MEDIA DATA AND SOCIAL SCIENCE RESEARCH: PROBLEMS AND APPROACHES

ABSTRACT

News media provide a unique source of information on important societal developments, both contemporary and historical. Consequently, over the past forty years, social scientists have attempted to utilize media data to study important questions in a number of fields. But these efforts have been subjected to sobering critiques in an on-going debate over the utility of media data in social science research. The advent of the Information Age has both raised the stakes of this sustained debate and restructured it. Over the past several decades we have seen the emergence of the Internet, the rise of news websites, the widespread availability of digitized news reports, and the creation of 24x7 news stations. These developments have led to unprecedented increases in the volume, scope and accessibility of news reports. Advances in data science and computational capacity have greatly enhanced the ability of researchers to process information embedded in those news reports.

The confluence of these developments has laid the groundwork for third-generation media data projects that have the potential to generate major advances in several fields of research. But the implications of these Information Age developments for the sustained debate over the utility of media data have not been explored and, without a better understanding of those implications, the potential of third-generation projects may never been fully realized. Thus, this paper re-examines the on-going debate over media data in light of these recent developments. We begin by summarizing the key issues raised by critics and asserting that they identify three sets of problems with media data: a comprehensiveness problem, an identification problem, and a distortion problem. In the second main section of the paper we decompose each of these issues and assess their implications for contemporary research employing media data. In this assessment we focus on civil strife research, but most of the main points pertain beyond this field. We also discuss the potential for the remediation of the problems that posed serious threats to the utility of media, with an emphasis on third-generation research efforts like the Social, Political and Economic Event Data (SPEED) project.

CONTRIBUTORS

Peter F. Nardulli
INTRODUCTION ............................................................................................................................... 1

MEDIA DATA AND SOCIAL SCIENCE RESEARCH: SOURCES OF CONCERN ...................... 2

   NEWS VALUE .................................................................................................................................................. 3
   EVENT CHARACTERISTICS .......................................................................................................................... 3
   INSTITUTIONAL CONTEXT OF THE NEWS GATHERING SYSTEM .......................................................... 4
   CONTEXTUAL FACTORS ............................................................................................................................. 5
   RESEARCH PROCEDURES .......................................................................................................................... 5

MEDIA DATA AND SOCIAL SCIENCE RESEARCH .................................................................... 6

   SYNTHESSES, IMPLICATIONS AND REMEDIATIONS ........................................................................... 6
   THE COMPREHENSIVENESS PROBLEM: DEFICIENCIES IN NEWS COVERAGE ......................... 7
      Implications for the Utility of Media Data ................................................................................................. 7
      Compensatory Efforts .............................................................................................................................. 9
   THE IDENTIFICATION PROBLEM: FINDING RELEVANT DOCUMENTS AND INFORMATION ........ 11
      Implications for the Utility of Media Data ................................................................................................. 11
      Compensatory Efforts .............................................................................................................................. 11
   THE DISTORTION PROBLEM: BIASES IN THE INFORMATION REPORTED AND IDENTIFIED .......... 13
      Implications for the Utility of Media Data ................................................................................................. 13
      Compensatory Efforts .............................................................................................................................. 15
      Irremediable Distortions ............................................................................................................................ 15
      Interacting Distortions .............................................................................................................................. 15
      Poor Research Choices ............................................................................................................................. 16
      Remediable Distortions .............................................................................................................................. 18
      Weighted and Unweighted Civil Strife Intensity Scores, 1980-2005 ..................................................... 20

SUMMARY AND CONCLUSIONS ................................................................................................. 21

REFERENCES .............................................................................................................................................. 22

Table 1 ......................................................................................................................................................... 20
INTRODUCTION

News media are a unique source of information on the world in which we live. As a result, shortly after the behavioral revolution in social science research, scholars began to employ newspaper records as a source of data. Much, but not all, of this work was in the field of civil strife/contentious politics and focused on such things as mass protests, political violence and strikes (Gurr 1968; Jenkins and Eckert 1977; Lieberson and Silverman 1965; Spilerman 1970; Shorter and Tilly 1974; Eisinger 1973). Most of this early research employed newspaper articles from a single newspaper (usually the New York Times), focused on a single country (usually the United States), and employed fairly crude approaches to identifying relevant articles (subject matter indexes). While this early work was initially well-received, a number of scholars raised telling methodological criticisms of the use of media data in social science research (Danzger 1975; Jackman 1979; Snyder 1977; Taylor and Hudson 1972; Hazelwood and West 1974).

These critiques notwithstanding, the appeal and potential of media data led to a wave of second generation studies that began to appear in the 1980’s (Koopmans 1993; McAdam 1982; McAdam and Su 2002; Taylor and Jodice 1983; Kriesi et al. 1995; Olzak 1989). These studies were more sophisticated theoretically and methodologically, they were more likely to employ multiple sources and/or digitized news services (Reuters, AP, Agence France Presse, etc.), as well as keyword searches to identify relevant articles. Moreover, more of the research was focused on multiple countries, or at least countries other than the United States. Finally, the 1990s saw the re-emergence of automated content analysis of media data using more powerful computers and more sophisticated algorithms. (Schrodt 2006; Schrodt, Davis, and Weddle 1994; Schrodt and Gerner 1994; Schrodt et al. 1994; Gerner et al. 2002; Jenkins and Bond 2001).

These advances did little to quiet the critics of media data and their continued use spurred a flurry of new reviews (Barranco 1999; Hocke 1998; Koopmans 1995; Mueller 1997; Myers 2004; Oliver 2000; Franzosi 1987). However, despite the sustained and multi-faceted critiques of the appropriateness of media data for social science research, they continue to be used by scholars; some have even offered well-reasoned defenses for its continued use (Earl et al. 2004). Moreover, there is substantial middle ground in the media data debate. Even the most critical commentators do not call for abandoning their use; most simply want users to take the shortcomings of media data seriously, give thought as to how the methodological issues can be addressed, and recognize the implications of those shortcomings for the inferences they draw (Ortiz et al. 2005, 412). Even the most ardent advocates recognize the limitations of media data and do not advocate ignoring them. Rather, they focus on the potential of media data and are less inclined to view them as uniquely or fatally flawed, but on a par with other sources of social science data (surveys, archival records, observations, etc.).

While the debate over the utility of media data for social science research has been conducted for over forty years, the stakes have never been greater. The advent of the Information Age, during which we have seen the emergence of the Internet, an explosion of news websites, the widespread availability of digitized news reports, and the creation of 24x7 news stations, has led to an unprecedented increase in the volume and scope of information provided by news media. Equally important is the emergence of advanced data science tools such as machine learning and natural language processing for analyzing unstructured data (text, images), such as is found in media data. These developments have shifted the ground upon which the debate over media data needs to be conducted and have laid the groundwork for third-generation research projects employing them. Third-generation media data projects have the potential to draw on a greater variety of news sources over a longer period of time, cover a greater geographic expanse, and employ more sophisticated tools for extracting information from news texts. But realizing this potential will require a better understanding of the research implications of media data’s shortcomings and how those
shortcomings can be addressed, using both Information Age advances and the innovative use of conventional social science methods.

The Social, Political and Economic Event Database (SPEED) project is an example of a third generation media data project. Its structure has been shaped by the availability of digitized news sources, advances in data science, and the media data debate. SPEED uses a hybrid workflow system to extract rich event data from a large digitized global news archive containing tens of millions of news reports from diverse news sources for the post WWII era; it is updated daily by scraping news reports from a global set of news websites. SPEED’s hybrid system consists of both automated components and human-centric components. Together they form what we call a *progressive supervised learning system* (Nardulli, Althaus, and Hayes 2014). In this system, human coders are presented with input data that have been automatically pre-processed and classified. Humans perform only the most difficult coding decisions, leaving the more mundane work to automated processes. Combining wide-but-shallow machine capabilities with deep-but-narrow human capabilities leverages the advantages of each while limiting their liabilities.

