Event Data, Civil Unrest and the SPEED Project

Peter F. Nardulli, Kalev H. Leetaru and Matthew Hayes

Cline Center for Democracy
University of Illinois
Suite 207 Strata Building
2001 South First Street
Champaign IL 61822

(nardulli@illinois.edu; leetaru@illinois.edu; mjhayes2@illinois.edu)

Within the social sciences and humanities there is a growing interest in automated analysis of textual materials. Most recently, the field of "culturomics" has exploited the massive volumes of available digital text to study everything from language evolution to political censorship (Michel 2011). Since the 1962 release of the General Inquirer, the first general-purpose computational content analysis platform (Stone 1962), scholars have leveraged increasing computing capacity to analyze vast volumes of text with unprecedented speed. Political scientists were among the early users of automated textual analysis (Holsti 1964) and this interest has continued to the present day (Monroe and Schrodt 2008). This research approach would be impossible without the extraordinary advances in natural language processing and automated content analysis. In the past analyzing textual materials has been limited by the prohibitive costs of manually transforming large bodies of text into quantitative data. Moreover, human content analysis can introduce a degree of unreliability into the transformation process that undermines its utility (Laver, Benoit, and Garry 2003; Mikhaylov, Lerner, and Benoit 2008; Schrodt, Davis, and Weddle 1994; Stone 1962).The vast increases in digitized texts and the development of automated methods of information extraction have addressed both of these problems – scholars can now rapidly generate quantitative datasets from novel sources with perfect reproducibility. Moreover, at least one experiment suggests that, at least with respect to making very basic textual distinctions, automated systems can perform some tasks on par with human coders (King and Lowe 2003).[1]

Despite the impressive progress that has been made in this field and the undeniable allure of automation, the undiscriminating use of fully automated methods of textual analysis comes at an often unacknowledged cost. To wit, the amount of useful information that is available in textual sources often far exceeds that which can be extracted accurately using fully automated systems. This is particularly true for information extracted from news reports, a source that is of particular interest to social scientists and has long been an important source of data on key political events. The amount of useful information in most news report far exceeds the current capacity of automated systems to extract it accurately. This is particularly true if historical materials (i.e., those that were not "born digital") are involved: optical character recognition (OCR) programs introduce typographical error that undermines the accuracy and utility of fully automated procedures. Thus, depending upon the needs of the research, automated information extraction may be not be optimal. When the costs of fully automated systems exceed their benefits, researchers should employ hybrid systems that use technological applications to enhance the speed, accuracy and reliability of human coders in extracting rich bodies of information from texts. It should also be stressed that developing hybrid systems can be an important intermediary step in building increasingly automated systems. They can provide insights into the utility of various types of data and provide the training data needed to develop fully automated systems.

This paper introduces a hybrid system that has been developed over the past five years: the Social, Political and Economic Event Database (SPEED) project.[2] SPEED is a technology-

[1] Although the automated system that was evaluated performed well in differentiating among very broad categories of events, it fared poorly at detecting non-events (i.e., textual references that did not belong in any category). Human coders correctly identified that news reports contained no event at least 92% of the time, compared to 23% for the computer.

[2] SPEED is an integral part of a broader effort to understand the role of national institutions and contexts in societal development that focuses on 165 countries in the Post WWII era (1946 – present), the Societal

intensive, protocol-driven system designed to generate a rich body of event data on a range of topics (civil unrest, property rights, political expression, supremacy of law, etc.) from a global archive of news reports that spans the post-WWII era. SPEED is a hybrid system in that the efforts of human operators are aided by a variety of technologies. Within SPEED various technologies are used to assemble a digitized news archive, identify news reports containing information on relevant events, and highlight relevant textual passages within individual news reports. Technology is also used to facilitate the extraction of geographic information, the identification of the proper names socio-cultural groups and individuals, and the linking of related events. In addition, SPEED provides for the monitoring of coder reliability.

The next section introduces the SPEED project, with a particular emphasis on one protocol: the Societal Stability Protocol (SSP). The SSP focuses on small-bore civil unrest events (protests, strikes, politically motivated attacks, disruptive state acts, irregular transfers of power, etc.). The second section provides a brief comparison between SPEED and other event data projects concerned with civil unrest. The third section uses SSP data to underscore the need for small-bore data on civil strife by critiquing the current focus on civil wars. The fourth section examines the utility of employing a hybrid system such as SPEED's SSP to the study of civil strife, an increasingly important field of inquiry. The fourth section provides some concluding remarks.

The SPEED Project and the Societal Stability Protocol

To design an project that would provide the basis for generating significant advances in our understanding of civil unrest we had to address several basic challenges. The responses to these challenges shaped SPEED and its SSP, providing the latter with a distinctive set of features that affects its capacity to generate high-quality event data that can be used to meet a variety of research needs. Our responses to five challenges were particularly important for shaping this capacity. First, we had to assemble a comprehensive archive of digitized news reports for virtually every country in the world for the entire post-war era. Second, we had to construct a comprehensive destabilizing event ontology that would structure the search for relevant events within the news archive. Third, we had to devise electronic search procedures to identify relevant news reports within the news archive and relevant textual passages within the news reports. Fourth, we had to: (1) decide what type of event-specific information was needed (and available) to advance our understanding of civil unrest, and (2) create efficient and reliable procedures for extracting that information from news reports. Fifth, we had to develop training and quality control procedures that would foster the generation of high-quality data. Our approach to each of these challenges is discussed below.

SPEED's Global News Archive

To meet the information needs of the SPEED project we had to assemble a comprehensive set of global news sources for the post-1945 period. Since 2006 SPEED's **SEARCH** routine has been crawling across news websites (over 5,000 news feeds in 120 countries) several times each day, scraping news reports and storing them on our server. It currently adds an additional 100,000 articles each day. Acquiring news sources before 2006 required a different approach. We secured

---

Infrastructures and Development project (SID); for more information on SID see: http://www.clinecenter.illinois.edu/research/sid-project.html.

the digitized archives of the New York Times and Wall Street Journal for the 1946-2006 period. However, these were not deemed to have sufficient international coverage. Thus, we secured microfiche and microfilm records for two intelligence agency news services: the Foreign Broadcast Information Service (CIA) and the Summary of World Broadcasts (BBC). These contain millions of news articles and broadcasts that were translated into English from scores of languages. These news reports were derived from tens of thousands of news outlets and cover a range of developments in every country in the world. Adding these news reports led to a highly inclusive historical news archive containing tens of millions of reports.

The Societal Stability Event Ontology

In order to maximize the utility of the Cline Center's global news archive in advancing our understanding of civil unrest, it was essential to have a comprehensive ontology of destabilizing events. To construct this ontology we began by surveying event categories that had been used in such diverse fields as political violence, terrorism, political instability, and social movements. The initial ontology was then used in conjunction with a multi-year pretest involving tens of thousands of global news reports randomly selected from the post-WWII era. This led to a more inclusive and refined ontology, which is reported in Figure 1. There are five tier-one categories in this ontology: political expression events, politically motivated attacks, disruptive state acts, political power reconfigurations, and mass movements of people. Each of these has between one and three tiers of categories below the first tier.[3] The most distinctive features of the SSP ontology are its scope, level of refinement, and its inclusion of small-bore events. Compiling event data on small-bore indicators is viewed as an efficient way generating insights into the factors and processes that generate more salient and destabilizing developments.

Automated Search Procedures – The BIN and EAT Programs

A basic challenge faced by all large-scale event coding projects is processing the volume of the textual information from which quantitative, event-specific data are extracted. Isolating those news reports that contain relevant information and identifying the relevant text within those news reports are particularly formidable tasks for the SSP, both because of the size of SPEED's global news archive and the scope of the destabilizing event ontology. To accomplish these tasks the SPEED project has invested in the development of technologies that employ natural language processing principles (NLP) to automate its search processes. The first of these is **BIN,** an automated text classification system that screens news reports. The second is the Event Annotation Tool (**EAT**). The **EAT** module scans the screened news reports and highlights relevant text.

BIN uses statistically based algorithms based on key words, word correlations, and semantic structures to identify relevant reports. It generates statistical probabilities that a news report contains information on an event that falls within the SSP's ontology. A news report is assigned to the "relevant" category if that probability is sufficiently high; if it falls below the inclusion threshold it is assigned to the "discard" bin. **BIN**'s algorithms were developed by using thousands of human-

[3] A document detailing the operational definitions of the event in this ontology can be accessed at the following address: http://www.clinecenter.illinois.edu/research/publications/SPEED-Definitions-of-Destabilizing-Events.pdf.

categorized reports to "teach" the computer to recognize the semantic attributes that characterize relevant reports.[4] It has proven to be very robust. Thresholds for inclusion were set low, so as not to discard relevant news reports. Consequently, tests examining random samples of discarded news reports, suggest that **BIN** has a false negative rate of 1-4%, depending upon the source of the news report (NYT, FBIS, SWB, etc.). Thus, it identifies about 96-99% of the relevant reports. Thus, at a very small cost, **BIN** enables human operators to process huge amounts of text in an efficient manner.

While the **BIN** system is absolutely essential to the success of SPEED's hybrid approach to event analysis, even perfectly binned reports leave humans with a formidable amount of text to process. This poses serious cognitive challenges for human coders.[5] Thus, the **EAT** module is being developed to meet these challenges. It employs a variety of NLP-based computational procedures to annotate (i.e., highlight) relevant segments of text within news reports. **EAT** operates at the level of words and phrases and builds cognitively based models that identify such things as "trigger" words that demarcate a reference to such things as events, geographic locations, dates, actors, etc. Building these cognitive models is a tedious process and it is based on training data generated by trained human coders who use **EAT**'s utilities to highlight key words and phrases that will be used in model building. When properly calibrated, **EAT**'s annotations will greatly enhance the efficiency, accuracy and reliability of information extraction within SPEED. It will cut processing time significantly and it will reduce disparities in event identification by human coders; all coders will be focusing on the same passages of annotated text.

