Analyzing International Event Data: A Handbook of Computer-Based Techniques

Philip A. Schrodt
and
Deborah J. Gerner

Draft book manuscript for the
Cambridge University Press

October, 2000

Table of Contents

Preface

Chapter 1: International Event Data

Chapter 2: Fundamentals of Machine Coding

Chapter 3: Statistical Characteristics of Event Data

Chapter 4: Clustering Methods

Chapter 5: Sequence Analysis Methods

Chapter 6: Hidden Markov Models

Chapter 7: Conclusion

Bibliography

Appendix: Event Data Coding Systems

Preface

Among the unusual features of the *Spruce Goose*, the giant plywood aircraft constructed by Howard Hughes in the 1940s, was the fact that its engines were designed to be accessible for maintenance while the aircraft was flying. With in-air refueling and sufficient spare parts, it could presumably stay aloft for a very long time, though in fact the plane flew only once.

Documenting the coding and analytical techniques developed in the Kansas Event Data System (KEDS) project has been rather like trying to repair the engines while the plane is flying. Unlike the *Spruce Goose*, the problem was not that the plane only flew once, but rather that it never seemed to land. While the various facets of the project were continually documented in a series of conference presentations, articles appearing in political science journals and chapters in edited volumes, it never clearly stopped at a point that shouted “Final Report.” In fact it still hasn’t: even as this preface is being written we are considering embarking on a new machine-coding system utilizing full-parsing capabilities. However, rather than putting off writing indefinitely, the start of the new millenium appeared as good a place as any to produce a book-length synopsis of our efforts.

The Development of KEDS

The heart of the KEDS project is an eponymous computer program written in Pascal[1] that contains about 16,000 lines of code, and will run in about 1.2 Mb of memory under the Apple Macintosh operating system. Much of that code deals with the interface; the core routines handling parsing and coding involve only about 3000 lines of code.

In order to understand the approach and limitations of KEDS, some history is appropriate. The roots of KEDS are found in Schrodt and Liebsohn's (1985) demonstration that a relatively short list of verbs—in their study these were derived using a simple machine-learning

[1] This decision made sense at the time because the Macintosh was originally a Pascal-oriented system. If we were to rewrite the system today, we would use C or C++.

algorithm—could be employed to assign WEIS event codes to the abbreviated natural-language summaries found in the ICPSR version of the WEIS data set with 70% to 85% accuracy. This exercise—combined with the NSF-sponsored “Data Development in International Relations” (DDIR) event data project and the discovery that Reuters was available on NEXIS—begat a short-lived computer program named WINR that served as a "proof of concept" for DDIR (Schrodt and Donald 1990).[2] WINR used statistical text indexing methods (see Salton 1989) to code events from the short English-language descriptions found in an index of the Foreign Broadcast Information Service (FBIS); the ICPSR WEIS set; and the longer event descriptions found in an extension of the WEIS data set coded in the early 1990s by the International Public Policy Research Center (IPPRC). Using an iterative technique, WINR attained about 65% coding accuracy.

The development of KEDS, with its dictionaries constructed by human coders, came with the recognition that statistical indexing methods—which were originally developed to classify documents containing a large amount of text—would not work very well on the relatively short lead sentences found in Reuters. An initial test of a system using a set of 500 actor and verb phrases derived from NEXIS leads achieved an 80% accuracy rate in coding the IPPRC WEIS descriptions, and Gerner (1990) demonstrated that KEDS-coded data generated from Reuters leads provided sufficient detail to differentiate Israeli policy shifts in response to criticism from Arab versus non-Arab sources during the first two years of the Palestinian intifada. Because Israel had been criticized from the left as being insensitive to criticism and from the right as being overly sensitive, this analysis was nontrivial.

The motivation behind developing the KEDS program was pragmatic: we needed the data! By 1990 publicly available event data sets such as WEIS and COPDAB were more than a decade

[2] WINR is a recursive acronym for “WINR Is Not Relatus.” Relatus was an artificial-intelligence system developed at MIT that was capable of constructing networks of political assertions from declarative texts constructed within a limited vocabulary and grammatical structure. The “recursive acronym” was made famous by the GNU Project—“GNU’s Not Unix.” It all seemed quite cute at the time.

out of date (WEIS was being maintained privately by Tomlinson, but we were not aware of this at the time) and virtually irrelevant to the topic we wanted to study, the Israeli-Palestinian dispute. Based on some earlier consulting experience on a privately-funded Middle East event data collection, we knew that human coding would not be able to keep pace with the events being generated by the Palestinian *intifada*, so it was machine coding or nothing. Necessity was in fact the mother of invention.

