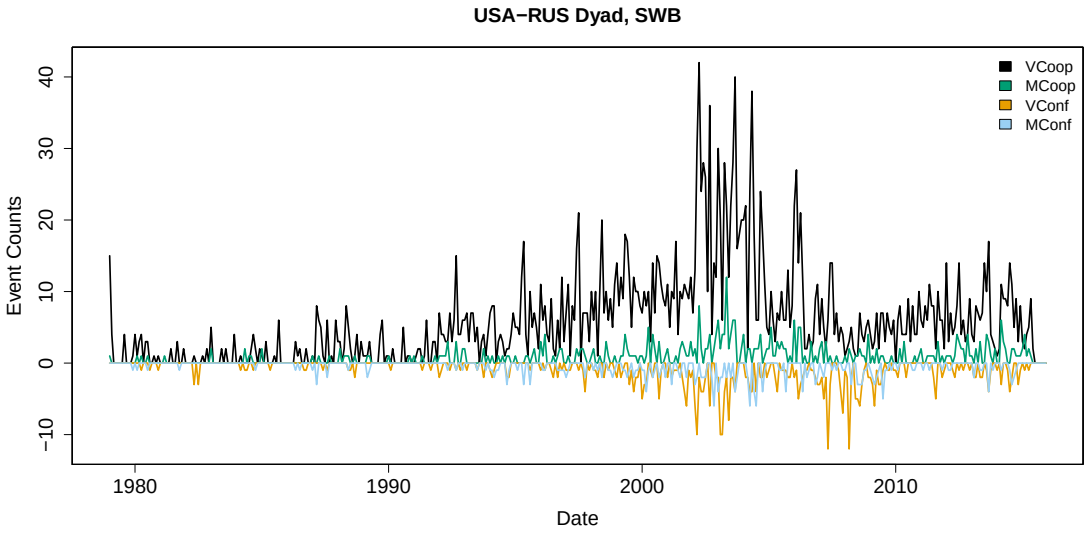

Figure 5: USA-RUS Time Series, SWB

Additional time series plots for the USA-ROK and USA-PRK dyads are shown in the Appendix. The takeaway is the same: ICEWS trends are roughly what one would expect, while the NYT and SWB have somewhat less face validity. SWB appears to have a large increase in the number of events across all dyads in the early 2000s, but this increase dissipates over time.

We compare these four time series for the USA-RUS dyad for each pair of event data collections. Simple product moment correlations or Pearson correlations are valid measures of association under assumptions about linear relationship and convergence to normal distributions. But that is not the case for the event data we have here, which tend of be sparse, serially correlated and cross-variable correlated. So simple measure of correlation will not serve us well here. Further, as Brandt et al. (Forthcoming) show, there are non-linear, non-normal mixture distributions that do a good job explaining these kinds of event data.

A more general measure of association for a multivariate time series is mutual information (MI). This measure, used in statistics / econometrics (e.g., Dionisio, Menezes and Mendes, 2004), and ecology (e.g., Cazelles, 2004; Ardón et al., 2017) applications generalizes the idea of a correlation coefficient when the marginal density of the data is unknown. For this non-parametric measure of association, we begin with two time-series data matrices X and Y, say each a set of pentaclass time series of length T. Let $H(X)$ and $H(Y)$ be the empirical marginal entropies for each data source matrix. The unstandardized mutual information is then defined for the densities over $p(x), p(y)$, and $p(x, y)$ as

$$I(X,Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \quad (1)$$

$$= \int\int p(x,y) \cdot log\left(\frac{p(x,y)}{p(x)p(y)}\right) \partial x \partial y. \quad (2)$$

This measure is bounded so that $I(X,Y) \geq 0$. So it functions like a covariance score. We can standardize the mutual information measure to be between zero and one via,

$$MI(X,Y) = \frac{H(X) + H(Y) - H(X,Y)}{\sqrt{H(X)}\sqrt{H(Y}} \tag{3}$$

The superiority of Equation 3 is that it captures both linear and non-linear dependence in the presence of serial correlation, without making strong parametric assumptions about the data being compared (Dionisio, Menezes and Mendes, 2004). Mutual information scores have been implemented in `R` via Scheuerell (2017) and allow for different lag structures across the datasets. This allows for both leads and lags of the data to see if there are differences across say different news wires or event data source reports over time. To get a critical value for the MI measure in Equation 3, a Monte Carlo procedure is implemented for an $\alpha = 0.05$ level of significance. We can then see if an estimated MI value exceed the critical value for this α level of significance. The default is to use a Monte Carlo sample size of 100 draws.