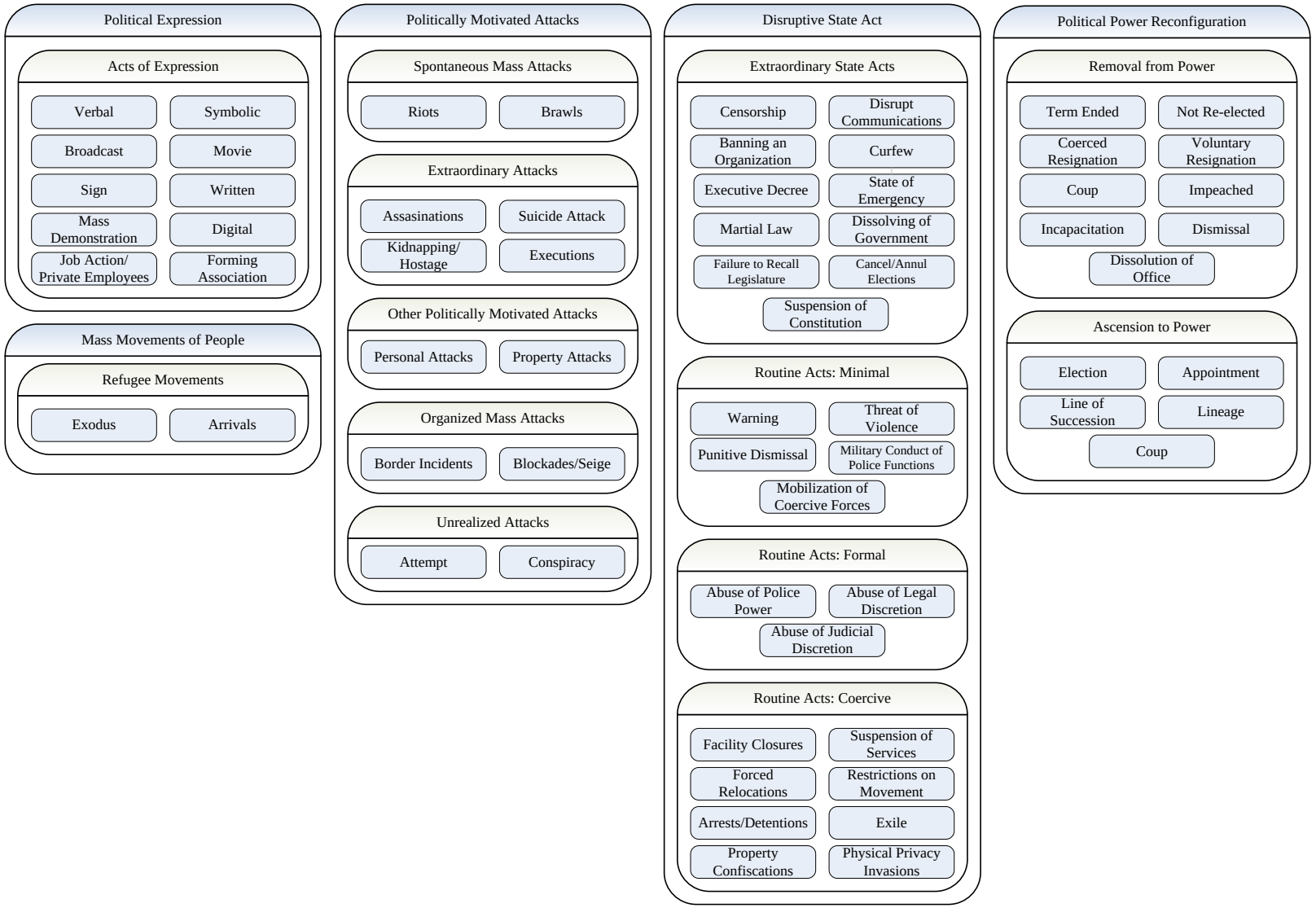
The Societal Stability Protocol and the EXTRACT Suite of Programs

Another distinctive component of SPEED is **EXTRACT**, a suite of programs designed to enhance the capacity of human operators to extract information reliably and efficiently. **EXTRACT** combines a web-based interface that integrates digitized news reports and domain-specific protocols with a direct entry database system. More importantly, **EXTRACT** contains a suite of electronic modules (discussed below) that facilitates information extraction. The SSP can be used to illustrate the role of these modules. It is composed of six substantive sections (who, what, how, where, when and why); each of these sections contains a series of integrated question sets designed to extract event-specific information efficiently and succinctly. The SSP contains 317 questions that human operators complete with the aid of **EXTRACT**; indeed, twenty-three of these questions (news source, publication date, word count, etc.) are automatically completed by **EXTRACT**. Most of the questions in the SSP are relevant only to specific event types or situations; over 93% of the questions are response-activated by 387 branching commands. The types of queries contained in the SSP can be used to illustrate its analytic potential:

[4] A document describing the **BIN** system can be accessed at the following address: http://www.clinecenter.illinois.edu/research/publications/SPEED-BIN.pdf.

[5] It should be noted that one of the costs of obtaining a low false negative rate is the inclusion in the "relevant" bin of a sizeable number of false positives – news reports with no information on relevant events. About 65% of the news reports screened by the **BIN** system contain no information on destabilizing events.

Figure 1
Societal Stability Event Ontology



- **Who**
  - Initiators; Targets/Victims
  - International involvement
- **What**
  - Event type
  - Impacts (people, property, society)
  - Consequences (for initiators)
  - Reactions (to event)
  - Subsequent events
- **How**
  - Weapon, mode of expression
- **Where**
  - Geo-spatial location, geo-physical setting
- **When**
  - Date
- **Why**
  - Societal context
  - Attributed origins

With respect to "who" is involved in the event, the SSP captures information on initiators, targets and victims. List sets were developed that capture thirty-seven types of non-governmental actors (socio-cultural groups, workers, civic leaders, clergy, etc.) and twenty-three types of government actors (public safety officers, soldiers, bureaucrats, presidents, dictators, generals, etc.). In addition, **EXTRACT**'s **PROPER NAME MODULE** uses various NLP-based techniques to capture the proper names of individuals, groups and countries efficiently and accurately. The module scans the news report and returns a list of names; operators merely select the appropriate name to record it automatically. With respect to "what" the event entailed and "how" it unfolded, the protocol extracts information on both the multi-tier event type and its scope/intensity. A series of scope/intensity question sets captures such information as the number of initiators and victims, the mode of expression (list sets containing over thirty options), the weapon used (a twenty-five item list set), and the event's effects (e.g., impact on individuals/communities/society, property damage, etc.). The SSP also has question sets that capture the direct consequences for initiators, post-hoc reactions (condemnations, boycotts, retaliatory attacks, protests, etc.) promulgated by entities not involved in the event (foreign governments, civic groups, international organizations, etc.). To enhance the utility of these data, as well as other SSP event data, **EXTRACT**'s **LINK** module creates electronic links between a focal event and related events. This transforms SPEED's event data from a set of unconnected events to an integrated database that can be used to conduct such things as network analysis and intervention analyses that gauge deterrent effects.

With respect to "where," geographic information is provided by **EXTRACT**'s **GEOCODER** module, which can generate geospatial data for events down to the city level. The **GEOCODER** scans the news report and uses NLP principles to identify geographic references. It then matches those references to a database of over 8 million place names constructed from the GEOnet Names Server (GNS) and the Geographic Names Information System (GNIS) and displays a list of identified place names to the operator. Operators simply select the appropriate place name to save all of the geospatial data in these databases as part of the event record. Finally, with respect to "why" an event occurred, the SSP captures an extensive amount of information on the societal context of the event and its attributed

origins. The contextual information in the SSP is captured within five broad categories (period of sustained violence; transitory governance situation; in the midst of a social movement; some other sort of ongoing societal turmoil; in the penumbra or anniversary of a prominent event); each has many embedded categories that are response-activated. More than one context can be identified. The event origins captured within the SSP are organized within nine broad categories, each of which has many embedded sub-categories. The nine tier-one categories are: anti-government sentiments, socio-cultural group animosities, economic or class-based concerns, desire for self-determination/ political rights, desire for political power, imminent threat to public order, imminent threat to personal security, desire for retribution, and eco-scarcities.

Training and Testing Procedures

Because SPEED uses human operators working within a relatively complex protocol, it is crucial to ensure that all operators are highly trained and that they operate at an acceptable level of proficiency throughout their coding tenure. To achieve these objectives a detailed training regimen has been developed that contains lectures, one-on-one training sessions, and group training sessions to familiarize each operator with the SSP. This training culminates in a series of tests that gauge their ability to implement the protocol. The first test covers general concepts covered by the SSP and a set of training documents; the second test gauges their capacity to identify events embedded in news reports. The last test is the "gatekeeper" test; it examines the operator's ability to extract key items of information accurately. Trainees must pass the gatekeeper test before they are allowed to generate "production data." After beginning production coding operators are regularly tested to insure that they are performing their processing duties in an accurate and reliable manner; to conduct these test we use **EXTRACT**'s capacity to blindly feed a set of pre-coded "test" articles to all operators.[6]

Summary Comparison with Other Academic Projects Collecting Civil Unrest Event Data

SPEED is one of several academic research projects that use event data to enhance our understanding of intra-state conflict/civil unrest. These include the Penn State Event Data Project (PSEDS, formerly the Kansas Event Data System, or KEDS); [7] a collaborative project between the Peace Research Institute Oslo (PRIO) and Uppsala University's Uppsala Conflict Data Program (UCDP), which includes the Armed Conflict Location and Events Dataset project (ACLED); the University of Maryland's Global Terrorism Database (GTD); and the Minorities at Risk project (MAR). Each of these projects has made major contributions to this field of study. But they differ from one another, and SPEED's SSP, in a number of important regards. The most important include their: (1) focus and objectives; (2) temporal and spatial reach; (3) information sources; (4) use of technology; (5) unit of analysis; (6) data collection practices; and (7) quality control procedures. Some of these event-based projects are concerned with specific types of events (civil war deaths, events involving minority groups, terrorist attacks, etc.) and do not aspire to compile a comprehensive body of data on civil unrest. Some are global projects that span the entire post WWII era; others focus on certain regions of

[6] A fuller discussion of SPEED's training and testing procedures, including the results of the reliability tests can be found at: http://www.clinecenter.illinois.edu/research/publications/SPEED-Reliability.pdf.