In designing SPEED’s hybrid workflows and analyzing the data generated by them, we have considered the methodological limitations of media data that have been raised and, where possible, have adopted measures to minimize their impact. This white paper outlines those measures. To better understand what we have done and why we did it, the next section provides a succinct overview of the on-going media data debate. This overview has been facilitated by a set of recent articles that have nicely summarized the enduring issues in the debate (Earl et al. 2004; Myers and Caniglia 2004; Ortiz et al. 2005; Althaus et al. 2011). In the second section we argue that, in coming to terms with the key issues in this debate, it is useful to think of it as having highlighted three major problems with using media data in social science research: a comprehensiveness problem, an identification problem, and a distortion problem. After introducing each problem we assess their implications for research, using research in the field of civil strife to provide concrete examples. We then discuss what has and can be done to address those problems, with a focus on the measures adopted in the SPEED project and its Societal Stability Protocol (SSP), which generates event data on civil strife.

**MEDIA DATA AND SOCIAL SCIENCE RESEARCH: SOURCES OF CONCERN**

The debate over the appropriateness and utility of media data stems from the undeniable observation that these data are not generated by trained social scientists for the purpose of conducting rigorous research on social matters. Ortiz et al., for example, observe that “… newspaper content is not created for the purpose of conducting social science research nor is it intended to capture or sample all protests or other political events, even in a limited geographic area (Ortiz et al. 2005, 397).” Their observations, obviously, extend to virtually any topic covered by the news media, not just protests or political events. Thus, media reports do not capture everything that happens, even within topics and areas of concern to their primary audience. Compounding this reality is an observation made by Myers and Caniglia, who note that “… recent studies make it plain that the media are far from transparent conduits of information about political events. The media use various filtering mechanisms that produce substantial representation of some events while neglecting others (Myers 2004, 521).” The reason the media do not provide an unbiased sampling of important happenings is that they are driven by concerns that are quite different from those that would motivate social scientists if they were in control of the resources and processes used to generate news documents. Indeed, Althaus et al. (2011), elaborating upon on reflections made nearly a century ago by Walter Lippmann, observe that news outlets “serve their own logic, and the logic of the news business creates important limits to … what can be broadly known through the news (Althaus et al. 2011, 1065).”
Media data critics argue that the logic of the news business gives rise to a media selection process. Because that process filters happenings that unfold across the world it can generate an incomplete and distorted view of domains that are of interest to the social sciences. Critics also contend that the issues posed by the media selection process can be compounded by the research procedures used by social scientists to identify relevant events in news reports and to extract information from them. To better understand the research implications of these issues, as well as how they can be addressed, requires a probing reassessment of the key issues in the media data debate in light of Information Age developments. To begin this reassessment the following four subsections provide a succinct overview of that factors that affect the media selection process (news value, event characteristics, institutional setting, context); the fifth addresses misgivings about the research procedures that have been used.

**NEWS VALUE**

According to Althaus et al. (2011, 1069) what journalists call “news value” is an important factor in the media selection process. They go on to note that the standard elements of news value include timeliness, novelty, geographic proximity, familiarity, audience impact and drama. These factors are important elements of the triage process news organizations use to determine which stories to include in scarce news space; the greater the news value of a story, the greater the likelihood that it will survive the triage. While all components of news value are important, they note that drama and novelty are especially important. Sometimes the novelty and drama of developments are so striking that their news value trumps other influences on the media selection process, thus offsetting the distortions that may result from institutional or contextual factors.

But factors related to news value can also work to exacerbate the bias generated by media selection processes. For example, in places like Africa, where conflict have been commonplace over the past several decades, it is less likely to be considered novel and, hence, less likely to be covered, all other things being equal (Althaus et al. 2011, 1069). Myers and Caniglia (2004), who emphasize the importance of news value factors such as geographic proximity and its interaction with audience interest and infrastructure matters, argue that

… attention to international events likely suffers even more from selection processes. In addition to even greater distances and the accompanying infrastructure challenges, cross-border political, cultural, and language differences reduce relevance to local audiences, thereby reducing coverage and comprehensiveness (Myers 2004, 538).

**EVENT CHARACTERISTICS**

There is uniform agreement that event characteristics are an important, and perhaps the best empirically documented, influence on the media selection process (Myers 2004, 524). Event intensity (violence, demonstration size, number killed or injured, duration of the event, etc.) is perhaps the most important event characteristic affecting the media selection process. But the political significance of the actors and the locale are also important, along with organizational sponsorship (Ortiz et al. 2005, 399). Althaus et al. (2011, 1068) differentiate between routine events (press conferences, speeches, annual events, etc.) and accidental events (natural disasters, scandals, crimes, domestic or international crises, etc.). Accidental events are more likely to survive the media selection process, in their view, because accidental events “empower journalists to define the news on their own terms, to glimpse how institutions respond to problems when the veil of spin is temporarily lifted, and to verify claims made by political leaders about the nature of social problems (Althaus et al. 2011, 1068)”
Event characteristics are important to note when thinking about the media selection process because they feed into “news value” and have the capacity to offset the bias that may result from other factors. Moreover, large, dramatic events also have the capacity to attract the attention and resources of news organizations to geographically distant locales, thus offsetting the bias introduced by proximity effects. Thus, one unsurprising effect of event characteristics is that intense, dramatic, or important events are more likely to survive the media screening process.

**Institutional Context of the News Gathering System**

Media organizations are profit-driven entities operating in a capitalist economy – or state instruments working in an authoritarian environment – not social science research institutes operating in a university setting. There is widespread agreement on the part of media scholars that this institutional context has both direct and indirect effects upon the media selection process. Ortiz et al. point out the most direct and obvious influence: self-interest (either economic or political) can affect the types of events covered, how the story frames the event, the diversity of those interviewed to provide perspective, etc. (Ortiz et al. 2005, 400). The mix of influences exerted by media operatives can be compounded by the contrasting interests of different advertisers and political patrons; they can change over time as ownership shifts, the media industry becomes more concentrated, and political regimes change.

The institutional context within which the news is produced also has an array of indirect effects. Althaus et al. talk about the impact of efforts to maximize the potential for generating an on-going flow of newsworthy stories under resource constraints (Althaus et al. 2011, 1067). This leads media organizations to locate strands of the news net at strategically important locations (e.g., key government institutions), as well as utilize beat reporters and wire services. These organizational strategies enable media groups to collect and report the news efficiently but it also has a distorting effect on the news because of what they are not capturing with this resource allocation strategy. Thus, with respect to political issues, if prominent individuals located in powerful institutions are not talking or thinking about a topic, then it is less likely to become newsworthy. Ortiz et al. also stress the important effect of what they call “media mechanics;” if something happens that is outside established beats and routines it is less likely to be reported (Ortiz et al. 2005, 401).

Attention cycles and the size of the “news hole” are other widely recognized institutional effects. When certain issues (political corruption, urban riots, drug offenses, etc.) gain prominence because of earlier happenings, news organizations are more likely to report on similar events. This is what is referred to as an attention cycle effect (Ortiz et al. 2005, 401). Because of it the probability of identical events being reported will vary depending solely upon when they occur, thus distorting media-generated views of reality. Also, as media fatigue sets in, events that occur toward the end of an attention cycle are less likely to be covered. The size of the news hole on a given day has an equally important distorting effect. Media organizations, particularly newspapers and television stations, have limited space for news and their audience has a limited interest in the news and a limited capacity for absorbing it. Thus, decisions must be made about what appears. When the density of newsworthy events is great, there is more for gatekeepers to choose from, and vice versa. Thus, the news hole effect can generate distortion.

One other set of institutional influences has to do with audience characteristics and tastes, as well as the self-definition of the media outlet. As Althaus et al. point out, the news is a business and this generates economic incentives and constraints (Althaus et al. 2011, 1068). In order to make a profit it must take into consideration the tastes of its audience. This can lead to underreporting of events that upset their audience or are not of interest to them. Also, different media outlets have different images of how they see
themselves. Some view themselves as liberal while others view themselves as conservative; these differences can lead them to under or over emphasize different types of stories (labor strikes, anti-government protests, gay right marches, etc.). Moreover, media outlets that market themselves as local outlets will emphasize different stories than those who define themselves as national or international in scope.