Table 1: History of KEDS Project

1985-1989	Development of event sequence methods using COPDAB, WEIS and BCOW data sets. Initial experiments with machine-coding WEIS summaries.
1990	WINR—machine learning of actor and verb phrases
1991, spring	KEDS parsing system with human-coded dictionaries for actors, verb phrases and stop-words; syntactic information used to identify actors and events; compound phrases. A parallel version was developed to code in German for the research reported in Gerner et al (1994).
1991, fall	Null codes, compound actors and verbs, pronoun dereferencing, subordinate phrase elimination
1992, spring	Agents (PANDA project) and compound actor codes
1993, spring	Phrase-specified issues (PANDA project)
1994, fall	Complexity filter, date-restricted actors
1995, spring	Content-analysis "profiles" (PANDA project)
1995, fall	Grammatical transformation rules, word classes and *grep*-like patterns
1996, spring	Pronoun forwarding across sentences
1995-1997	Development of cluster-based forecasting methods; KEDS_Count aggregation utility.
1998-1999	Development of hidden Markov model techniques; development of standard .verbs and .actors dictionaries; Actor_Filter utility.

Table 1 shows the subsequent development of the various features in KEDS.[3] The computer program followed a classical software development curve: By late 1991 it was producing usable event data but another two years of work were required to eliminate bugs (some of the file-destroying variety) and extend the feature set. We initially focused on the problem of coding Reuters lead sentences using the WEIS system, but for a time also worked on coding specialized foreign affairs chronologies, including a German-language chronology (we expected these would provide a denser event stream than Reuters, which did not turn out to be true; see Gerner et al, 1994). In 1992, we began a close electronic collaboration with the PANDA project at Harvard (Bond, Bennett and Vogele, 1994) that was critical in debugging the system and resulted in the addition of a number of features that provide for complex coding beyond the basic source-event-target framework of WEIS. PANDA's efforts at developing a global event data set also tested the program in ways that were different than our project at Kansas, which focused primarily on the Middle East. The development of the program was accompanied by extensive dictionary development work—we estimate about four person-years—at Kansas and Harvard.

A couple of points from this history are relevant to the design decisions made on KEDS. First, going into the project the prevailing academic opinion was that the task of fully automating machine coding was impossible—for example the DDIR event data project spent almost all of its funds on human-coded projects. This erroneous assessment was due to several factors, including

- failure to realize that event data coding was a much simpler problem than the general problem of natural language parsing;
- inexperience with Reuters, which proved to be a denser and more standardized source than the newspapers used in most of the earlier event data projects;

[3] Table 1 shows only the evolution of the linguistic features of the program; many additional changes were made in the interface and the internal structure, to say nothing of extensive debugging and some features that were experimentally developed but dropped from the final program.

- over-estimation of the accuracy of human-coded data, particularly the failure to recognize the tendency of fatigued human coders overlook some of the events in a complex sentence and the impossibility of maintaining consistency in a decades-long time series coded across multiple institutions and multiple human generations of coders and supervisors.

Developing a machine coding system that could replace human coding proved to be a much easier problem than anyone had originally anticipated. Throughout the exercise, simple rules applied to complex problems—for example KEDS's system for resolving pronoun references, an extraordinary difficult problem to solve generally—have tended to work quite well for the task at hand.

Because of this initial pessimistic assessment (and limited computing resources), KEDS was developed very conservatively, so that a minimum of effort was invested before results could be evaluated. This is most evident in two design decisions that form the core of the "sparse parsing" approach: KEDS assigns parts of speech only to the words that are required for coding (for example, it ignores most adjectives), and KEDS focuses only on the subject-verb-object structure of the sentence (for example it ignores—or tries to ignore—subordinate phrases). Following “Moore’s Law”, personal computer speed, memory and mass storage capacity have increased by more than a factor of ten since we began this work, and language models that are far more complex than that used in KEDS are now practical. KEDS should be viewed as demonstrating the minimum requirements needed in a system that can produce event data sets comparable in quality to those produced by human coders; it certainly does not represent the pinnacle of achievement in that field.