Table 4 shows the MI measure for each pair of event datasets for the USA-RUS dyad, for each pentaclass type. This normalized MI ranges from 0 to 1, and is significant if the value exceeds the threshold value. Only one MI measure for this dyad is significant, and the MI measures are considerably low. This pattern—very few significant MI measures with very low MI values—is consistent across the dyads we have examined. This includes: USA-ROK; USA-PRK; CHN-JPN; CHN-VNM; RUS-UKR; RUS-SYR; and, USA-SYR. Pearson correlations are shown in Table 5. Here, we can see similarly low values. As with MI, other dyads we have examined have low correlation values.

Table 4: Mutual Information: USA-RUS Dyad

	sources	Comments	VCoop	VConf	MCoop	MConf
1	nyt-fbis	**0.165**	0.045	0.112	0.067	0.075
2	nyt-swb	0.016	0.035	0.038	0.01	0.036
3	nyt-icews	0.027	0.028	0.081	0.038	0.062
4	nyt-phox	–	–	–	–	–
5	fbis-swb	0.056	0.038	0.071	0.103	0.1
6	fbis-icews	0.04	0.03	0.045	0.06	0.069
7	fbis-phox	–	–	–	–	–
8	swb-icews	0.026	0.01	0.025	0.032	0.024
9	swb-phox	–	–	–	–	–
10	icews-phox	–	–	–	–	–

Figure 6 shows the rolling correlations for SWB-ICEWS. Addition rolling correlations for NYT-SWB and NYT-ICEWS may be found in the Appendix. The window for rolling correlations has been set to five. The horizontal dashed line is zero. Again, we see very little evidence of consistency for the USA-RUS dyad across any of the pentaclasses. If the data were similar, we would see consistent, positive correlations. This would

Table 5: Pearson Correlations: USA-RUS Dyad

	sources	Comments	VCoop	VConf	MCoop	MConf
1	nyt-fbis	0.173	0.079	0.168	0.026	-0.037
2	nyt-swb	**0.157**	-0.016	0.007	0.006	0.011
3	nyt-icews	-0.094	0.023	0.015	0.06	-0.034
4	nyt-phox	–	–	–	–	–
5	fbis-swb	0.098	0.03	0.07	0.106	0.082
6	fbis-icews	0.117	0.115	-0.061	-0.058	-0.117
7	fbis-phox	–	–	–	–	–
8	swb-icews	-0.02	**0.167**	-0.017	0.026	0.024
9	swb-phox	–	–	–	–	–
10	icews-phox					

indicate that as we window over the time range, the two time series being compared are trending together. With the exception of a few small patches (which could be due to random chance), this is not the case in any of the five rolling correlations shown in Figure 6.

There are three related explanations for the low MI and correlation values. First, the Cline Center data and the Real Time Phoenix data have been coded with Petrarch 2, which is known to produce fewer events than Petrarch 1 or its predecessor, TABARI. The drop in the number of events will lower values for associations. Second, with the exception of ICEWS and SWB, the temporal overlap is not extensive. Third, each Cline Center set has been coded from a single source, as opposed to ICEWS and the Real Time Phoenix datasets which use many sources. Each of these three explanations mean that there are simply fewer events per dyad, and thus weaker associations. Further, the ICEWS data are coded with the BBN Accent coder, which itself is a variant of TABARI with different dictionaries and coverage.[4]

The takeaway from this comparison has a simple but important point to reinforce: automated event coding benefits from large volumes of data. This includes larger volumes of source materials, which in turn means larger volumes of coded events. It is necessary to build and support infrastructure for processing big event data efforts.

5 Conclusion

We provide access to large scale political event datasets through our database and API at `eventdata.utdallas.edu`. Currently, we host ICEWS, Cline NYT, Cline SWB, Cline FBIS, and Phoenix Real-Time, and hope to expand to additional new event datasets as well as existing event datasets that have been used in past research (e.g., Shellman, Levey and Young (2013), Schrodt and Gerner (2004), Pevehouse and Goldstein (1999)). By improving access and the availability of event data collections, we hope to increase use of these

[4] `https://www.iarpa.gov/index.php/newsroom/press-releases-and-statements/882-iarpa-announces-bbn-accent-release-to-the-resear`