**CONTEXTUAL FACTORS**

A number of contextual factors also affect the media selection process, although some media scholars intermix contextual factors with those reviewed above. Some contextual factors introduce temporal bias, others introduce a spatial bias. For example, both Ortiz et al. and Althaus et al. view the geographic location of events as factors affecting coverage. More “important” countries, urban areas, regions of the country (coasts), and proximity to news agencies and targeted audiences are examples of operative geographic factors. Events not occurring in these prime locations are less likely to be covered, which leads to a distortions in how the news portrays reality. Temporal biases can be introduced in at least two ways. First, Althaus et al. (2011) argue that the current political context will affect coverage in that what is perceived to be important or relevant will change over time. Thus, when an event transpires could have an effect on whether it is captured by media. The emergence of an authoritarian regime or the imposition of martial law will also lessen coverage of some types of events. A second source of temporal bias is technological. News technologies have changed dramatically in recent decades. The emergence of news websites with greater capacity for coverage and 24x7 TV news networks greatly enhances the chances for an event to be covered by someone somewhere. Because of competition and technological developments, the probability of an event being reported with vary, which could lead to misleading temporal comparisons.

**RESEARCH PROCEDURES**

Ortiz et al. (2005, 398) stress the importance of research procedures in affecting whether the happenings that get covered by the media are captured in social science databases. They argue that researchers themselves can produce selection effects because

> Only in rare cases can social scientists collect information from all newspapers that might have covered the event of interest (citations). Instead, scholars usually extract information by using some kind of sampling technique that will reduce the time and cost of data collection (Ortiz et al. 2005, 402).

They correctly point out that there is the universe of events, the happenings that appear in the newspaper, and the events that are captured in media data projects. How those events are assembled in social science databases and analyses can distort the results and inferences drawn from media data. Ortiz et al. touch on six aspects of the research process that can have the potential to distort the representation of reality they are trying to capture: (1) the selection of news sources; (2) the sampling techniques used to select the news reports to be studied; (3) the procedures used to identify the news stories with information on relevant events (strikes, protests, assassinations, etc.); (4) the procedures used to obtain the desired information in the relevant news stories (dates, participants, weapons, injuries, etc.); (5) the coding of the information identified in the news stories selected; and (6) the handling of missing data (i.e., desired information that is not available or detected in the news reports analyzed).

They correctly point out that most social science researchers cannot afford to analyze a comprehensive set of news sources and often employ only one. Most researchers studying the United States have employed just the New York Times, which has been shown to be the most comprehensive source for U.S. events. But the
Times has been shown to have significant selection biases and, they note, even using the best source does not eliminate bias. Moreover, research has suggested that adding a second source does not do much to expand the coverage of relevant events. Employing electronic media bases (e.g., Lexis-Nexis) is a way to overcome this problem but they do not include all media sources. Sampling schemes used to select news reports to be analyzed are also a potential source of bias because of how media organizations report the news (slow news days, prospective or delayed reporting of events, heavy news days). Thus, the sampling scheme adopted can lead to misrepresentations of the events that unfolded on a particular day.

Once a sampling strategy for a project has been selected, researchers must find the news stories that pertain to the project’s focus and then locate relevant information within the story. News indexes are notoriously poor tools for identifying relevant stories and Ortiz et al. report “undocumented” problems in using electronic search mechanisms across databases (Ortiz et al. 2005, 403). But even when the relevant stories are found the cognitive limitations of human coders leads to events and/or information about events that are not identified. Then there is the problem of coder proficiency in capturing the information that is identified and the issue of intercoder reliability. Finally, there is the problem of missing data. Even if a representative sampling of events is achieved and coders are well-trained, there is the problem of desired information that is simply not reported. How this is handled, Ortiz et al. note, can have an effect on the substantive conclusions and inferences.

MEDIA DATA AND SOCIAL SCIENCE RESEARCH

SYNTHESSES, IMPLICATIONS AND REMEDIATIONS

Synthesizing the various critiques of media data suggests that critics identify three major problems. The first is a comprehensiveness problem: news organizations fail to capture all of the relevant happenings that unfold in the world. The second is an identification problem: researchers seldom capture all of the information on relevant happenings that the news organizations report; this is especially true for more ambitious projects (i.e., those studying complex topics, collecting longitudinal data, or employing cross-national designs). The third is a distortion problem: the structure of the process by which media organizations select happenings to include in news reports, and the procedures used by social scientists to collect and analyze information derived from those reports, can combine to generate an unrepresentative depiction of the social phenomena being studied.

In this section we examine each of these problems with an eye toward clarifying them and assessing their implications for the utility of media data in social science research. Because of the nature of media data limitations, their research implications will vary across topics (crime, elections, corruption, human rights abuses, etc.) and research designs (objectives, timeframe, geographic focus, information sources, etc.). Surveying these implications across all possible topics and applications would be unwieldy, overly general, and highly speculative. Thus, we focus on civil strife applications. Historically, the bulk of the work using media data falls within this field. Moreover, focusing on civil strife allows us to draw on data and examples from the SPEED project’s Societal Stability Protocol (SSP), which has been used to code nearly 100,000 destabilizing events. In assessing the research implications of media data within this field, we differentiate among irremediable and remediable limitations. With respect to the former we focus on the reach and severity of the issues posed, the validity threats posed by employing media data, and the opportunity costs of not employing them. With respect to remediable limitations we briefly review the remedies that have been used previously and introduce the remedies adopted within the SPEED project, some of which

---

1 For more information on the SSP see: http://www.clinecenter.illinois.edu/research/documents/AnOverviewoftheSSP.pdf
employ data science tools. We also speculate on solutions that third generation event data projects can adopt in the future.

**THE COMPREHENSIVENESS PROBLEM: DEFICIENCIES IN NEWS COVERAGE**

Perhaps the most fundamental criticism of media data is that they do not capture every event of interest to social science researchers that transpired somewhere in the world on a given day. If the media did, in fact, report every relevant happening then it would be an invaluable – and indisputable – source of social science data. But no news source, nor any aggregation of news sources, will capture every event of interest to social scientists, regardless of the field of inquiry. Moreover, this is an inherent, irremediable shortcoming of media data: there is nothing that researchers can do to change this, either for historical or current events. It is simply a “given” that the media does not, cannot, and will not cover everything of interest.

**Implications for the Utility of Media Data**

In assessing the implications of this irremediable flaw it is useful to think both about the specific limitations of media data in civil strife research and relative strengthens and weaknesses of media data in general. With respect to the first point it is useful to begin by thinking about the characteristics of civil strife events that are most likely to be missing.² Because the vast majority of civil strife events are unscheduled events that can involve violence, drama, large numbers of demonstrators, prominent and/or coercive state actors, etc., their news value will be high and many will be covered. Those that do not survive are likely to be relatively minor and isolated events. Whether these events create validity problems will depend on the objectives and design of the research. For example, if the concern is with mapping every manifestation of discontent in a particular venue or time frame, then the comprehensiveness problem will be a debilitating flaw; it will also be a concern if the researcher desires to use minor events as precursors to predict major disruptions.

Fortunately, the validity of many research applications in the study of civil strife does not depend on identifying every destabilizing event. Moreover, the civil strife events that are of most interest to most researchers in the field are not randomly distributed over time and space. Rather, they are highly concentrated in discrete spatial and temporal clusters (civil wars, episodes of political instability, social movements, etc.). The societal effects of these concentrated episodes of strife are much more consequential than random and isolated strife events and studying them is an important concern for civil strife researchers. To illustrate the extent to which civil strife events are clustered in strife episodes we examined the distribution of events from a project that identified major and middling episodes of instability within 165 countries during the post-WWII era. The data for the project was derived by implementing SPEED’s SSP on a random sample of New York Times articles between 1946 and 2005. The analysis revealed that civil strife events are highly clustered: of the 56,000 non-US civil strife events, 84 percent fell within an episode.³

The comprehensiveness problem is not a significant validity threat for research focused on civil strife episodes, for several reasons. One derives from the nature of the research. Trying to identify and gauge temporally and spatially clustered surges in civil unrest does not require that every civil strife event be captured in news reports, as long as there is not uncorrectable bias in country coverages. Indeed, in such

---

² The potential for distortions due to the media selection process will be assessed later.

³ For more information on this effort see: [http://www.clinecenter.illinois.edu/research/documents/SSP_Episodes_Demarcation.pdf](http://www.clinecenter.illinois.edu/research/documents/SSP_Episodes_Demarcation.pdf)
research initiatives random and isolated events may be little more than white noise. Perhaps the biggest validity threat posed by the comprehensiveness problem pertains to the episodes parameters. In some instances neither the start and end dates nor the patterns of spatial dispersion patterns will be as precise as desired. Consequently, researchers must be sensitive to this issue and be cautious in making inferences about these temporal and spatial parameters, especially in relatively minor episodes in “unimportant” countries.