[7] With the introduction of CAMEO, PSEDS coding scheme is very similar to another event data project that uses machine coding, IDEA (Integrated Data for Event Analysis; Bond, et al., 2003). The focus here is on KEDS/PSEDS because it is an academic project, but most of the observations made here would hold true for IDEA.

the world and/or cover a shorter time span. Some use targeted information sources while others are more eclectic in their sources. There is also a good deal of variation in their use of technology, as well as their unit of analysis (the individual event, the country-year, the group country-year, etc.), the depth of event-specific information collected and the format used to report that data (data base, narrative summaries, mixed). Finally, these projects differ in their efforts to assure quality control in collecting event-based data (coder training, reliability tests, etc.).

The differences across these projects are reported in Table 1, which summarizes a more detailed comparison.[8] Table 1 shows that, of the projects with a global scope (SSP, GTD, UCDP/PRIO), the SSP has the most encompassing focus (i.e., a wide range of small-bore civil unrest events); KEDS/PSEDS also has a broad focus but it presently focuses on a handful of regions. While most of the projects draw from an extensive set of news sources, only the SSP and UCDP/PRIO cover the entire post-WWII era; the other projects begin at 1970 or after. The SSP employs a variety of advanced tools, but the only fully automated project is KEDS/PSEDS. Only the SSP, KEDS/PSEDS, ACLED and the GTD employ the event as their unit of analysis and only the SSP, ACLED and the GTD can provide geographic information to the city-day level. The SSP, ACLED and the GTD provide the most extensive amount of event-specific data, but much of the detailed information in ACLED is reported in narrative form in the "Notes" field – rather than recorded as quantitative variables in a database format. The SSP and KEDS/PSEDS are the only projects that provide reliability statistics on their data collection efforts.

Expanding the Scope of Research on Civil Unrest: The Need for Small-bore Data

Most of the projects discussed above, as well as others, have been initiated within the last decade. The reason for the recent surge of scholarly interest on civil strife is well articulated by Kahl (Kahl 2006). He summarizes the views of many observers when he notes that

> Civil strife in the developing world represents perhaps the greatest international security challenge of the early twenty-first century. Three-quarters of all wars since 1945 have been within countries rather than between them…Wars and other violent conflicts have killed some 40 million people since 1945, and as many people have died as a result of civil strife since 1980 as were killed in the First World War (Kahl 2006:1)

In operationalizing this concern with civil strife almost all quantitative cross-national studies have focused on civil wars. Moreover, a number of highly regarded research projects have generated data on the existence of a civil war, with the country-year as the unit of analysis. The most prominent of these data series are the UCDP/PRIO conflict data discussed above (Gleditsch et al. 2002; www.prio.no); a new dataset from the Correlates of War Project (www.correlatesofwar.org), and a project directed by James Fearon and David Laitin (Fearon and Laitin 2003; www.stanford.edu/group/ethnic/publicdata). There is ample justification for the prevailing focus on civil wars: they are the most destabilizing and violent manifestation of civil unrest and they account for a disproportionate amount of the fatalities rooted in intra-state concerns. Moreover, their saliency should make it possible to identify them objectively, which can provide the basis for the cumulative development of knowledge in this area. Intellectual progress is often fostered when independent researchers with diverse perspectives and approaches can work from a common and empirically well-grounded information base.

[8] A detailed comparison of these projects can be found at: http://www.clinecenter.illinois.edu/research/publications/SPEED Comparison-With-Other-Projects.pdf.

Table 1
Comparison of Civil Unrest Event Data Projects

Project	Focus	Objectives	Temporal Reach	Spatial Reach	Information Sources	Unit of Analysis	Use of Technology	Scope of Data Collection	Coder Training Procedures	Reliability Checks
SPEED's SSP	Civil unrest	Identify and analyze patterns in civil unrest over time and space	1946-Present	Global	Extensive, multiple sources	Event	Identifying relevant news reports and textual passages; automation of date, location and socio-cultural group information; linking related events	Extensive information on large set of event types, geographic location, date, actors, consequences, reactions, contexts, and origins.	Multiple weeks of intensive training with selected news reports; culminates with several tests of coding proficiency	Two different reliability checks during training; regular, blind reliability checks during coding
PSEDS	Interstate interactions and intrastate conflict	Develop early warning indicators for political change and civil unrest	1979-Present	Selected countries in the Balkans, Middle East, West Africa	Extensive; largely wire service reports (mainly Reuters)	Event	Full automation of event coding using the TABARI program; normally limited to lead sentence	Event counts for events included in CAMEO coding scheme; heavily focused on international interactions. Includes actor and country information.	N/A	Early tests of KEDS system demonstrated 75-85% agreement with human coders; tests mainly on lead sentence parsing
UCDP/PRIO	Intrastate conflict involving government and rebel/insurgent groups	Identify and analyze patterns in intrastate armed conflict over time and space	1946-2008	Global	Varies by country; primarily Keesing's and various regional reports	Conflict-year	Minimal	Deaths counts by year for intrastate conflicts involving governments and insurgents that result in at least 25 battlefield deaths	Unspecified	Unspecified
ACLED	Battle events that are part of civil war	Identify and analyze information on discrete events involving intrastate conflicts over time and space	1997-2010	50 developing countries in Africa, Central Europe, Asia	Extensive; varies by country; incl. Keesing's, various regional reports, and international news sources (Reuters, AFP, etc.)	Event	Minimal	Includes information on event type, actor, date, geographic location; no minimum threshold for inclusion	Coders provided with summary materials about each conflict prior to coding	Codings reviewed by senior staff
GTD	Acts of terrorism or activities by terrorist organizations	Track and analyze terrorist activity over time and space	1970-2008	Global	Pinkerton, publicly available news sources	Event	Minimal	Any incident meeting 2 out of 3 criteria of terrorism; extensive information on location, date, event type, actors, consequences.	Unspecified	Codings undergo a systematic review process by GTD staff
MAR	Politically active minority groups that are or have been at risk within a county	Provide systematic comparative information relating to conflict involving politically active minority groups	Group data is for 2004-2006	283 ethnic groups in 117 countries	Unspecified open source information	Minority group-year	Minimal	Includes all groups meeting MAR criteria for "at-risk." Extensive information on group characteristics, current status, repression, etc.	Coders undergo a rigorous training procedure	Coding is reviewed by senior staff prior to release

Despite the prevalence of the current focus on civil wars, this focus is unlikely – by itself – to provide the basis for intellectual progress in a field that includes some of the most important threats to human well-being in the contemporary era. The deficiencies of a civil war focus stem from its inability to generate type of broad and empirically well-grounded knowledge base that is needed for intellectual progress in the study of civil strife. These deficiencies are rooted in three factors. First, civil wars are challenging to delineate and scholars have not reached a consensus on how to identify them. As a result, due to the use of different criteria and information sources, a great deal of disparity exists across efforts to identify them.[9] Second, capturing civil wars as a dummy variable measured at the country-year level masks an enormous amount of variation in conflict, which compromises its utility as an analytic focus. Third, civil wars, because they are an extreme form of civil discontent involving only a small subset of actors, account for only a small proportion of civil unrest. This last point suggests that the deficiencies with a civil war focus cannot be addressed simply by improvements in efforts to gauge civil wars more precisely. Rather, a much more encompassing focus on civil strife is needed, one that captures a broader range of small-bore destabilizing actions initiated by a wider set of domestic actors.

A comparison of the data on civil wars from the three projects discussed above for the period from 1946 to 1999 can be used to illustrate the first point. During this period one or more of these projects identifies 1272 country-years of civil war. Unfortunately, the level of agreement across them is quite low, as seen in Figure 2. All three sources agree on only 357 of these years, 28% of the total. Dyadic comparisons are somewhat better. For example COW and Fearon and Laitin agree on 49% on 968 country-years of war; COW and UCDP/PRIO agree on 35% of 1105 country-years. Fearon and Laitin and UCDP/PRIO agree on 51% of 1246 country-years. Table 2 presents some insights into the nature of the disagreements for five prominent countries (Angola, China, Guatemala, India, and Iraq) that are fairly representative of the level of agreement across the three data sets (i.e., the different sources agree on about one-third of the country-years). With respect to Angola, in twenty-five consecutive years (1975-1999), either the UCDP/PRIO, COW, or Fearon-Laitin data indicate internal conflict. However, the three sources of data agree as to the presence of internal conflict in just six years (24%). China has twenty-one years of internal conflict but there is agreement between the conflict data sets in only one-third of the cases. Similar patterns of agreement are found in the data for the remaining countries in Table 2 (Guatemala 29%, India 25%, and Iraq 24%).

Even if scholars were able to reach a consensus on the identification of civil wars, a solitary focus on a dummy civil war variable measured at the country-year level would still be problematic: such an operationalization masks a substantial amount of variation in conflict. This point can be illustrated using a set of data generated by SPEED's SSP. These data were derived from a "saturation study" (i.e.,

[9] The UCDP/PRIO project employs the most liberal criteria for identifying intra-state conflict which it defines as "a contested incompatibility that concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in 25 battle-related deaths." Fearon and Laitin have slightly more restrictive criteria for internal conflict. To qualify as a violent civil conflict must be: 1) involve fighting between agents of a state and organized non-state groups, 2) there must be at least 1000 deaths over the course of the conflict and a yearly average of at least 100 killed, and 3) both sides of the conflict must suffer at least 100 deaths. The COW data has the most stringent criteria. In order to qualify as an intra-state conflict it must involve "…sustained combat, involving organized armed forces, resulting in a minimum of 1000 battle-related combatant fatalities within a twelve month period."