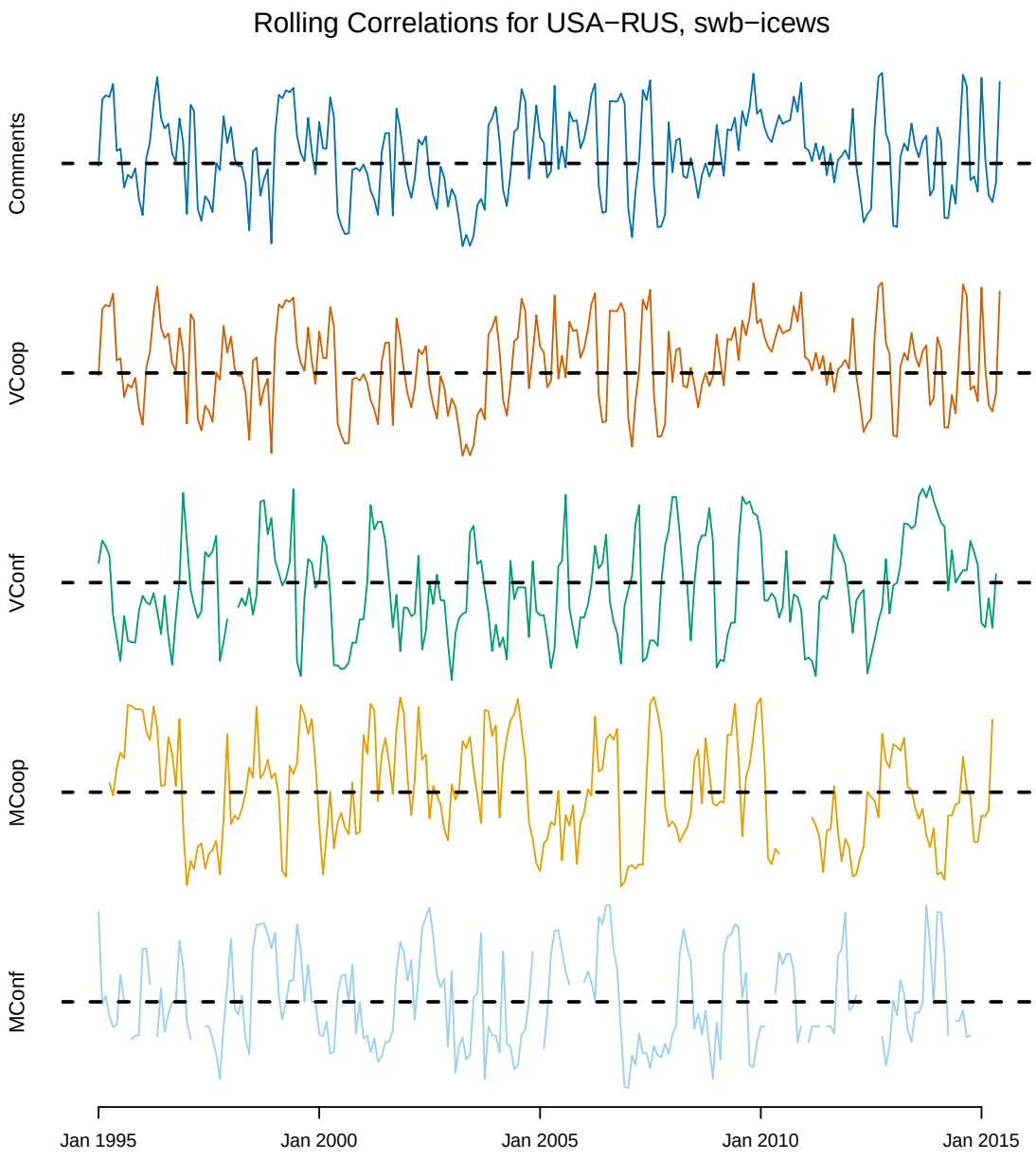

Figure 6: USA-RUS Rolling Correlations, SWB-ICEWS

resources to address complex social research questions.

The five datasets were compared with basic descriptive statistics. All five have global coverage, although we show that countries are disproportionately represented in each. Each collection has different temporal coverage, with the NYT collection being the longest from 1945-2005. The ICEWS data contains the most events, due to the large number of sources and extensive actor dictionaries that were part of this effort.

We built dyadic time series and compared the collections using mutual information and Pearson correlations. The ICEWS data shows reasonably strong face validity, while the NYT and SWB less so. Comparisons indicate that differences exist in the time trends across the collections. We expect this is due primarily to the sparse event counts in the single-source collections. Overall, the lack of similarity over the time series reinforces the need for infrastructure and open-source software development to improve event data research.