A second reason has to do with the operation of the media selection process, which can lead to extensive coverage of on-going strife episodes. Sustained periods of unrest have high news value. They often involve drama and the types of intense activities (deaths, injuries, mobs, tragedies, etc.) that will draw the attention and resources of news organizations. While audience interest may vary depending upon the episode location and duration, a well-designed research strategy can employ media data to generate valuable and valid insights into civil strife episodes.

One last reason that the comprehensiveness problem is less of a validity threat in studies of sustained conflict derives from a misconception of the media selection process as it is portrayed in most critiques of media data. An implicit assumption in most depictions of this process is that it is a “one-shot” matter. That is, because timeliness is such an important component of news value, most discussions of the media selection process implicitly assume that if an event unfolds and it is not included in a news report immediately, it is never captured. This is central to criticisms anchored in such things as the operation of media mechanics, the size of the news hole, and scarce media resources. Fortunately, this is an oversimplification of journalistic practices. Reviews of thousands of news reports show that an important component of the journalistic repertoire is the “recapitulation passage.” A recapitulation passage is a summary statement of prior events that often appears in news reports about related events. The events described in them are often less detailed than the report’s focal event, but much valuable information is provided. To illustrate, consider the example of the street vendor in Tunisia whose plight initiated the Arab Spring. Very few, if any, major outlet initially covered this event. But it was widely reported as background information as the strife endured and diffused.

Recapitulation passages are extremely valuable devices in that they provide news organizations with the means to play “catch up” in a geographic region that is outside their normal interests/beats and in which they were not investing many resources. It minimizes the impact of the undeniable observation that these organizations do not have the capacity to cover important developments everywhere they unfold. While resource constraints are real and important, the attention and resources of media organizations will be focused on situations where strife endures. In covering those unfolding episodes they will inevitably be drawn to capturing recapitulations of prior events, which will minimize the impact of the comprehensiveness problem.

In assessing the implications of the comprehensiveness problem for the utility of media data it is also important to evaluate its relative utility: is a lack of comprehensiveness unique to media data. If so, then it would suggest that analyses based on them would justifiably be relegated to an inferior status within social science research. But a moment’s reflection suggests that media data are not unique in this regard. Are measures of GDP based on a compilation of every widget and service produced in every country since measures of economic activity have been produced? Are educational statistics based on the progress of every student enrolled in every school in every reporting district in the world? Are census data derived from interviews in every household? Are experimental (survey) data drawn from a random sample of subjects (respondents) from the population to which the researcher wants to make inferences? The answer to each of these questions is an unequivocal “No” and the notion that media data are somehow uniquely flawed is
simply without merit. Thus, when admittedly incomplete media data are compared with other sources of social science data, which are also incomplete, the trade-offs do not seem as stark.

Also without merit is the notion that “In few others fields of the social sciences does the researcher have so little control over the data on which analysis begins (Mueller 1997, 821).” This is only true if the social sciences are limited to data sources such as self-designed and executed surveys or experiments. The notion that social scientists have more control over Nigerian bureaucrats who complete educational attainment surveys or the Paraguayan minister that generates economic statistics simply does not withstand scrutiny. And to limit social science research to surveys and experiments would drastically curtail its reach and relevance. Indeed, some research questions cannot be examined quantitatively without relying on media data. This is certainly true in the field of civil strife, particularly if a comparative or longitudinal research design is needed. In these instances, Casey Stengel’s admonition to “Look it up” has no bearing; there is nowhere else to look; a Day Book of important societal happenings does not exist.

Do these observations mean that media data are always appropriate and “just as good as” other sources of social science data? Here again the answer is an unequivocal “No.” Rather, it depends on the question being asked. Our point is simply that, forty years of criticism of notwithstanding, media data should not be rejected out of hand simply because they provide incomplete accountings of relevant phenomena. Instead, researchers should consider combining different sources of data (media, census, archival, etc.), as well as conduct a probing cost-benefit analyses of alternative sources of flawed data.

Compensatory Efforts

In conducting these cost-benefit analyses researcher should consider the type of methodological “fixes” available for addressing methodological limitations of the available data, as these could influence the choice among suboptimal alternatives. With respect to media data, prior researchers have addressed the comprehensiveness problem by employing multiple news sources or electronic media databases that compile reports from various sources. This approach can reduce the impact of the comprehensiveness problem but it cannot wholly compensate for it. While it has been observed that adding an additional source may not markedly enhance coverage (Ortiz et al. 2005, 403), electronic media databases offer a better alternative as they are more inclusive. While no compilation captures all major news sources, using them can greatly enhance the coverage of events.

Because the information needs of the SPEED project are broad and deep (we desire media data for most countries in the world for the entire post WWII era), we devised a strategy that was more encompassing than prior efforts and provided for highly sophisticated search, indexing and clustering capacities. This strategy led to a unique electronic news archive containing tens of millions of news documents. These documents are preprocessed using natural language processing (NLP) techniques to extract information on such things as dates, places, entities, etc.; the preprocessed documents are stored in a SOLR data archive, which is a powerful data management system for storing and processing unstructured data (i.e., text and pictures). The archive contains both historical documents (1946-2005) and contemporary documents (2006 -).

The historical component of the archive contains documents from four sources. The first two are digitized documents from traditional newspaper sources, the New York Times (NYT; 5.3M documents) and the Wall Street Journal (WSJ; 1.4M documents). To provide for broader international coverage we secured the entire

---

4 These archives were obtained under a written agreement with ProQuest, a firm that owns the distribution rights to them.
set of microfiche, microfilm, and “born-digital” open-source records from two intelligence agency aggregation services: the Foreign Broadcast Information Service (FBIS) and the Summary of World Broadcasts (SWB). These aggregations were compiled by field agents who monitored thousands of broadcast, print, and internet news outlets; many were translated into English from scores of languages by culturally sensitive native speakers. The born-digital records for the FBIS and SWB compilations contain over 8 million documents. But the majority of their holdings were never digitized. To incorporate them into our repository we: (1) scanned over 50,000 microfiche and 1,000 microfilm reels; (2) used an OCR program to digitize the scans; and (3) segmented the digitized output to identify document boundaries. The segmentation component of this project was a major, interdisciplinary effort that was managed by a team of computer scientists from the University of Illinois’ National Center for Supercomputing Applications (NCSA) and benefited from a grant of supercomputing resources from the Extreme Science and Engineering Discovery Environment (XSEDE) project, an NSF initiative. These efforts added more than 20 million documents to our archive.

We invested in the digitization of the FBIS and SWB records because several experiments suggested that their international coverage was far greater than the NYT. For example, a comparison of the NYT and SWB archives for a comparable set of climate-related, rapid-onset disaster episodes suggests that the SWB archive has information on thirteen times as many destabilizing events (Nardulli, Peyton, and Bajjalieh 2013). A more limited experiment involving just Liberia and Sierra Leone shows that FBIS contains 5-6 times as many civil strife events as SWB. That experiment also showed that the combined coverage of violent attacks in FBIS, SWB and the NYT compared favorably with those attacks identified by the Armed Conflict Location and Event Data (ACLED) project, even though ACLED focused mainly on Africa and relied on as many as fifty news sources, including local outlets.

To construct a contemporary news archive we employed a Heritrix-based web crawling system to access global news websites, download news reports, and eliminate duplicate reports. The remaining documents were preprocessed and stored in the Cline Center’s SOLR news archive. To optimize this strategy, much effort was invested in identifying a comprehensive set of RSS news feeds from English-language websites from across the world; almost 900 were identified. These news sources include: (1) major news aggregators such as Google News, Yahoo News, and the Big News Network; (2) wire services such as Reuters, AP, and UPI; (3) regional news feeds that such as AllAfrica News, Africa and Middle East Info; (4) national and local news feeds such as the Afghanistan Sun, the Times of India, Xinhua, London Financial Times, Philippine Times, and the Malaysia Sun; and (5) specialty news feeds (business, politics, security, law, humanitarian matters, etc.), such as Amnesty International USA, JURIST – World Legal News, MENAFN.com Politics, etc. Since 2006 millions of news reports have been stored and we are currently downloading nearly ten thousand non-duplicate news reports a day.