Figure 2
Summary of Comparisons across Civil War Projects

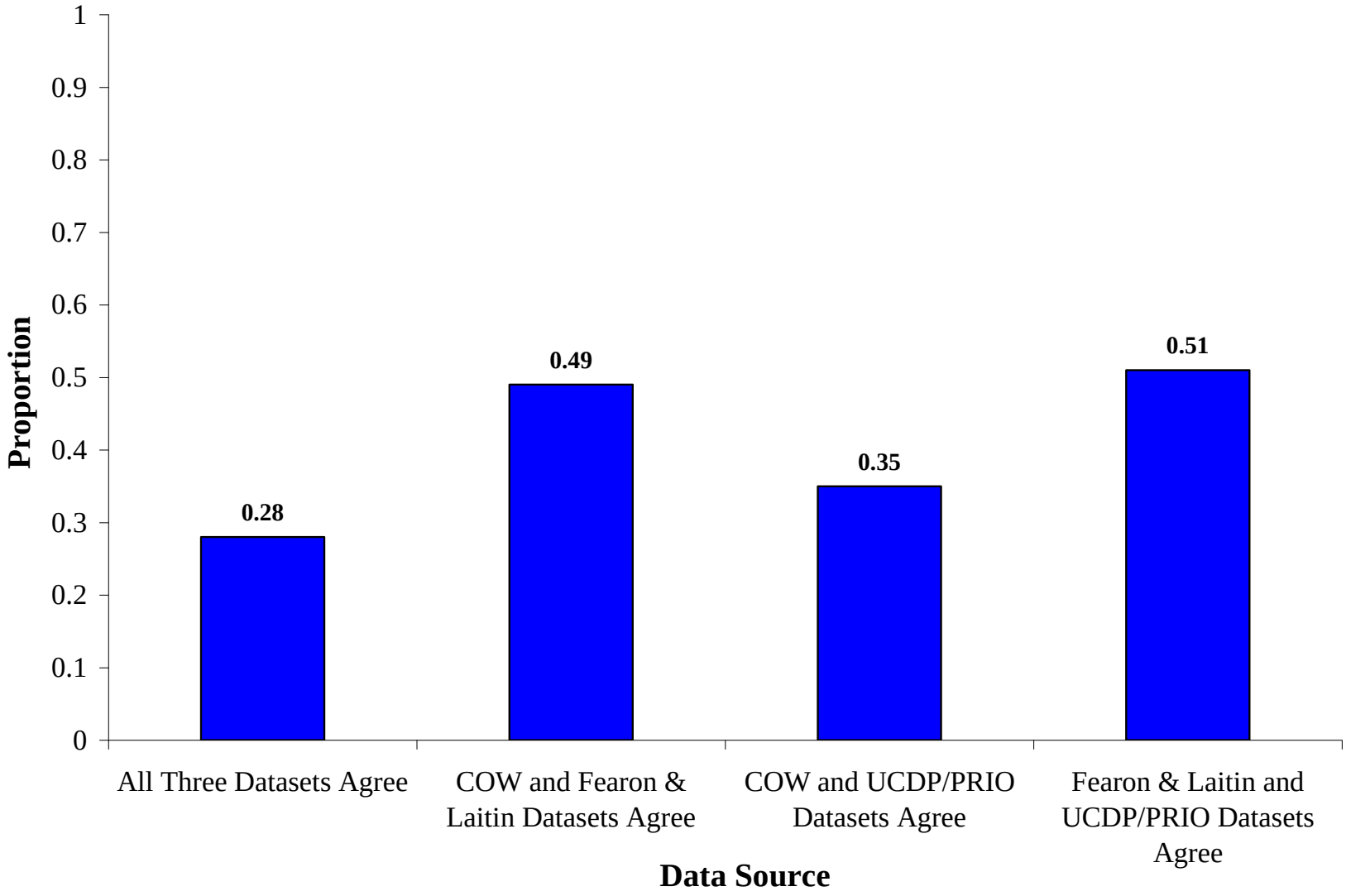


Table 2
Comparison of Civil War Data Sets for Five Countries

COW	Fearon-Laitin	UCDP/PRIO	Agree	COW	Fearon-Laitin	UCDP/PRIO	Agree	COW	Fearon-Laitin	UCDP/PRIO	Agree	COW	Fearon-Laitin	UCDP/PRIO	Agree	COW	Fearon-Laitin	UCDP/PRIO	Agree
1946	.	.	.	.	1	1	1	Yes	0	0	0	.	0	0	1	No	0	0	0	.
1947	.	.	.	.	1	1	1	Yes	0	0	0	.	0	0	1	No	0	0	0	.
1948	.	.	.	.	1	1	1	Yes	0	0	0	.	0	0	1	No	0	0	0	.
1949	.	.	.	.	1	1	1	Yes	0	0	1	No	0	0	1	No	0	0	0	.
1950	.	.	.	.	1	1	1	Yes	0	0	0	.	0	0	1	No	0	0	0	.
1951	.	.	.	.	0	1	0	No	0	0	0	.	0	0	1	No	0	0	0	.
1952	.	.	.	.	0	0	0	.	0	0	0	.	0	1	0	No	0	0	0	.
1953	.	.	.	.	0	0	0	Yes	0	0	0	.	0	1	0	No	0	0	0	.
1954	.	.	.	.	0	0	0	Yes	0	0	1	No	0	1	0	No	0	0	0	.
1955	.	.	.	.	0	0	0	Yes	0	0	0	.	0	1	0	No	0	0	0	.
1956	.	.	.	.	1	1	1	Yes	0	0	0	.	0	1	1	No	0	0	0	.
1957	.	.	.	.	1	1	0	No	0	0	0	.	0	1	1	No	0	0	0	.
1958	.	.	.	.	1	1	0	No	0	0	0	.	0	1	1	No	0	0	1	No
1959	.	.	.	.	1	1	1	Yes	0	0	0	.	0	1	1	No	1	1	1	Yes
1960	.	.	.	.	0	0	0	.	0	0	0	.	0	1	0	No	0	0	0	Yes
1961	.	.	.	.	0	0	0	.	0	0	0	.	0	1	1	No	1	1	1	Yes
1962	.	.	.	.	0	0	0	.	0	0	0	.	0	1	1	No	1	1	1	Yes
1963	.	.	.	.	0	0	0	.	0	0	0	.	0	1	1	No	1	1	1	Yes
1964	.	.	.	.	0	0	0	.	0	0	0	.	0	1	1	No	0	1	1	No
1965	.	.	.	.	0	0	0	.	0	0	1	No	0	1	1	No	1	1	1	Yes
1966	.	.	.	.	0	0	0	.	1	0	1	No	0	1	1	No	1	1	1	Yes
1967	.	.	.	.	1	0	0	No	1	0	1	No	0	1	1	No	0	1	1	No
1968	.	.	.	.	1	0	0	No	1	1	1	Yes	0	1	1	No	0	1	1	No
1969	.	.	.	.	0	0	0	.	0	1	1	No	0	1	1	No	1	1	1	Yes
1970	.	.	.	.	0	0	0	.	1	1	1	Yes	1	1	1	Yes	1	1	1	Yes
1971	.	.	.	.	0	0	0	.	1	1	1	Yes	1	1	1	Yes	0	1	0	No
1972	.	.	.	.	0	0	0	.	0	1	1	No	0	1	0	No	0	1	0	No
1973	.	.	.	.	0	0	0	.	0	1	1	No	0	1	0	No	0	1	1	No
1974	.	.	.	.	0	0	0	.	0	1	1	No	0	1	0	No	1	1	1	Yes
1975	0	1	0	No	0	0	0	.	0	1	1	No	0	1	0	No	1	0	1	No
1976	1	1	0	No	0	0	0	.	0	1	1	No	0	1	0	No	0	0	1	No
1977	1	1	0	No	0	0	0	.	0	1	1	No	0	1	0	No	0	0	1	No
1978	1	1	0	No	0	0	0	.	1	1	1	Yes	0	1	1	No	0	0	1	No
1979	1	1	0	No	0	0	0	.	1	1	1	Yes	0	1	1	No	0	0	1	No
1980	1	1	0	No	0	0	0	.	1	1	1	Yes	0	1	1	No	0	0	1	No
1981	1	1	0	No	0	0	0	.	1	1	1	Yes	0	1	1	No	0	0	1	No
1982	1	1	0	No	0	0	0	.	1	1	1	Yes	0	1	1	No	0	0	1	No
1983	1	1	0	No	0	0	0	.	1	1	1	Yes	0	1	1	No	0	0	1	No
1984	1	1	0	No	0	0	0	.	1	1	1	Yes	1	1	1	Yes	0	0	1	No
1985	1	1	0	No	0	0	0	.	0	1	1	No	0	1	1	No	1	0	1	No
1986	1	1	0	No	0	0	0	.	0	1	1	No	0	1	1	No	1	0	1	No
1987	1	1	0	No	0	0	0	.	0	1	1	No	0	1	1	No	1	0	1	No
1988	1	1	0	No	0	0	0	.	0	1	1	No	0	1	1	No	1	0	1	No
1989	1	1	0	No	0	0	0	.	0	1	1	No	0	1	1	No	0	0	1	No
1990	1	1	1	Yes	0	0	0	.	0	1	1	No	1	1	1	Yes	0	0	1	No
1991	1	1	1	Yes	0	1	0	No	0	1	1	No	1	1	1	Yes	1	0	1	No
1992	1	1	1	Yes	0	1	0	No	0	1	1	No	1	1	1	Yes	0	0	1	No
1993	1	1	1	Yes	0	1	0	No	0	1	1	No	1	1	1	Yes	0	0	1	No
1994	1	1	1	Yes	0	1	0	No	0	1	1	No	1	1	1	Yes	0	0	1	No
1995	0	1	1	No	0	1	0	No	0	1	1	No	1	1	1	Yes	0	0	1	No
1996	0	1	1	No	0	1	0	No	0	1	0	No	1	1	1	Yes	1	0	1	No
1997	0	1	1	No	0	1	0	No	0	0	0	.	1	1	1	Yes	0	0	0	.
1998	1	1	1	Yes	0	1	0	No	0	0	0	.	1	1	1	Yes	0	0	0	.
1999	1	1	0	No	0	1	0	No	0	0	0	.	1	1	1	Yes	0	0	0	.

all civil unrest events identified in SPEED's SWB archive were coded) of four countries that experienced a civil war in the post-1979 era: El Salvador, Nicaragua, the Philippines and Sierra Leone.[10] Figure 3 (a-d) aggregates, by month, the number of killings of insurgents or soldiers for these countries; at the top of each graph is the period defined as a civil war by each of the three projects discussed above. While it is not (and cannot) be asserted that every death involving a soldier or an insurgent is captured in these saturation codings, it is also unlikely that there are systematic gaps in coverage. Thus, Figure 3 suggests that an enormous amount of monthly variation exists in the intensity of conflict. There are clearly periods of intense strife and periods of relative calm. Indeed, on average, there were no reported casualties in about 60% of the months in which all three projects agreed that a civil war was on-going. It is also clear that most violence stops well before the war is acknowledged to have ended. A geospatial analysis would reveal a great deal of variance across space as well.