Appendix

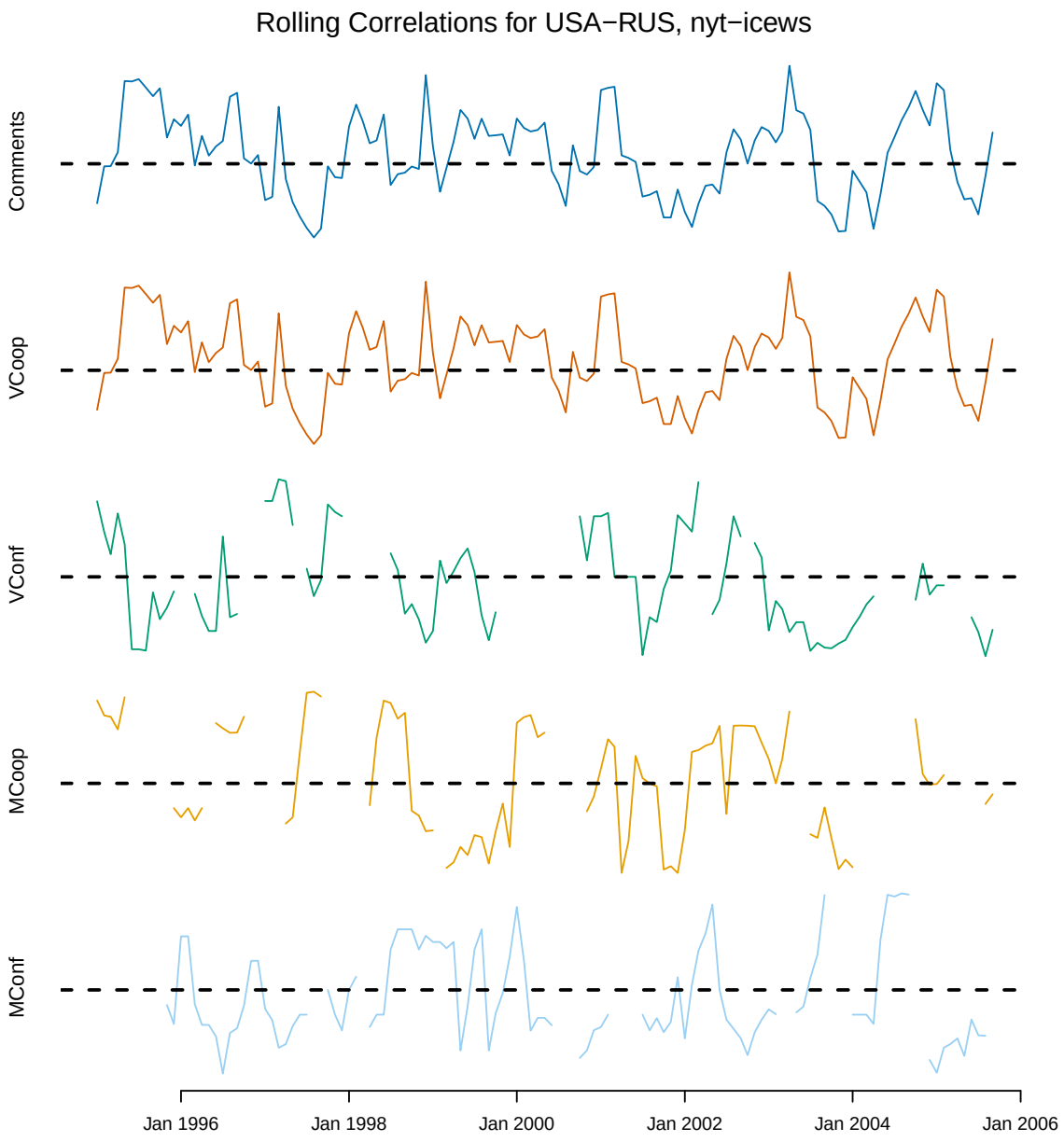

Figure 7: USA-RUS Rolling Correlations, NYT-ICEWS

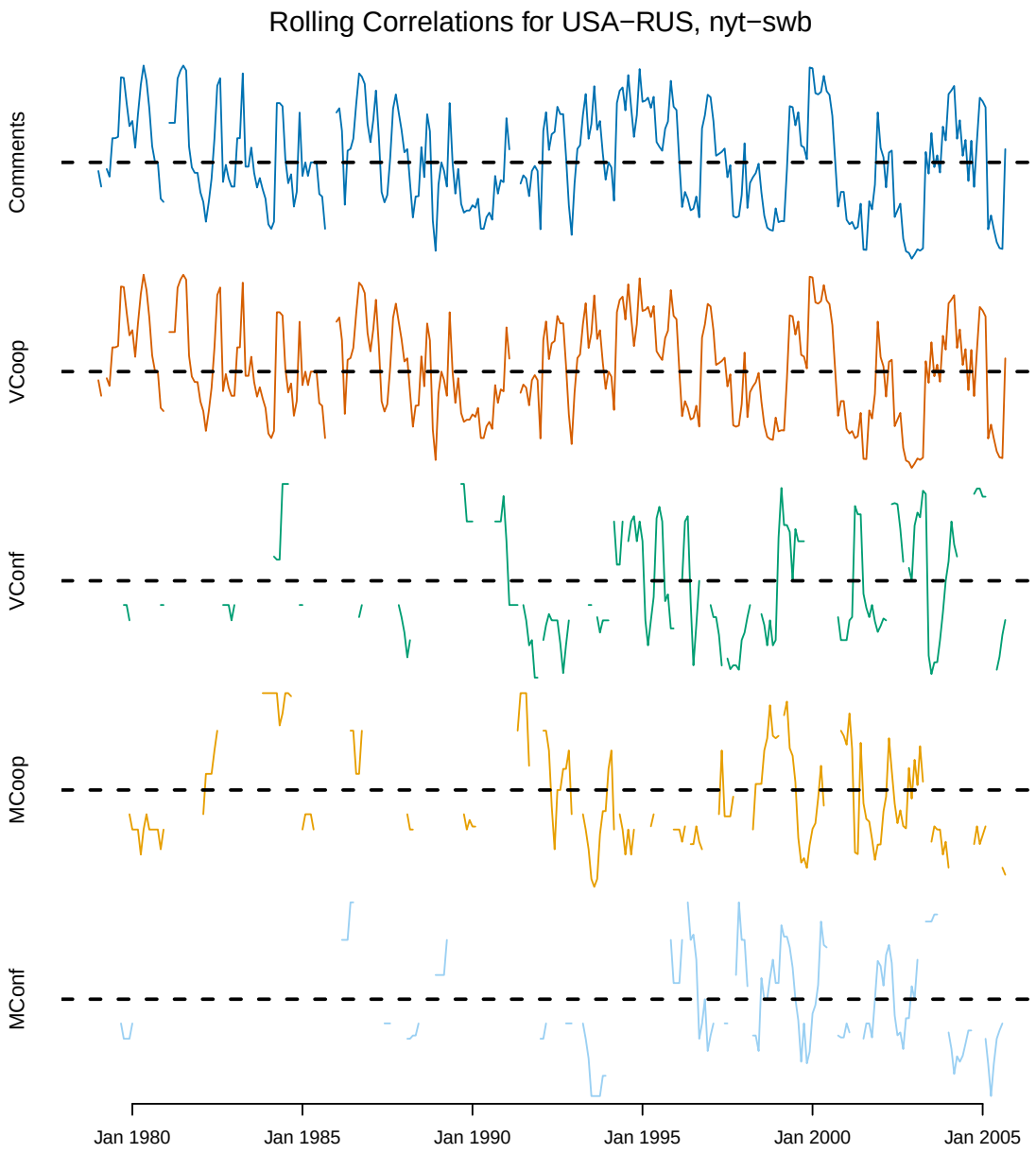

Figure 8: USA-RUS Rolling Correlations, NYT-SWB

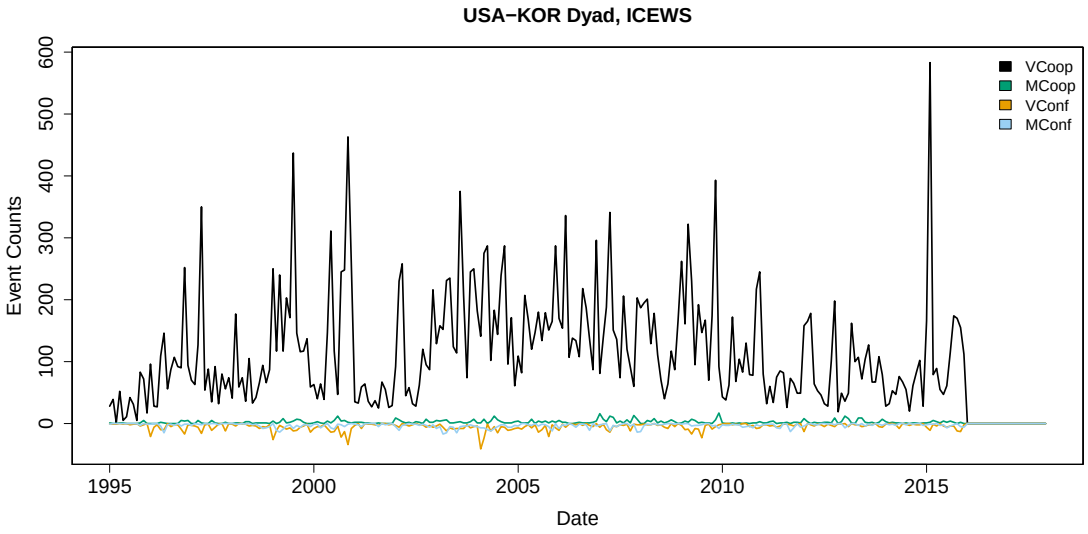

Figure 9: USA-KOR Time Series, ICEWS

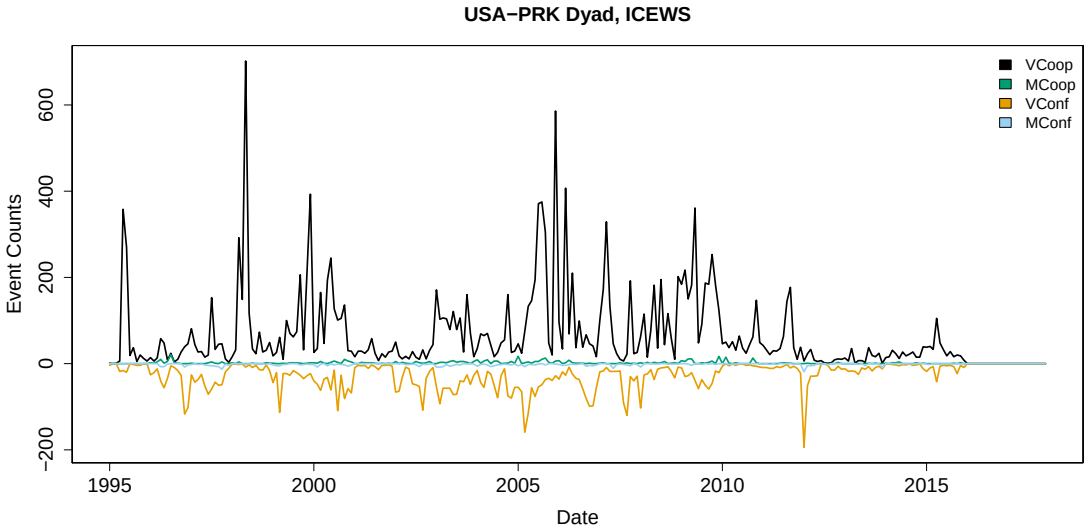

Figure 10: USA-PRK Time Series, ICEWS

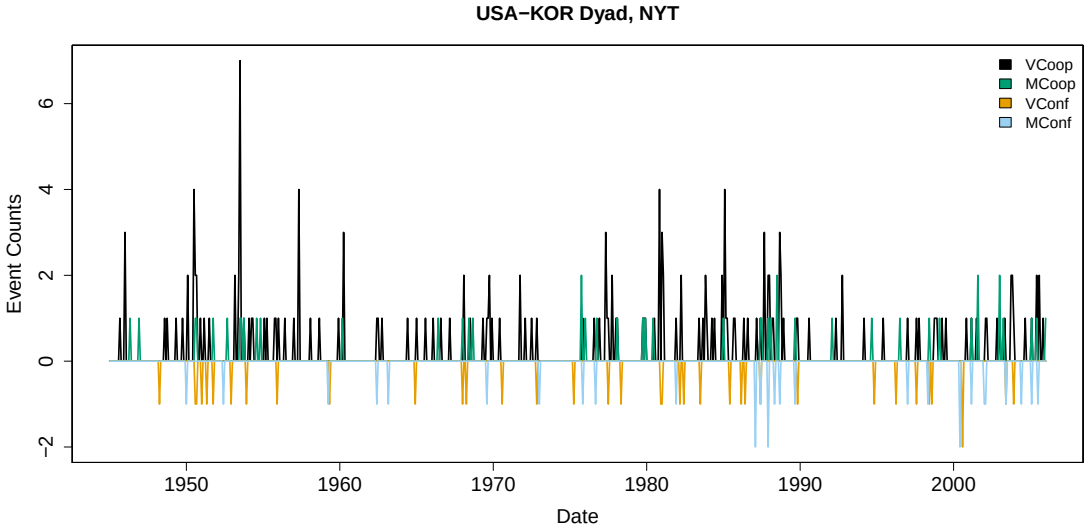

Figure 11: USA-KOR Time Series

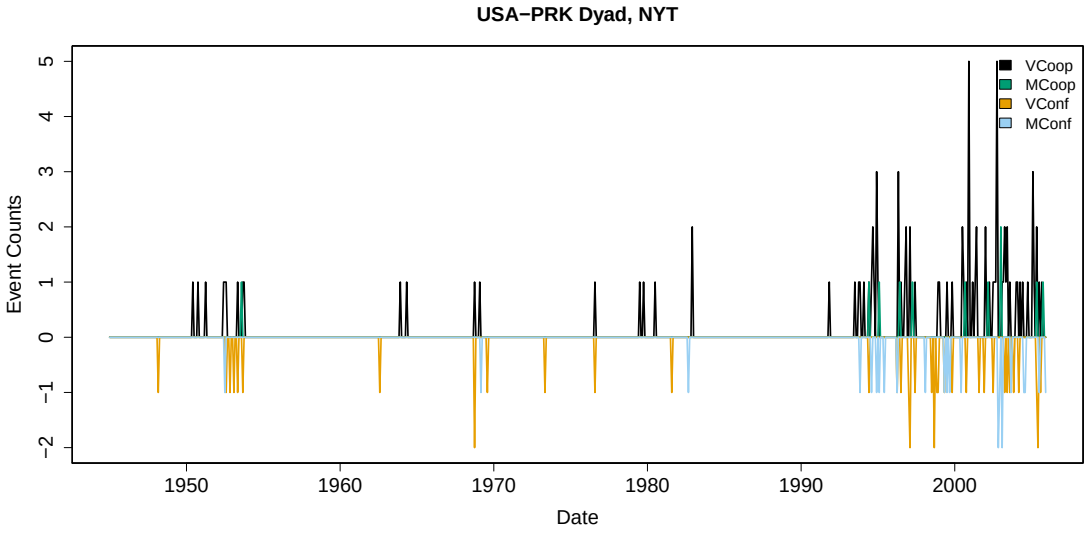

Figure 12: USA-PRK Time Series, NYT

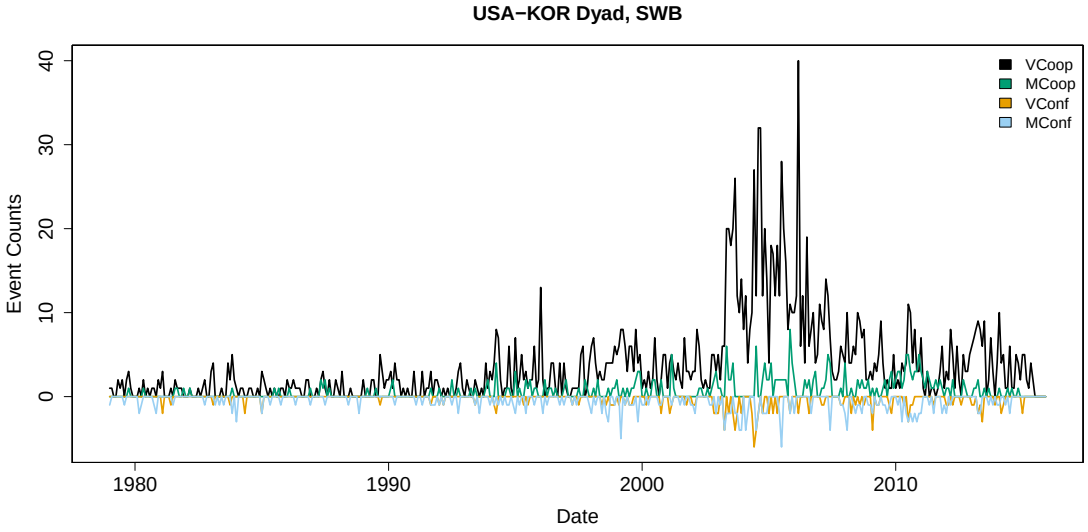

Figure 13: USA-KOR Time Series, SWB

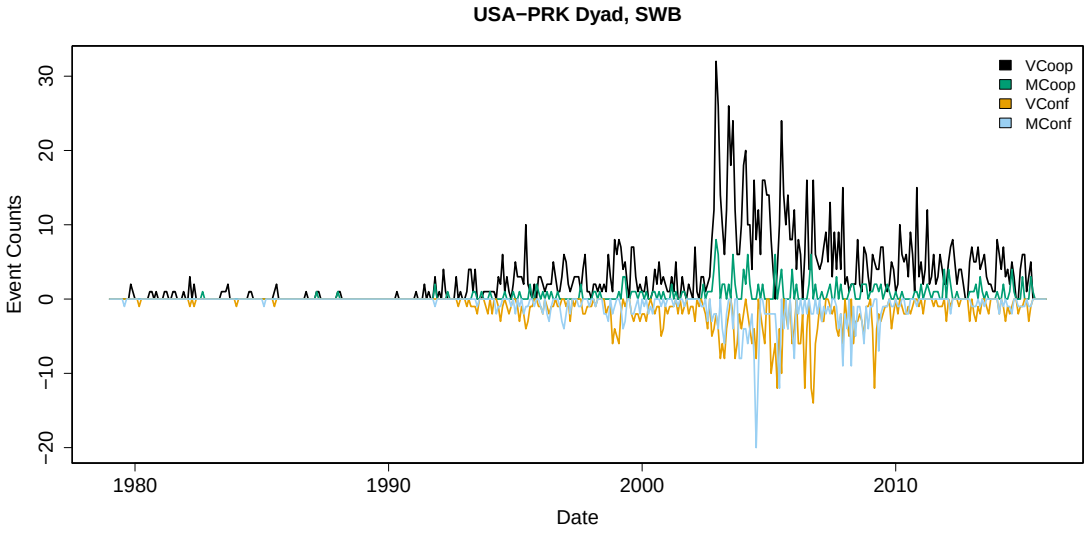

Figure 14: USA-PRK Time Series, SWB

References

Ardón, Marcelo, Ashley M Helton, Mark D Scheuerell and Emily S Bernhardt. 2017. "Fertilizer legacies meet saltwater incursion: challenges and constraints for coastal plain wetland restoration." *Elem Sci Anth* 5.