---

5 For more information on ACLED see: http://www.prio.no/Data/Armed-Conflict/Armed-Conflict-Location-and-Event-Data/

6 For more information on this comparison see: http://www.clinecenter.illinois.edu/research/documents/SPEEDInformationBase.pdf.

7 In terms of the overall composition of the website list, there are one hundred thirty-eight general news feeds with a global or no specific geographic focus; eighty general news feeds with a regional focus that includes every region in the world; three hundred and thirty-one general news feeds with a national focus that includes one hundred seventy-five countries; one hundred forty-three news feeds with a subnational focus; and one hundred eighty-six specialty feeds.
To maximize the extraction of information from these sources the SPEED project has adopted procedures that capture information contained in recapitulation passages. A review of recapitulation codings demonstrates that they are an important tool for mitigating the impact of the comprehensiveness problem, particularly for distant geographical locales. For example, while recapitulation codings from the New York Times account for only 7.5% of U.S. codings, they accounted for 12.5% of codings for all other countries. Moreover, they are a valuable source of information. Recapitulation codings account for 17% of mass protests, 21% of all politically motivated attacks by non-state actors, and 64% of the killings that were reported.

THE IDENTIFICATION PROBLEM: FINDING RELEVANT DOCUMENTS AND INFORMATION

Acquiring as comprehensive a set of news documents as possible is only the first challenge in enhancing the social science utility of media data. Equally important, though often unappreciated, is the challenge of finding information within those documents that is relevant to the focus of the research (e.g., mass protests, elections, corruption, political violence, etc.). Even with highly sophisticated strategies not every relevant document in an archive will be identified, nor will every relevant passage within a document be uncovered. Moreover, as the size of news archives assembled increases in order to minimize the comprehensiveness problem, the challenges involved in identifying relevant articles and texts will also increase. Fortunately, this is a remediable problem and some of the data science tools that power third generation event data projects can be used to address it.

Implications for the Utility of Media Data

Identifying relevant units within a larger population is not unique to research that utilizes media data. Researchers who work with archival data often confront similar problems, as do survey researchers. Moreover, as long as the failure to identify relevant news reports and passages is random (i.e., due to human cognitive deficiencies, fatigue, etc.), the impact of the identification problem will simply be the introduction of noise in the data. But the comprehensiveness problem already introduces noise into the data and the use of suboptimal identification procedures, or the poor execution of effective procedures, will only compound the problems caused by deficiencies in news coverage. Too much noise will weaken underlying relationships in the data, enhancing the likelihood of false negatives and leading to null findings and misleading inferences. Thus, it behooves researchers working with media data to address the identification problem as effectively as possible.

Compensatory Efforts

First generation works employing media data used newspaper indexes to identify relevant articles and trained coders to identify and record needed information from relevant passages. This was clearly a suboptimal strategy. The indexes were not intended to meet the needs of social scientists and often lacked the precision required for rigorous research; moreover, the available classifications seldom matched the researcher’s needs. Moreover, regardless of the training regimens, human coders suffer from cognitive deficiencies resulting in missed information (i.e., information on main events, secondary events, recapitulations, etc.) within the documents identified. While second generation works made no progress in dealing with human limitations, they were able to use simple key word searches in electronic databases and this improved the identification process. But keyword searches are crude tools with limited capacity for conducting sophisticated searches (political attacks in Turkey initiated by Kurds between 1985 and 1995). It is also difficult to know, a priori, the complete set of key words needed for an effective search, which will generate false negatives (incorrectly discarded documents). Moreover, language ambiguities are such that
crude key word searches will generate many false positives (incorrectly included documents), which will increase both cost and coder fatigue. Finally, key word searches can work differently across databases (Ortiz et al. 2005, 403).

Third generation works can capitalize on recent developments that will allow researchers to address the identification problem and greatly enhance their capacity to generate more refined subsets of relevant articles to analyze. These advances result from recent developments in fields such as machine learning and the availability of more powerful cyberinfrastructures for managing unstructured data. The SPEED project has taken advantage of both developments. For example, SPEED employs a progressive supervised learning system (Nardulli, Althaus, and Hayes 2014) to identify relevant news reports. A subset of training and test documents is developed by humans who separate documents into “relevant” and “irrelevant” categories. The training data set containing both types of documents is used to derive statistically based algorithms based on key words, word correlations, and semantic structures. The algorithms are then analyzed with the test data set and refined until they produce a satisfactory classification of documents (i.e., one with minimal levels of false positives and false negatives). Once derived, the algorithms are used to automatically classify the tens of millions of news documents that reside in the Cline Center's SOLR archive. The human processing of the “relevant” documents generates additional training data, leading to progressive improvements in the algorithm.

The initial implementation of SPEED’s progressive supervised learning system was used in conjunction with events that fell within its civil strife ontology (protests, politically motivated attacks, state repression, etc.). When assessed in light of the identification problem, the results are encouraging. Analyses of random samples of news documents classified as “irrelevant” revealed a false negative rate of 1-4%, depending upon the source of the news report (NYT, FBIS, SWB, etc.). Thus, it identified about 96-99% of the relevant reports. Unfortunately, analyses of the documents classified as “relevant” showed a false positive rate of 65%, demonstrating that two out of every three documents classified as “relevant” were not, imposing significant human costs. Subsequent classification efforts, building on the review of the automatically classified and humanly processed documents, reduced the false positive rate to 20%, while only increasing the false negative rate to 4%. Thus, the progressive nature of SPEED's automated classification procedures generated enormous efficiency benefits, with only marginal costs.

To evaluate the impact of SPEED’s progressive supervised learning approach on the identification problem we compared its results with a first-generation effort at identification, the World Handbook of Social and Political Indicators (WHSPI, Taylor and Jodice 1983). WHSPI events were identified by trained students employing the New York Times Index (WHSPI, Taylor and Jodice 1983, 418). To conduct a meaningful comparison, we limited the analysis to: (1) those countries (except the US) and years (1948 to 1982) covered by WHSPI; and (2) five straightforward event types for which there is an extended time-series (mass demonstrations, political strikes, politically motivated attacks, assassinations and coups). Because the SSP data represent a sample of only 12.8% of the relevant events identified in the NYT archive, and WHSPI data reflect all events identified from the New York Times Index, a weight of 7.8 is applied to SSP events. This weighted numbers reflects the number of events that would have been generated if all of the NYT articles identified as relevant had been coded. Then we calculated a ratio of weighted SSP to WHSPI event counts

---

8 For more information on this analysis see: http://www.clinecenter.illinois.edu/research/documents/SPEEDInformationBase.pdf.
(ratio=weighted SSP events/WHSP events). The ratios show that SPEED’s identification procedures identified nearly seven times as many events, on average, as WHSPI; the ratios range from 4.2 to 8.5.9

The WHSPI comparison demonstrates that SPEED’s approach to the document identification problem represents a significant advance. However, despite the benefits of automation, there are still huge volumes of text to process within the relevant documents, which poses serious cognitive challenges. To reduce these challenges SPEED’s cyberinfrastructure provides an annotation tool that coders use to manually highlight relevant information in the document. The availability of this tool separates the information identification process from the coding process, thus reducing the cognitive challenges to coders. Moreover, we have been working with computer scientists at the Cognitive Computation Group (http://cogcomp.cs.illinois.edu/) to develop an automated event annotation tool (EAT). EAT employs machine learning techniques in conjunction with tens of thousands of humanly processed event annotations (trigger words for different events, dates, places, actors) to build models that automatically annotate relevant segments of text within news documents. A prototype of EAT has been developed and is being fine-tuned. When properly calibrated, EAT’s annotations will greatly enhance the efficiency, accuracy and reliability of information extraction within SPEED, further minimizing the effects of the identification problem that plagues the use of media data.