Reorienting the study of civil wars to focus on disaggregated death totals would enhance the analytic utility of a civil war focus. But this reorientation would not change the fact that a civil war focus captures only a small slice of civil unrest by a relatively narrow set of actors (soldiers and insurgents). This can be illustrated by examining a random sample of over 40,000 SSP events that were derived largely from New York Times historical archive (1946-2005).[11] While politically motivated attacks constitute about 55% of all civil unrest events (and all civil war-related events), political expression events (speeches, symbolic acts, demonstrations, strikes, etc.) constitute 20% of these events; disruptive state acts (excluding violent attacks against insurgents) constitute about 17%. Irregular transfers of power account for another 6% of the destabilizing events and mass movements of people account for about 2%. However, only 10.5% of the politically motivated attacks identified involved soldiers and insurgents. These attacks account for just 5.7% of all destabilizing events. Moreover, if we consider only politically motivated attacks involving soldiers and insurgents in which someone is killed (the core piece of data used to identify civil wars), the percentage drops to 5.8% of all attacks and 3.2% of all destabilizing events. The unrest that unfolded in Northern Africa and the Middle East during the latter part of 2010 and the early part of 2011 underscores the import of the larger point here. Virtually none of the unrest that toppled governments would have been captured with a civil war focus.

SPEED's SSP and the Study of Civil Unrest

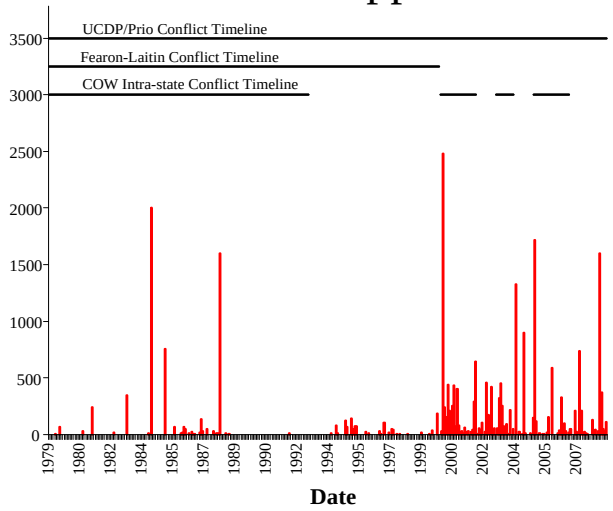
The above assessment of a civil war focus in the study of civil strife underscores the need for a broader analytic focus. This section examines the capacity of SPEED's SSP to contribute to the creation of a broader, yet empirically well-grounded, knowledge base in this field. As noted in the discussion surrounding Figure 1 and Table 1, the SSP has a number of distinctive features that could yield both useful and unique insights to the study of civil unrest. But, as with most things these features involve trade-offs. The first two subsections examines some of these trade-offs and the third subsection illustrates the potential value of using SSP data to enrich the study of civil strife.

[10] These data were collected and analyzed as part of a collaborative agreement, entitled "Natural Resources, Climate Change and Societal Stability: the Role of Human Interventions," with the U.S. Army Construction Engineering Research Laboratory; Great Rivers Cooperative Ecosystems Studies Unit Agreement No. W9132T-10-2-0014.

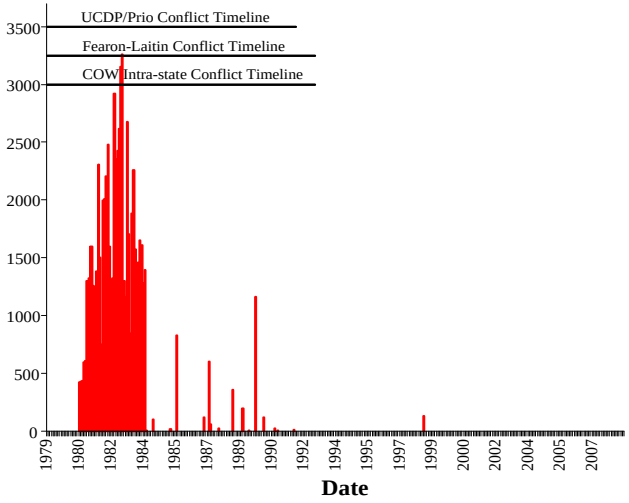
[11] This analysis excludes US events because the New York Times has disproportionate coverage of US events and it would skew the distributions of destabilizing events.

Figure 3
Monthly Death Totals for Soldier and Insurgents

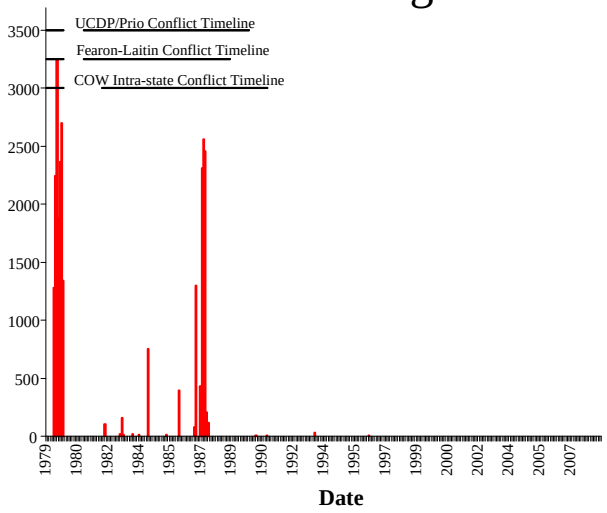
a. Philippines



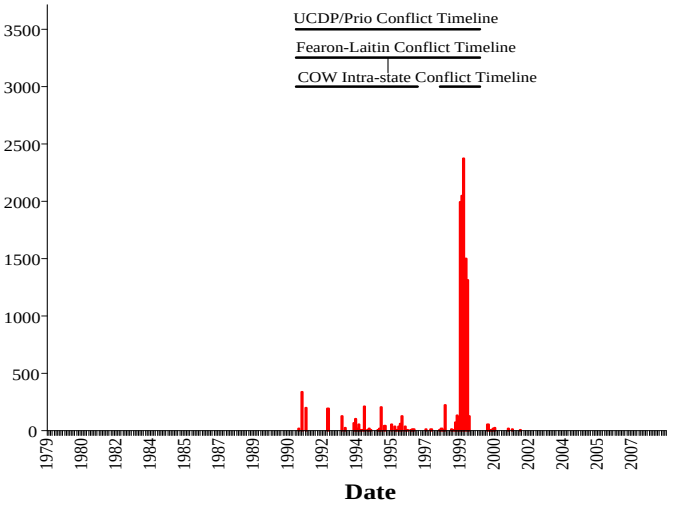
c. El Salvador



b. Nicaragua



d. Sierra Leone



The SSP Event Ontology and Information Base

As noted above, two of the most distinctive features of SPEED's SSP are the comprehensiveness of its event ontology and the richness of the information it extracts from news reports. However, while these features lead to the generation of a great deal of data on an array of civil unrest events, they are also the features that impede more highly automated information extraction. Thus, a legitimate inquiry is whether the cost of these features outweighs the benefits of a system that focuses on event counts, which can more readily be automated.

The SSP's event ontology was developed in a year-long pretest and refined at various points in its development. While no ontology can capture the full range of destabilizing events that have unfolded across scores of nations during the post WWII era, the schema depicted in Figure 1 does reasonably well. Coders have the option of categorizing destabilizing events in an "Other" category, yet they exercise this option in less than 1% of the event codings. The SSP's event ontology's capacity to provide a sense of the relative frequency of soldier/insurgent clashes illustrates its utility. But its real potential will be realized only when it is reduced into a more parsimonious operationalization of destabilizing actions and integrated with data on relevant event characteristics (# of demonstrators, # killed, type of weapons involved, type of participants, etc.). The integration of event types and event characteristics can provide the basis for a more refined analysis of the dynamics of episodes of destabilizing events as well as a more meaningful comparison of civil unrest across time and space. Indeed, the costs of the SSP's emphasis on compiling extensive amounts of event-specific data can only be justified by the contributions of this integration to more refined and sophisticated analyses of civil strife.

The proposition that analytic power can be gained by integrating data on event characteristics with a more parsimonious event ontology is premised on the existence of meaningful, event-specific differences within similar event types. If event characteristics do not vary much across similar types of events, then there is little value-added from collecting extensive amounts of event-specific data. To examine this premise Table 3 and 4 uses the global random sample of events derived from the NYT archive to present some SSP data on differences in four broad types of destabilizing events initiated by non-state actors. Table 3 provides event-specific data on two broad types of political expression events: small-scale events (provocative speeches/postings/ writings, symbolic actions, broadcasts, etc.) and mass political expressions (demonstrations, marches, strikes). Table 4 provides data on two types of politically motivated attacks: non-lethal and lethal.

Table 3 demonstrates that there is a great deal of event-specific variation within these broad categories of political expression. One of the most salient distinctions is the number of participants. By definition, small-scale events involve fewer participants. But even within this category there is a good deal of variation. While the mode and median is '1,' the (highly skewed) mean is 40; 90% of the events involve fewer than 600 participants. Probably a more important distinguishing feature within small-scale expression events is the type of expression involved. Just over one-third involved verbal expressions, about a quarter involved written expressions, and forty percent involved some type of symbolic action (sit-in, self-immolation, picket, etc.). **EXTRACT**'s **LINK** module reveals that nearly one-third of these events were an integral part of a more complex sequence of actions; 13% involved some type of post hoc reaction by either the government or other private actors.