Beger, Andreas, Cassy L Dorff and Michael D Ward. 2016. "Irregular leadership changes in 2014: Forecasts using ensemble, split-population duration models." *International Journal of Forecasting* 32(1):98–111.

Beieler, John, Patrick T Brandt, Andrew Halterman, Philip A Schrodt and Erin M Simpson. 2016. "Generating political event data in near real time." *Computational Social Science* p. 98.

Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz and Michael Ward. 2017. "ICEWS Coded Event Data.".
URL: *http://dx.doi.org/10.7910/DVN/28075*

Brandt, Patrick T., John R. Freeman, Tse min Lin and Philip A. Schrodt. Forthcoming. "A Bayesian Time Series Approach to the Comparison of Conflict Dynamics." *Political Science Research and Methods* .

Cazelles, Bernard. 2004. "Symbolic dynamics for identifying similarity between rhythms of ecological time series." *Ecology Letters* 7(9):755–763.
URL: *http:https://doi.org/10.1111/j.1461-0248.2004.00629.x*

Dionisio, Andreia, Rui Menezes and Diana A Mendes. 2004. "Mutual information: a measure of dependency for nonlinear time series." *Physica A: Statistical Mechanics and its Applications* 344(1-2):326–329.

D'Orazio, Vito, Marcus Deng and Michael Shoemate. 2018. TwoRavens for Event Data. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE.

Pevehouse, Jon C and Joshua S Goldstein. 1999. "Serbian compliance or defiance in Kosovo? Statistical analysis and real-time predictions." *Journal of Conflict Resolution* 43(4):538–546.

Scheuerell, M. D. 2017. *muti: An R package for computing mutual information.*