The machine coding elements of the KEDS project are sandwiched between two periods of work on political forecasting using sequence analysis methods. The first set of efforts, which primarily used COPDAB and BCOW data, was strongly influenced by the “artificial intelligence” work on computational modeling during the 1980s (Sylvan and Chan 1984, Cimbala 1987, Hudson 1991, Schrodt 1995). Following a hiatus of five or so years during which we developed the KEDS program and coding dictionaries, organized the 1995 International Studies Association meetings, and dealt with other distractions, we returned to the early warning problem, now focusing on the Levant, where we had spent considerable time doing field research. Our forecasting work during the 1990s focused first on statistical clustering methods, and more recently on hidden Markov models. These efforts are continuing.

Outline of the Book

This book deals with two broad topics: the creation of event data and the analysis of event sequences, with a focus on early warning problems. Throughout the book we have provided extensive examples and analyses of actual event data: most of these examples focus on KEDS-generated data for the Middle East, though some use other data sets.

Chapter 1 is a broad survey of development and use of event data in international relations research. It discusses the historical development of the technique, the rise-and-fall-and-rise of interest in event data in the policy community, and briefly summarizes a number of event data sets. This chapter also addresses issues of the validity of event data and general issues such as the effects of regional biases in news sources.

Chapter 2 looks specifically by the challenges of machine coding, focusing on the sparse-parsing approach employed in KEDS. Most of the problems discussed here will be relevant to any machine coding system, although some may be solved in future implementations. This chapter compares human and machine-coded data, and also looks at some problems that are intrinsic to news reports, with a focus on Reuters.

Chapter 3 is a technical chapter that looks at a number of general statistical issues involving event data. These include the various sources of error that can contaminate data—coding error is only one!—and suggest some techniques that could be used to reduce this. We discuss the issues of scaling and aggregation, and report experiments that indicate event data may be relatively insensitive to scaling decisions. Finally, we deal with the general problem of statistical early warning models in political analysis

The second half of the book discusses various forecasting and classification methods. Chapter 4 deals with statistical clustering techniques: while these were actually a later development in our own work, the techniques used are likely to be more familiar are to most political analysts familiar with quantitative methods, and thus are presented first. Some of these methods can be implemented using standard statistical packages such as SAS, SPSS and Stata.

The downside of clustering is that events must be converted to numerical values and aggregated before analysis. This restriction is relaxed in Chapters 5 and 6, which present techniques that can work directly with sequences of discrete events. These methods, adapted from work on speech recognition and DNA sequence analysis, require customized programs and are likely to be less familiar to individuals who have are accustomed to working with interval-level data. We nonetheless hope to demonstrate that they are promising approaches for forecasting and monitoring political behavior. Chapter 5 deals with deterministic sequence comparison methods; Chapter 6 discusses the hidden Markov model approach.

Chapter 7 summarizes the progress that has been in event data analysis during the 1990s, with an emphasis on current trends in computational power and data availability that are likely to make the technique even more useful in the next decade. We conclude with three “grand challenges” that go well beyond today’s applications of event data: identification of policy changes from behavioral changes, the development of automated chronology generators, and the induction of organizational rules of natural language sources. Appendices to this volume show the event coding schemes of four systems—BCOW, COPDAB, IDEA, and WEIS—and current sources of event data sets.

Acknowledgements

This project has benefited from a number of people along the way. Shannon Davis, Christopher Donald, Phillip Huxtable, Jon Pevehouse, Julia Pitner, Anthony Nownes, Erin Simpson, Scott Savaiano, Uwe Reising, and Judith Weddle were extensively involved in the vocabulary development and testing of the KEDS program at the University of Kansas; Brad Bennett, Doug Bond, Joe Bond, and William Vogele performed a similar role at the PANDA project at Harvard. We are particularly grateful to the PANDA project for helping us demonstrate that academic research need not be a zero-sum game. At various points in the

project we have also benefited from the advice and experience of Hayward Alker, Paul Diehl, Joshua Goldstein, Harold Guetzkow, Ted Gurr, Michael Haxton, Richard Herrmann, Craig Jenkins, Heinz Krummenacher, Edward Laurance, David Leibsohn, Renée Marlin-Bennett, Richard Merritt, Will Moore, Mohan Penubarti, Susanne Schmeidl, Fritz Snyder, Donald Sylvan, Charles Taylor, Robert Trappl, Michael Young, and Dina Zinnes, as well as our Kansas "Beer and Politics" colleagues Ryan Beasley, Paul D'Anieri, Ronald Francisco, Thomas Heilke, Juliet Kaarbo, Jarek Piekalkacweiz, Gary Reich, and Leo Villalón. Raja Abu-Jabar provided expert assistance in copy-editing drafts of the manuscript and helping bring it into final form.