Table 3
Differences Across Expression Events

	Small-scale Expression Events	Mass Expression Events
Number of Participants/Initiators		
Mean	40	77300
Median	1	2500
Mode (proportion)	1 (.52)	2500 (.14)
90th Percentile	600	95000
Type of Expression		
Verbal	0.36	.
Written	0.24	.
Symbolic Action	0.4	.
Demonstration	.	0.71
Strike	.	0.29
Intensity Indicators		
Involved Linked Event?	0.32	0.4
Elicited a Post hoc Reaction?	0.13	0.09
Involved a Weapon?	0.04	0.02

Table 4
Differences Across Political Attacks

	Non-lethal Attacks	Lethal Attacks
Number of Initiators		
Mean	408	89
Median	3	2
Mode (proportion)	2 (32%)	2 (.48)
90th Percentile	200	29
Type of Attack		
Attack Against Person	0.45	.
Attack Against Property	0.25	0.03
Intensity Indicators		
Involved a Linked Event?	0.42	0.37
Elicted a Posthoc Reaction?	0.06	0.03
Weapon Type		
No Weapon	0.39	0.31
Crude Weapon	0.11	0.03
Small Arms	0.16	0.33
Explosives	0.29	0.29
Miltary Grade Weapons	0.05	0.04
Personal Injury?	0.24	1
Number of People Killed		
Mean	.	255
Median	.	2
Mode (proportion)	.	1 (.41)
90th Percentile	.	46

The second column in Table 3 demonstrates the significant variation across mass events. While the median and the mean for these events are 2500, the mean is 77,300; 10% of these events had more than 95,000 participants. Forty percent of these mass events were an integral part of a more complex sequence of actions; 9% involved some type of post hoc reaction by either the government or other private actors. More than 70% of these events were demonstrations or marches; the rest were strikes. Weapons were involved in only a small fraction of any type of political expression events and injuries were even rarer. While not shown in Table 3, the SSP data reveal that nine categories of non-state

actors accounted for at least 5% of the small scale political expression events: members of a political group (10.6%), non-descript citizens (10.1%), members of an insurgent group (8.4%), journalists (8%), students (8%) members of a socio-cultural group (7.6%), clergy (6.3%), political dissidents (5.6%) and workers (5.2%). An examination of the categories of private individuals accounted for at least 5% of the mass political expression events revealed that the same categories were involved – and with the same rank order – with one minor exception: workers just missed the 5% cut-off.

Table 4 presents some data on politically motivated attacks. Most attacks involve just a handful of initiators (2-3), with non-lethal attacks involving a marginally larger number of initiators. Around 40% of both lethal and non-lethal attacks are part of a more complex sequence of events; however, only a handful (3-6%) involved some type of post hoc reaction. Fewer than half of the non-lethal attacks targeted individuals; by definition all lethal attacks did so. Unsurprisingly, lethal attacks involved more sophisticated weapons than non-lethal attacks. At the same time, however, there is a great deal of variance across weapon types within each category. Non-lethal attacks involved no weapon or a crude weapon in 50% of all events, small arms were involved in 16% of the events and explosives were involved in 29%. Lethal attacks involved no weapon or a crude weapon in 34% of all events, small arms were involved in 33% of the events and explosives were involved in 29%. Military grade weapons were relatively rare in both categories (4-5%). Non-lethal attacks involved personal injuries in about one-quarter of the events. Much more variance exists in the number killed in lethal attacks. While the mode is '1' and the median is '2,' the highly skewed mean is 255. Ten percent of these events involved more than 46 people killed. While not reported in Table 4 the SSP data reveal that four categories of non-state actors accounted for at least 5% of the non-lethal political attacks: members of an insurgent group (42.8%), members of a socio-cultural group (13.9%), non-descript citizens (8.2%), and members of a political group (7.8%). Only two categories of non-state actors accounted for at least 5% of the lethal political attacks: members of an insurgent group (53.3%) and members of a socio-cultural group (14.3%). These are not mutually exclusive categories; thus, some actors could be both members of an insurgent group and a socio-cultural group.

The data presented in Tables 3 and 4 demonstrate that similar types of events differ in important ways. This, of course, suggests that SSP data can provide the basis for conducting more refined and sophisticated analyses of civil strife, analyses that make use of important event-specific characteristics. But before the SSP's potential contributions to the study of civil strife can be assessed, the implications of another distinguishing feature must be examined: its use of a global news archive. As noted in the discussion surrounding Table 1 most event data projects concerned with civil unrest, especially those that collect a good deal of event-specific data, focus on only certain regions of the world. In some cases this allows them to utilize a wider range of news sources, including local sources. In contrast, the SSP draws from SPEED's global news archive. The historical component of that news archive, while huge, includes only a handful of news sources, even though the FBIS and SWB components compile reports from a wide range of local sources. This notwithstanding, it is important to examine the implications of this for the SSP's relative capacity to capture reports of destabilizing events.

SPEED's Global News Archive

No news source, or set of news sources, will capture all of the important events unfolding in the world or even in selected regions of the world. Limitations in media coverage and media bias are issues that all event data projects must deal with if they use news reports as their source of information. Coverage limitations and news bias are also challenging issues to address in a rigorous and empirically based

manner. To address these issues properly would require a major and multi-faceted long-term effort. However, some limited insights into the relative capacity of the SPEED project's historical news archive to identify civil unrest events can be gained by a comparison of the saturation codings (see note 10 and surrounding discussion) of two African countries, Liberia and Sierra Leone, with data collected by ACLED. ACLED focuses on African countries and draws on an encompassing set of news sources. While ACLED provides largely event counts for a smaller set of event categories than the SSP it is possible to collapse some of their categories in ways that make it possible to compare their event counts with SSP event counts compiled from one of SPEED historical archives that was used to conduct the saturation study (SWB). Two key event categories were constructed for this analysis: civil war attacks and civilian attacks. To construct the civil war attack variable from ACLED data we collapsed three categories (Battle-Government gains territory; Battle-No Change of territory; Battle-Rebels overtake territory). To construct this event category from SSP data we included all political attacks involving government actors and insurgents. To construct the civilian attack variable from ACLED data we used their "Violence against civilians" category. This event category was constructed from SSP data by including any attack on private citizens who were not insurgents.

To conduct this comparison we summed the events in each of these categories, by month, for every year between 1990 and 2008 (ACLED data does not begin before 1990; the saturation study ended in 2008). The results are depicted in Figure 4 (a-d). As is evident, the results are mixed. ACLED captured considerably more civil war attack events than SPEED's SSP in both Liberia (115 vs. 747) and Sierra Leone (1139 vs. 351). However, SPEED captured a comparable number of civilian attacks in both countries: 367 SSP events in Sierra Leone compared to 374 ACLED events; 303 SSP events in Liberia compared to 191 ACLED events. This suggests that drawing from a variety of news sources, including local sources, can have a significant effect on the events identified. It is puzzling, however, that the effect only emerges with respect to civil war attacks. This may reflect differences in event definitions as well as the variety of news sources. Even more puzzling is the fact that the number of SSP event counts for civil war attacks declines drastically after 2000 especially in Liberia. Indeed, while the correlation between SSP and ACLED monthly event counts for civil war attacks in Sierra Leone is .65, it is only .32 in Liberia. Joining SPEED's two other sources of global news (the New York Times and FBIS), will undoubtedly affect the event counts depicted in Figure 4 but it is impossible to determine the magnitude of the effect until the other sources are examined.

To provide more insights into this analysis, the SSP event counts were weighted by the number killed in each of the categories analyzed in Figure 4 (a-d). The result is a gauge of the number of people killed during each month. These data are then displayed in conjunction with ACLED event counts in Figure 5 (a-d), using different scales for event counts and the number killed. Also, the number killed variable is truncated at 100 because a handful of outliers (13 months or 1% of the total) resulted in a scale that obscured the range in variation for the other 99% of the events. What the data in Figure 5 reveal underscores a point made in the context of Table 3 and 4. Because of the difference across events, event counts do not always reflect the intensity of civil unrest. Thus, the correlation between ACLED event counts for civil war attacks and SSP death counts for civil war attacks is only .01 in Liberia and .25 in Sierra Leone.