URL: *https://github.com/mdscheuerell/muti*

Schrodt, Philip A. 2012. "Precedents, Progress, and Prospects in Political Event Data." *International Interactions* 38(4):546–569.
URL: *https://doi.org/10.1080/03050629.2012.697430*

Schrodt, Philip A and Deborah J Gerner. 2004. "An event data analysis of third-party mediation in the Middle East and Balkans." *Journal of Conflict Resolution* 48(3):310–330.

Schrodt, Philip A. and Omur Yilmaz. 2007. "Conflict and Mediation Event Observations (CAMEO) Codebook." `http://eventdata.parusanalytics.com/data.dir/cameo.html`.

Shellman, Stephen M, Brian P Levey and Joseph K Young. 2013. "Shifting sands: Explaining and predicting phase shifts by dissident organizations." *Journal of Peace Research* 50(3):319–336.

Solaimani, Mohiuddin, Rajeevardhan Gopalan, Latifur Khan, Patrick T Brandt and Bhavani Thuraisingham. 2016. Spark-based political event coding. In *Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on*. IEEE pp. 14–23.

Stewart, Craig A., Timothy M. Cockerill, Ian Foster, David Hancock, Nirav Merchant, Edwin Skidmore, Daniel Stanzione, James Taylor, Steven Tuecke, George Turner, Matthew Vaughn and Niall I. Gaffney. 2015. Jetstream: A Self-provisioned, Scalable Science and Engineering Cloud Environment. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*. XSEDE '15 New York, NY, USA: ACM pp. 29:1–29:8.
URL: *http://doi.acm.org/10.1145/2792745.2792774*

Towns, John, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D Peterson et al. 2014. "XSEDE: accelerating scientific discovery." *Computing in Science & Engineering* 16(5):62–74.