Another way in which the identification problem can be addressed is to take advantage of the unsupervised capacity of standard NLP tools to extract key pieces of information (dates, places, named entities) which, within the SPEED project, is done at the preprocessing stage. This automatically extracted information, when joined with the availability of powerful full-text search capacities of data management systems such as SOLR, provides the basis for conducting far more refined searches and sampling of news documents. For example, the SPEED archive has been used in projects that identified news documents that fell within a three-year window surrounding selected climate related disasters (Nardulli, Peyton, and Bajjalieh Forthcoming, 2014), as well as documents that appeared within the time frame of civil wars in such countries as the Philippines, Nicaragua, Sierra Leone and Liberia (Rhodes et al. 2013). These tools provide the capacity to generate more sophisticated sampling strategies than are possible using only paper records or working within the confines of a commercially managed electronic database – without incurring the methodological problems associated with earlier sampling efforts.

THE DISTORTION PROBLEM: BIASES IN THE INFORMATION REPORTED AND IDENTIFIED

The distortion problem facing media data derives from two sources. The first is the media selection process. As noted earlier, media organizations cannot capture every event of interest to social scientists and the information they select to include in news reports does not constitute a random sample of global happenings. Rather, the information selected to appear in news reports is a function of a selection process that incorporates some factors that can distort reality. Among the most important of these distorting factors are event attributes, the location and timing of events, and contextual factors. The second source pertains to the research procedures that social scientists use to transform news content into quantitative data. The most important distortions here derive from data collection procedures and the handling of missing information.

Implications for the Utility of Media Data

9 These results are consistent with other research examining the New York Times Index as a human-centric tool to mine unstructured data (Althaus, Edy, and Phalen 2001; Woolley 2000).
The most important event attribute affecting the media selection process is event intensity. In the study of civil strife intensity reflects such event characteristics as the level of violence (number killed or injured), the presence of lethal weapons, the number of demonstrators, the duration of a sit-in, and the involvement of coercive state actors (military, police). Such event characteristics as the prominence of the actors involved and organizational sponsorship can also be important in the selection process, but these factors likely play a more marginal role: less intensive events involving prominent actors or organizations may be more likely to be covered, but their absence is unlikely to affect the coverage of intense events. Event intensity is perhaps the most important factor in the media selection process because it affects such things as the event's news value, its audience appeal, the likelihood that the media will be aware of them, the self-image of news organizations, and the organization’s profit motive. The most important distorting effect of event intensity is that less intense strife events are less likely to be selected.

Event location affects the media selection process and reflects such things as the event's proximity to the news outlet/wire services/reporter beats; the “importance” of the country where the event occurred, population density, and the prominence of the city/region of the country. These factors affect the selection process because they reflect the news’ institutional setting: audience interest, event visibility, and efficiency concerns. The most important distorting effect of event location is that events that unfold in “unimportant” countries, remote regions and rural areas are less likely to be captured in news reports. The timing of events is reflected in the media selection process because of the importance of attention cycles and the size of “news hole,” both of which institutional realities. There is only limited space, especially for traditional outlets, and choices have to be made about what fills it. The most important distorting effect of event timing is that random, isolated events are less likely to be reported.

A range of contextual factors also affect the media selection process. One is the prevailing ideological climate. This reflects judgments about what is newsworthy but it is probably less consequential with respect to civil strife events. Because strife events constitute a threat to stability and security, which are core human concerns, ideological factors are unlikely to have marked effects on their coverage. Political regimes are also important. Events that might get reported in open, liberal societies may not get reported in authoritarian regimes because reporters do not have access to where events are unfolding or because reporting them will generate adverse consequences. With some extreme exceptions (e.g., North Korea), the impact of authoritarian regimes may have limited effects on the reporting of major civil strife events. The existence of large demonstrations, extreme political violence and massive state repression is hard to conceal – and the emergence and diffusion of mobile media devices is making it increasingly harder to conceal them. Thus, the most important distorting effect of political regimes is that less intense events are less likely to be reported in authoritarian regimes.

Another important contextual factor pertains to barriers to reporting posed by such things as natural disasters and war zones. These barriers could suppress the early reporting of some events, but disasters and wars are magnets for media attention and their relative attractiveness, in the medium to long term, will likely mitigate the impact of adverse conditions. One last contextual effect derives from technological changes that have greatly enhanced the capacity of news organizations to detect and report on newsworthy events. Distortions can emerge when making temporal comparisons that span these technological developments because events that would not have been detected and reported in an earlier era are more likely to be covered after their advent.

Finally, the structure of the data collection process used to gather media data has the potential to generate distortions. One potentially distorting factor pertains to the news sources used. Selecting only one source could introduce biases in such things as its ideological orientation, geographic focus, or intended
audience. A lack of fit between the research objectives and the news sources used could also generate distortions. For example, attempting to study small-bore civil strife events in targeted sub-regions of a country using only national media outlets could be distorting, as would using local newspapers to study international strife events. The sampling procedures used in collecting media data can also cause distortions. They must comport with the idiosyncrasies of the gathering/reporting process (slow news days, heavy news days, norms for covering scheduled events, etc.). The procedures used to code information obtained from the news reports sample are also important. Poor training of coders and the failure to insure intercoder reliability can result in inconsistencies and errors that fail to reflect news content accurately. One final factor pertains to the procedures for handling missing data. In many instances desired information is not available or detected in a news report; how that missing information is treated can affect the image of reality generated.

Compensatory Efforts

The variety of factors asserted to have distorting effects in the on-going debate over the utility of media data make it difficult to outline compensatory efforts succinctly, if they are even available. However, an assessment of the factors just reviewed suggests that there are at least five categories that can be identified and addressed separately. Within the field of civil strife research the effects of these factors can be categorized as: (1) largely irremediable, at least with respect to some research applications; (2) unclear due to interactions among factors affecting either the media selection or research process; (3) more likely to produce noise than create distortions; (4) reflecting poor research choices rather than intrinsic to media data (5) eminently remediable by either conventional research methods or data science tools. Each is discussed below.

Irremediable Distortions

Perhaps the most consequential irremediable distortion has to do with the intensity of civil strife events. News organizations employ intensity thresholds in selecting which civil strife events to report on (number killed, number of protesters, military intervention, presence of weapons, etc.). Moreover, these thresholds vary across news organizations, as well as with attention and news cycles, across countries, and across regions and cities within a country.¹⁰ If relatively minor civil strife events are underrepresented in news documents there is little that researchers can do to compensate. The effect of this distortion is limited, however, because capturing minor strife events is not a concern in all civil strife research, as noted earlier. Its most deleterious effects would be in projects concerned with generating a comprehensive overview of different types of strife events (small-bore expression, violence, demonstrations, repression, coups, etc.) within a locale or timeframe; or with those attempting to gauge the temporal sequences of destabilizing events. Moreover, it may be possible, by accessing local news websites, to minimize the distortions. Because space is less of an issue and they have a local audience, they will capture relatively more intensity events. This fix is limited, however, because these websites are recent developments and their geographic availability is uneven.

Interacting Distortions

¹⁰ For example, it is clear that the New York Times has thresholds and that they vary by region of the world. No one died in 73% of the politically motivated attacks that occurred in North America and were reported in the Times; the comparable figures are 64% in Europe, 59% for Sub-Saharan Africa, and 54% in Asia. Thus, the intensity threshold is much greater in locales that are more distant from the Times’ primary audience.
The long-term debate over the utility of media data has generated a laundry list of distorting factors, as noted above. But there is much less discussion of the interactions between those factors or the impact those interactions may have on the utility of media data. This is unfortunate because these interactions can either enhance or reduce the distortions that each could independently produce. These interactive effects are particularly noteworthy in research on civil strife. It has already been noted that attention cycles can offset the deleterious impact of proximity effects and the neglect of low intensity strife events, but other interactive effects exist.

For example, many of the distorting factors embedded in the institutional setting of news organizations interact with the news value of civil strife events. Media data critics have rightfully asserted that distortions can emerge from strategies media organizations employ to maximize the generation of an on-going flow of news under resource constraints. Events that happen outside of established beats and routines are less likely to be reported. Moreover, because news resources are often allocated to cover important institutions populated by powerful people, matters that are unimportant to those individuals are less likely to be covered. But contentious modes of political behavior (protests, riots, violent attacks, etc.) are the tools of the weak and dispossessed and they occur despite the best efforts of the rich and powerful to control them. Moreover, their emergence is unlikely to conform to any regular news beats – yet they demand and get the attention of news organizations.