Figure 4
Comparison of SSP and ACLED Data for Sierra Leone and Liberia

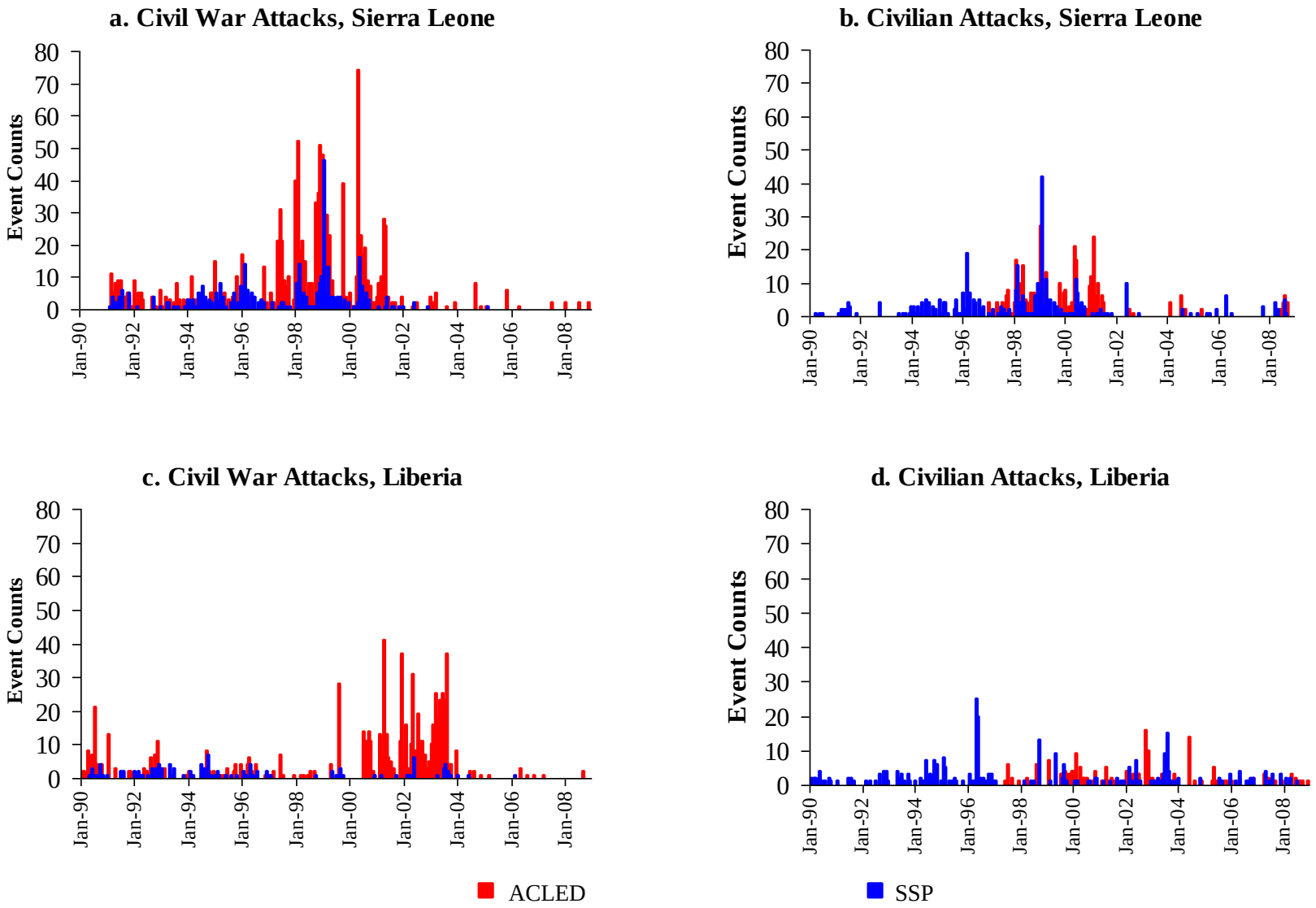
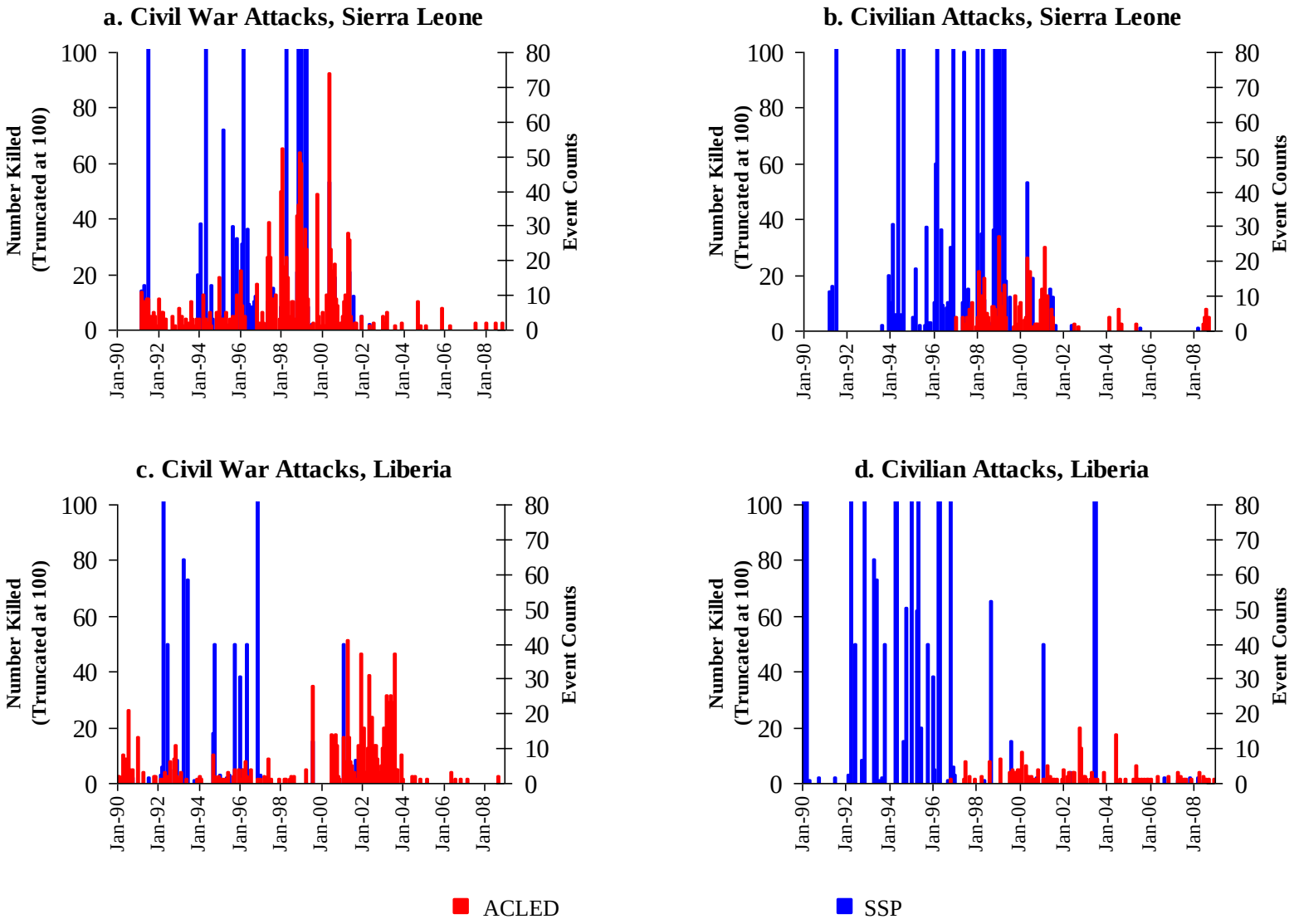


Figure 5
Comparison of SSP and ACLED Data for Sierra Leone and Liberia, weighted by Death Counts



SSP Data and Global Trends in Civil Unrest

To illustrate more concretely the analytic power can be gained by integrating data on event characteristics with a parsimonious event ontology, this section uses the global random sample of NYT events analyzed earlier to map some general trends in civil unrest outside the US. To do this it uses intensity measures of two broad categories of destabilizing events: political protests (demonstrations, strikes, symbolic acts, etc.) and political violence (bombings, attacks, assassinations, kidnappings, etc.).[12] These intensity measures were summed, by year, to capture the relative magnitude of civil unrest. These sums are reported in Figure 6, which demonstrates that the intensity of political violence begins to increase steadily in the early 1950's.[13] The intensity of political protest measure, on the other hand, is relatively stable for the early part of this time frame. This, of course, suggests that political violence was becoming the mode of choice to express domestic discontents. The upward trend in political violence continues unabated until the mid-1980. Then, after a brief respite, the trend in political violence begins to increase again at the turn of the century; in 2005 it approaches its post-war high.

Because the SSP captures information on event origins it is possible to provide more refined insights into these global Post-WWII trends. Analyses of the origins data revealed five principal origins of civil unrest: anti-government sentiments (actors, actions, policies), socio-cultural animosities (religious, ethnic, racial, tribal, national origins), class-based concerns (e.g., communist insurgencies, labor unrest), desire for political liberties (independence movements, political reforms), and concerns over basic human needs (food, water, land).[14] Using the same summation procedure as in Figure 6, Figure 7 depicts the relative importance of these different origins, by year, for the political violence measure. As can be seen in Figure 7, much of the sustained increase in political violence reported in Figure 6 is due to two factors: anti-government sentiments and socio-cultural animosities. Socio-cultural factors begin a marked increase around 1968 and skyrocket for a period after 1980. They drop significantly in the mid-1980's but then begin a steep incline in the early 1990's. In contrast, class-based factors begin to recede in importance by the mid-1970's. The desire for political rights begins to decline in the early 1980s, but evidences a modest surge again in the 1990's.

Conclusion

The importance of studying civil strife is self-evident; in the post-WWII era, intrastate conflict has risen to become a greater threat to life and security than even interstate wars. Over the past several decades, advances in computing technology have made the study of civil strife more feasible than ever. This has allowed the automation of much of the data collection needed to create a knowledge base about civil strife. Moreover, such automation can achieve high degrees of precision and reliability, and

[12] Details on the derivation of the intensity measures can be found on the Cline Center website at: http://www.clinecenter.illinois.edu/research/publications/SPEED-Gauging-Intensity-of-Civil-Unrest.pdf.

[13] These figures depict a moving average of the aggregated intensity scores for every country in the world (excluding the US) for a given year.

[14] Details on the derivation of the event origins categories are provided at: http://www.clinecenter.illinois.edu/research/publications/SPEED-Origins-of-Destabilizing-Events.pdf.

dramatically reduce the labor costs involved with traditional human coding. Making use of these technological developments will allow us to increase both the breadth and depth of civil strife data.

SPEED's approach to compiling event data is unique in a number of ways. SPEED's SSP employs a hybrid approach that joins the automated processing of text with human coding. Automated search procedures, such as our **BIN** system, allow us to quickly and efficiently accumulate a vast database of relevant text from which to work. Achieving comparable results using only human coding would take an inordinate amount of time. Moreover, using NLP-based techniques facilitates our extraction of detailed information such as precise geo-spatial event locations and actor names. Although these automated techniques aid in the extraction of such information, relying on a group of well-trained and regularly tested coders allows SPEED's SSP to collect much more information than would be possible exclusively using automation. As documented above, these features provide important insights into event-based differences across similar categories of events.

This underscores the importance of expanding the focus of research on civil strife. If we are restricted to event counts or similarly rough measures of strife, we cannot tell a dozen citizens petitioning their government from tens of thousands of protestors demanding revolution. Thus, many types of events that would be relevant to civil strife can become meaningless and are often ignored. The event ontology and structure of the SSP allows us to collect meaningful information on a particularly broad range of events relevant to civil strife. The basic event types are grounded in theoretical concerns, but the vast numbers of subcategories have been derived from thousands of pre-test news reports and as such reflect the types of events that are likely to occur in real-life situations. By using a protocol that is rich in details, we can determine empirically which features of different event types are relevant. This frees us from determining *a priori* which pieces of information are most crucial, and instead allows us to focus on developing models of event intensity that capture the scope and severity of civil strife.