Another example of offsetting effects pertains to political bias. For various reasons related to the news’ institutional setting, some news outlets may not want to publicize such things as strikes or gay rights parades. However, most news organizations are profit-driven entities operating in a capitalist economy and most journalists abide by an ethos that compels them to cover important happenings. Thus, the potential effects of political bias can be offset by the pressures from their competitors and the news value of the event. Even when these interactions do not eliminate political bias there are other fixes available to mitigate its effects. The most obvious is to study an array of news sources rather than just one. Also important are coding instructions and procedures that direct coders to focus on factual matters, disregarding framing efforts as well as the views of self-interested commentators, both of which are used in the SPEED project.

Interactions also exist that exacerbate the effect of distorting factors. Perhaps the most obvious is between event intensity and spatial factors. The relative neglect of low intensity strife events is likely to be much greater in distant or unimportant locales. Moreover, it is unlikely that any journalistic ethos or competitive pressures will offset these effects. Also relevant are interactions within locales governed by repressive regimes. In these settings the press is often an instrument of the state rather than a watchdog, and this will exacerbate the effects of political bias.

**Poor Research Choices**

Some of the distortions attributed to media data are not intrinsic to the data themselves. Rather, they are simply poor choices made by researchers. Perhaps the best examples here are selections of newspapers to be studied and sampling techniques used to study them. It is clear that many early media data studies made suboptimal choices with respect to both. Noteworthy here also is the fact that technological developments have made it much easier for social scientists to make much better choices. The availability of electronic repositories of documents from diverse sources, the use of NLP techniques to preprocess the documents, and powerful data management tools allow researchers to access a rich array of news sources and employ sophisticated sampling techniques.

**Random Noise**
The effects of some factors asserted to distort reality are more likely to blur it. One example here is the news hole effect. Because this effect is driven by developments that are independent of a study’s focus (e.g., major protests), the impact of the news hole effect is likely to be a less precise view of reality than a distorted view. The same can be said for poor training/testing practices for human coders. By generating poor and inconsistent codings these practices will yield noisy data that cloud reflections of reality but not systematically distort it. Moreover, while sound research procedures are essential for any research initiative, the use of poor research procedures is not intrinsic to media data. The same can be said for procedures to handle missing data. Not only are they more likely to generate noise than systematic distortions, but they are ubiquitous in all types of social science data, not a unique characteristic of media data.

While the noisy data is different from distorted data it is still an impediment to advancing social science research. Fortunately, there are approaches available that can reduce the noise generated by the use of media data; several have been adopted by the SPEED project. Its capacity to capture information from recapitulation passages, which reduces the noise introduced by the news hole effect, has already been mentioned. Also important are the training and testing procedures employed to ensure that coders are operating proficiently and reliably. Coders working on SPEED’s SSP begin their tenure by participating in an extensive training and testing regimen that requires nearly 70 hours to complete. This regimen includes lectures, one-on-one training, and group training sessions. Training culminates in a series of tests that gauge the coder’s ability to implement the protocol in accord with established norms and understandings; the tests gauge their capacity to identify events and to code them properly. Trainees must pass these “gatekeeper tests” before they are allowed to generate “production” data. Reliability testing continues after coders begin production coding: they are blindly fed “test” articles at established intervals to detect slippages in reliability.11

Critics of media data have also rightfully identified missing information as a significant problem with media data. More precisely, however, the problem is with both missing and ambiguous information. With respect to civil strife events, missing and ambiguous information is a particularly pressing problem with respect to such variables as the number of actors involved (initiators, targets, victims, soldiers, demonstrators, etc.), dates, weapons, and injuries. Thus, during the developmental and pretest phase of the SSP we invested much effort in responsibly addressing this problem; a two-fold approach was developed.

The first component involved the development of a standard specification routine. It provides coders with the capacity to provide: (1) exact numbers, when possible; (2) precise estimate ranges, when necessary; and (3) “ballpark” estimates as a last resort. The routine for providing potentially ambiguous information varies somewhat across variables. But generally the coder is asked to specify what type of information is available (exact information, a limited range of possibilities, a broad range of possibilities, none at all), which prompts a screen that allows them to provide the type of information they have. Ballpark estimates are used when coders do not have precise information but can discern some information from contextual information included in the news report. For example, teenagers assaulting police officers in situations where no one was seriously injured were not likely using military grade weapons or even small arms. While not a perfect solution, using ballpark estimates is superior to post hoc imputations not informed by contextual information.

11 A fuller discussion of SPEED’s training and testing procedures for the SSP, including the results of the reliability tests can be found at: http://www.clinecenter.illinois.edu/research/publications/SPEED-Reliability.pdf.
The second component of the solution involves imputing values for missing data; it is only employed for where no contextual information exists. This component builds on empirical relationships and regularities identified by examining observations with complete information. These observed relationships and regularities make it possible to provide well-grounded imputations for observations with missing data by using information on event traits that is available. For example, more than 90% of politically motivated attacks are committed by 1-2 assailants. Thus, when the number of initiators is unknown and the event type is a garden variety attack, it is possible to provide a defensible estimate. Crosstabulations of event traits also provided the basis for making imputations. For example, good information on the event type (sit-in, boycott, violent attack, etc.), the existence of an injury, and the actors involved (students, clerics, police, soldiers, etc.) can provide guidance for imputing the type of weapon involved.

**Remediable Distortions**

Finally, there are a set of factors that critics have identified that have the potential to distort reality but for which remediation exist. Some of these remedies derive from the creative use of the information that is available in news documents. The role of recapitulation codings to mitigate the impact of attention cycles and proximity effects has already been mentioned. Others involve the thoughtful use of conventional research design techniques. For example, media outlets that market themselves as local outlets will emphasize different stories than those that define themselves as national or international in scope. This suggests the common sense strategy of insuring that news sources used in a study comport with its objectives. Still others involve the use of more advanced data science tools. Perhaps the most important use of these tools in SPEED pertains to our efforts to address country bias in the New York Times. Simple cross-national comparisons of a global random sample of over 65,000 civil strife events from the New York Times during the Post-WWII era (1946 to 2005) showed an alarming amount of bias in country coverage. Simple event distributions suggested that the United States and a number of other Western European nations (Great Britain, Germany, France and Italy. African nations and other Middle Eastern nations looked far more stable. These distributions were almost certainly due to country biases rooted in proximity effects, audience interest, and global prominence. To address this bias we developed a weighting scheme, implemented at the country-month level, by using NLP tools that extracted place names in conjunction a highly inclusive place name database. The place names database included cities, provinces, country names and “nicknames.”

We used these tools to derive, for each month between January 1, 1946 and December 31, 2005: (1) the number of New York Times articles in which each country (or a city or province in the country) was mentioned; and (2) the total number of news articles published. We then calculated the proportion of total articles in a month that included a reference to a country for each country-month; this proportion formed the denominator for the weighting scheme. Next, using linear interpolation, we calculated populations for each country-month and expressed them as a proportion of the global population; this proportion was the numerator for the weighting scheme. The effect of this scheme is to reduce the impact of country-based media bias on our analysis. To illustrate, consider a country that constituted only 1% of the global population in a given month, yet was mentioned in 4% of the NYT articles in that month. Its weight would

---

12 The use of such an expansive set of place names was important because, in many instances, mentions of cities such as Paris, London, Rome, and New York are made without any references to the country within which they reside. Nicknames are important because there are many ways to refer to countries, country names and spellings change over time. The failure to capture these nuances would greatly diminish the validity of the weighting scheme. There were costs associated with using nicknames as some were erroneously assigned to countries, leading to
be .25. In contrast, a country that constituted 4% of the world’s population, but was mentioned in only 1% of the New York Times articles, would have a weight of 4.

Two examine the impact of our weighting scheme we used a per capita measure of civil strife, which captures the overall intensity of protests, political violence, state repression and coup activity for a given country month.\(^\text{13}\) We calculated a weighted and unweighted sum of this composite variable for the period between 1980 and 2005, a period in which the vast majority of our countries were independent. The correlation between the two aggregated measures is modest (\(r=.58; \text{Kendall’s Tau-b}=.34\)). More informative, however, for evaluative purposes is a comparison between the most and least strife-torn countries, which is reported in Table 1.