In sum, data from SPEED's SSP can contribute to our knowledge and understanding of civil strife because of several innovations in our methods of data collection. Our hybrid approach to the use of information technology and automation allow us to compile a vast store of raw data from which to extract event information. This hybrid approach also improves the efficiency and accuracy of our human coders by automating difficult tasks including precisely locating events and identifying actors involved. Our use of human coding facilitates the collection of detailed information on a much broader range of events. As a result, we can begin to expand the focus of research in civil strife from large-bore events such as civil wars to small-bore events such as protests and low-level political violence. In so doing, we may begin to unravel the complicated relationships between lower levels of civil strife and the escalation of violent, destabilizing unrest manifested by civil wars and revolutions.

Figure 6
Indices of Violence and Protest Intensity

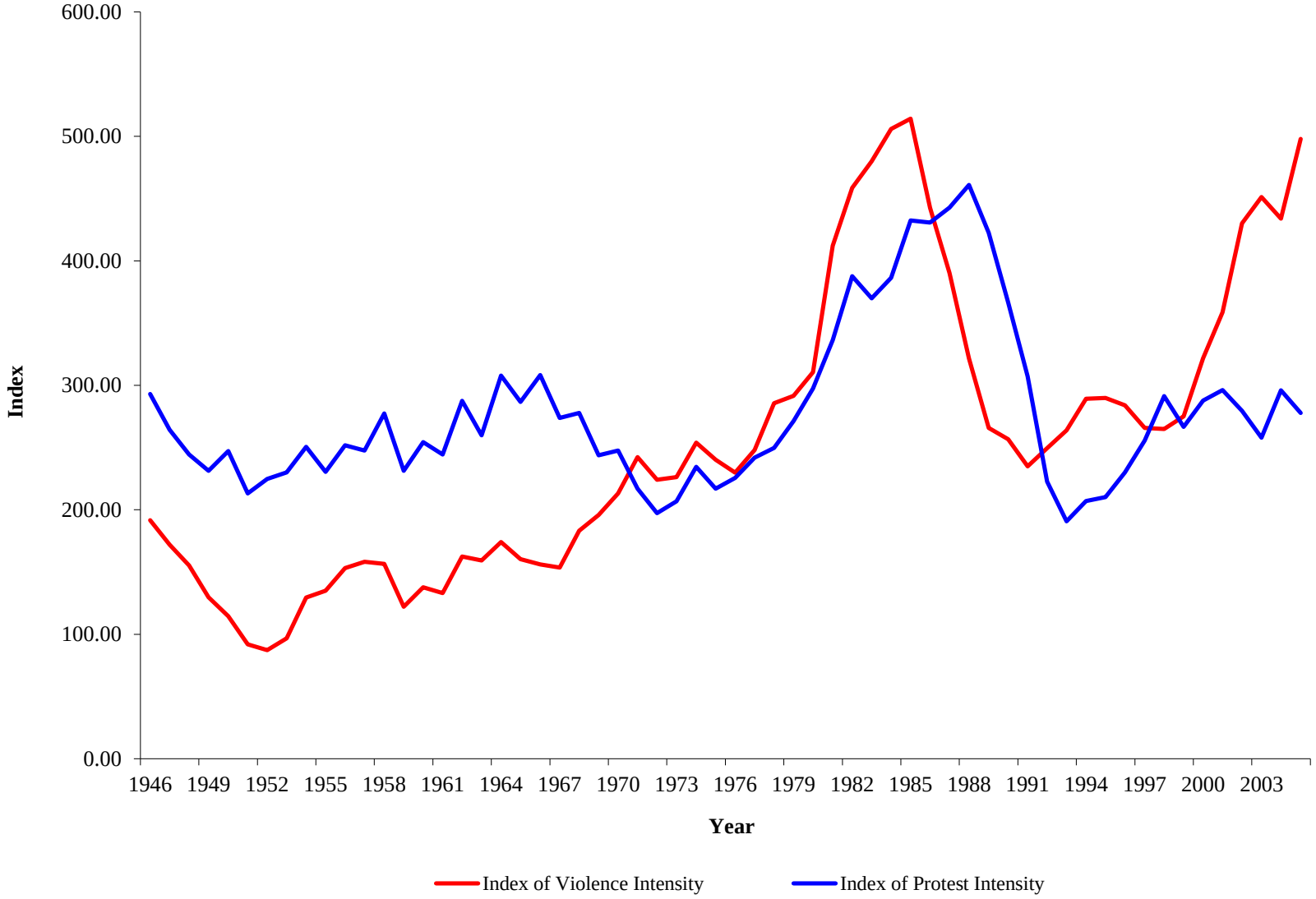
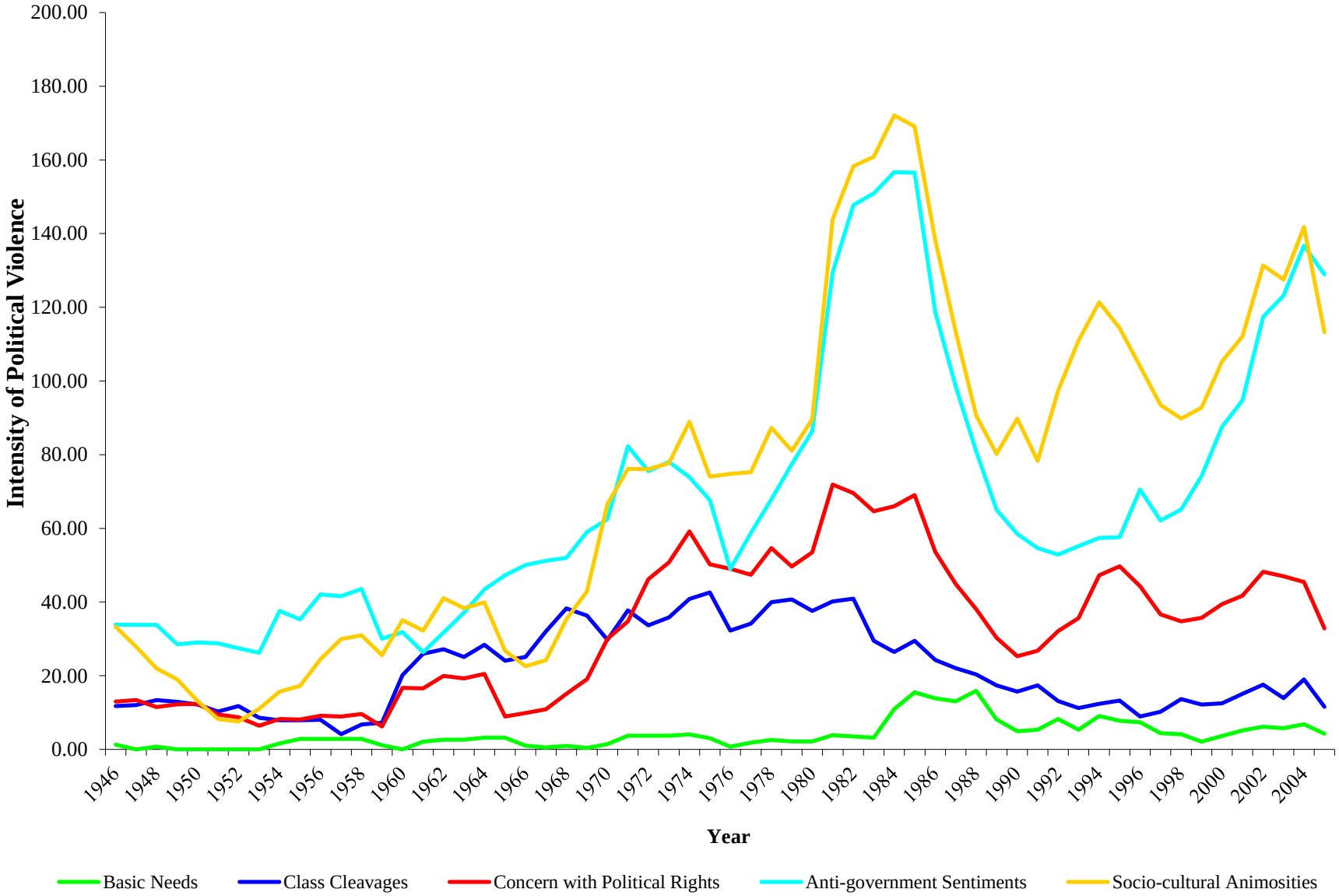


Figure 7
Intensity of Political Violence by Origins
(Five-year moving average)



References

Fearon, James D., and David D. Laitin. 2003. Ethnicity, Insurgency and Civil War. *American Political Science Review* 97:75-90.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Havard Strand. 2002. Armed Conflict 1946-2001: A New Dataset. *Journal of Peace Research* 39:615-637.

Holsti, R. 1964. An adaptation of the "General Inquirer" for the systematic analysis of political documents. *Behavioral science* 9 (4):382-8.

Kahl, Colin H. 2006. *States, Scarcity and Civil Strife in the Developing World*. Princeton and Oxford: Princeton University Press.

King, Gary, and Will Lowe. 2003. An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. *International Organization* 57 (3):617-642.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review* 97 (2):311-331.

Michel, J. B. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *SCIENCE - NEW YORK THEN WASHINGTON*- 331 (6014):176-182.

Mikhaylov, Slava, Michael Lerner, and Kenneth Benoit. 2008. Coder Reliability and Misclassification in Comparative Manifesto Project Codings. In *66th MPSA Annual National Conference*. Chicago, IL.

Monroe, Burt L., and Philip A. Schrodt. 2008. Introduction to the Special Issue: The Statistical Analysis of Political Text. *Political Analysis* 16 (4):351-355.

Schrodt, P. A., S. G. Davis, and J. L. Weddle. 1994. Political Science: KEDS-A Program for the Machine Coding of Event Data. *SOCIAL SCIENCE COMPUTER REVIEW* 12 (4):561.

Stone, Philip J. 1962. *The general inquirer : a computer system for content analysis and retrieval based on the sentence as a unit of information*. Harvard: Laboratory of Social Relations, Harvard University.