In comparing the most strife-torn countries across the unweighted and weighted columns, two observations are noteworthy. The first is that six of the countries make both lists: Afghanistan, Iraq, Lebanon, Rwanda, Angola and Guatemala. Those that are on the unweighted list but not the weighted list (the United States, Israel, Russia, South Africa, China, India, the United Kingdom, Pakistan, and Poland) are all regionally or globally prominent countries that are not notable for having high levels of civil strife, but would be of interest to the New York Times primary audience. Those countries that are only on the unweighted list (Burundi, Sierra Leone, Guinea, Uganda, Congo, Sri Lanka, Liberia, Somalia, and the Central African Republic) are mostly small African nations that have experienced extended periods of violent strife, but are not of intrinsic interest to the Times’ primary audience. The composition of the highest ranking countries on the unweighted lists underscores both the power that news value has on attracting media attention to obscure locations and the influence that the news’ institutional setting has on coverage. The composition of the highest ranking countries on the weighted lists suggests that the weighting scheme developed here has the capacity to temper the impact of country bias on news coverage. The list is purged of questionable countries and captures only countries that have experienced major periods of strife since 1980.

---

\(^{13}\) This measure is a weighted sum of the intensity of a set of individual intensity measures. Its derivation is reported at: [http://www.clinecenter.illinois.edu/research/documents/SSP_Episodes_Demarcation.pdf](http://www.clinecenter.illinois.edu/research/documents/SSP_Episodes_Demarcation.pdf).
# Table 1

Weighted and Unweighted Civil Strife Intensity Scores, 1980-2005

<table>
<thead>
<tr>
<th>Country</th>
<th>Unweighted Intensity Measure</th>
<th>Country</th>
<th>Weighted Intensity Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fifteen Lowest Ranking Countries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cameroon</td>
<td>2.12</td>
<td>Finland</td>
<td>1.26</td>
</tr>
<tr>
<td>Guyana</td>
<td>2.27</td>
<td>Australia</td>
<td>1.55</td>
</tr>
<tr>
<td>Laos</td>
<td>2.69</td>
<td>Laos</td>
<td>2.37</td>
</tr>
<tr>
<td>Finland</td>
<td>2.98</td>
<td>Switzerland</td>
<td>2.41</td>
</tr>
<tr>
<td>Gabon</td>
<td>3.08</td>
<td>Canada</td>
<td>2.96</td>
</tr>
<tr>
<td>Botswana</td>
<td>3.09</td>
<td>Guyana</td>
<td>3.26</td>
</tr>
<tr>
<td>United Arab Emirates</td>
<td>3.82</td>
<td>Trinidad and Tobago</td>
<td>3.79</td>
</tr>
<tr>
<td>Slovenia</td>
<td>3.83</td>
<td>Japan</td>
<td>4.14</td>
</tr>
<tr>
<td>Solomon Islands</td>
<td>4.47</td>
<td>Netherlands</td>
<td>4.32</td>
</tr>
<tr>
<td>Malawi</td>
<td>4.73</td>
<td>Norway</td>
<td>4.93</td>
</tr>
<tr>
<td>Qatar</td>
<td>4.76</td>
<td>Denmark</td>
<td>5.09</td>
</tr>
<tr>
<td>Mongolia</td>
<td>5.21</td>
<td>Hungary</td>
<td>8.11</td>
</tr>
<tr>
<td>Djibouti</td>
<td>5.42</td>
<td>Jamaica</td>
<td>8.43</td>
</tr>
<tr>
<td>Trinidad and Tobago</td>
<td>5.49</td>
<td>Benin</td>
<td>9.57</td>
</tr>
<tr>
<td>Fifteen Highest Ranking Countries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guatemala</td>
<td>656.29</td>
<td>Central African Rep</td>
<td>556.78</td>
</tr>
<tr>
<td>Poland</td>
<td>672.28</td>
<td>Iraq</td>
<td>675.62</td>
</tr>
<tr>
<td>Pakistan</td>
<td>685.12</td>
<td>Lebanon</td>
<td>682.01</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>798.12</td>
<td>Somalia</td>
<td>710.63</td>
</tr>
<tr>
<td>India</td>
<td>905.92</td>
<td>Liberia</td>
<td>777.27</td>
</tr>
<tr>
<td>Angola</td>
<td>956.89</td>
<td>Sri Lanka</td>
<td>1306.94</td>
</tr>
<tr>
<td>Rwanda</td>
<td>960.18</td>
<td>Congo</td>
<td>1581.57</td>
</tr>
<tr>
<td>China</td>
<td>1040.75</td>
<td>Rwanda</td>
<td>1718.66</td>
</tr>
<tr>
<td>South Africa</td>
<td>1081.22</td>
<td>Uganda</td>
<td>1743.28</td>
</tr>
<tr>
<td>Russia (Soviet Union)</td>
<td>1285.81</td>
<td>Guinea</td>
<td>2330.08</td>
</tr>
<tr>
<td>Israel</td>
<td>1391.67</td>
<td>Guatemala</td>
<td>3707.66</td>
</tr>
<tr>
<td>Lebanon</td>
<td>1502.14</td>
<td>Angola</td>
<td>4668.58</td>
</tr>
<tr>
<td>Iraq</td>
<td>1628.08</td>
<td>Sierra Leone</td>
<td>6073.01</td>
</tr>
<tr>
<td>United States</td>
<td>2625.38</td>
<td>Burundi</td>
<td>12324.14</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>5203.96</td>
<td>Afghanistan</td>
<td>14133.31</td>
</tr>
</tbody>
</table>

Comparing the least strife-torn countries also generates insights. The unweighted list is composed almost exclusively of small, rather obscure countries that are not of intrinsic interest to the Times’ core audience. In contrast, the weighted list includes more prominent countries (Australia, Canada, Switzerland, Japan, Norway, Netherlands, Denmark, etc.) many of whom have a reputation as stable, open societies. Thus, the
data in Table 1 suggests that SPEED’s weighting scheme has considerable face validity in mitigating the impact of country bias. Moreover, using a similar approach, weighting schemes could be developed to reduce within country media biases across spatial units (e.g., cities, provinces, regions, etc.).

**Summary and Conclusions**

Media scholars have conducted an on-going debate over the utility of data derived from news sources for over forty years, but Information Age developments have raised the stakes in this debate considerably. These developments have led to unprecedented increases in the availability of digitized news documents and have greatly enhanced our capacity to analyze the unstructured data embedded in news documents and transform them into structured data that can be analyzed by conventional statistical methods. These developments have restructured the debate over media data and have laid the groundwork for third-generation research projects employing them.

In this paper we re-examined the debate over media data in light of these developments. We summarized the key assertions in the debate and argued that they identify three problems with media data: a comprehensiveness problem, an identification problem, and a distortion problem. We then decomposed each of these problems and assessed their implications for the utility of media data in social science research, with a focus on civil strife research. We also discussed the potential for the remediation of the problems that posed serious threats to the utility of media, with an emphasis on solutions available to third-generation research efforts such as SPEED.

Our analysis makes it clear that some problems with media data are irremediable even in a field of research such as civil strife, which is highly amenable to research using media data. Perhaps the most serious irremediable problem is the relative lack of attention that the media pay to low intensity strife events. It is equally clear, however, that this problem is limited in its impact and it has only minimal effects in some of the most important civil strife applications. Most of the other issues identified can be addressed, though not wholly eliminated, by various tools and research procedures that are available to third-generation media data projects.

These third-generation tools include such things as large digitized news archives constructed from multiple sources, refined and highly automated identification procedures, and powerful data management systems that provide for unsupervised preprocessing and full-text searches of news documents. Enhancing the impact of these tools are advances in data collection and analyses procedures that can be used in processing media data. These advances include such things as powerful cyberinfrastructures with embedded NLP capabilities, high-tech training and testing procedures, approaches that maximize the extraction of relevant information (e.g., annotate relevant text, code recapitulation passages, capture multiple events), missing data procedures that capitalize on contextual information, and weighting schemes that address spatial and temporal biases in media coverage. Taking advantage of these advances will greatly enhance the capacity of researchers to capitalize on a truly unique source of information. They provide third-generation media data projects with the potential to draw on a greater variety of news sources over a longer period of time, cover a greater geographic expanse, and extract more information from them. Thus, to the extent that they are committed to employing these advances, social scientists should not shy away from the use of media data.
REFERENCES