Our thanks to Rodney Tomlinson for providing the 1982-1991 WEIS data for use in validating KEDS against human-coded data (Chapter 2). The Behavioral Correlates of War data set used in Chapters 5 and 6 was originally collected by Russell J. Leng and obtained through the Inter-university Consortium for Political and Social Research (ICPSR). Neither the original collector nor the ICPSR bear any responsibility for the analyses or interpretations presented here.

This research was supported in part by the National Science Foundation through grants SES89-10738, SES90-25130 (Data Development in International Relations Project), and SBR-9410023, and by several grants from the University of Kansas General Research Fund Grant 3500-X0-0038. The hidden Markov model estimation in Chapter 6 was partially supported by an EPSCoR start-up grant from the National Computational Science Alliance and utilized the NCSA SGI/CRAY Origin2000

Dedication

To the peoples of Egypt, Israel, Jordan,
Lebanon, Palestine, and Syria,
in the hope that the next twenty years of events
may be more peaceful than those
we have analyzed.

Obtaining the software and data

Most of the software and data sets discussed in this volume are available on the KEDS web site: `http://www.ukans.edu/~keds`. At the time of this writing, we are continuing to maintain several of our data sets in near-real-time (they are typically updated quarterly); these include almost all of the data used to generate figures in this volume. Several of these data sets are also archived in the Inter-University Consortium for Political and Social Research in Ann Arbor.

We are also engaged in an on-going effort to make our software more accessible, for example shifting from Pascal to C/C++, shifting from the Macintosh to Linux as our reference platform, and improving the internal documentation of the code. The source code for the better-documented programs in available on the web site under GNU General Public License provisions; we will gladly provide other source code, as is, to interested researchers. The Pascal source code for KEDS itself is available on request, although the data structures in the program are rather dated and the internal documentation is less than thorough.

Most of the other data sets mentioned here are also available through the ICPSR or through the individual researchers. In some cases—for example CREON and the original COPDAB—the data collection efforts ended a number of years ago, and the final version of the data are available through the ICPSR. In other cases, the ICPSR's collections have been superceded by other projects that, as of this writing, maintain their own web sites—this is true of CASCON, and to an extent of COPDAB, which is being carried forward (with a ten-year interuption in the time series) by the GEDS project. Finally, there are several data sets that have circulated for a number of years among the international relations research community, but have not been archived—these include the 1969-1992 extension of WEIS, the extensions of the SHERFACS data set, SAFED, and PANDA.

Appendix 2 gives a list of the URLs for all of the event data sets that we know to be accessible through the World Wide Web at the time of this writing. This includes data at the

ICPSR: access to these data sets requires institutional membership in the ICPSR or other arrangements; the inconvenience of this is more than offset by the knowledge that ICPSR data is systematically archived and will not disappear. Because URLs are notoriously transient, we intend to keep an up-to-date list of these links at the KEDS web site.

List of Acronyms

AFP	Agence France Presse (news service)
AP	Associated Press (news service)
BBC	British Broadcasting Corporation (news service)
BCOW	Behavioral Correlates of War (event data set)
CASCON	Computer-Aided System for the Analysis of Local Conflicts (event data set)
CD$_t$	Cluster density (sequence analysis method)
CD-ROM	Compact disk – read-only memory (data storage method)
COPDAB	Conflict and Peace Data Bank (event data set)
CREON	Comparative Research on the Events of Nations (event data set)
DARPA	Defense Advanced Research Projects Agency (United States government agency)
DDIR	Data Development in International Relations (data development project)
FBIS	Foreign Broadcast Information Service (US government news service)
GA	Genetic algorithm (sequence analysis method)
HMM	Hidden Markov model (sequence analysis method)
ICPSR	Inter-University Consortium for Political and Social Research (data archive located at the University of Michigan, Ann Arbor)
IDEA	Integrated Data for Events Analysis (event coding scheme)
IGO	Inter-governmental organization (political actor)
IPPRC	International Public Policy Research Center (consulting group)
GEDS	Global Event Data System (event data set)
KEDS	Kansas Event Data System (computer program)
KWIC	Keyword-in-context (data display technique)
LML$_t$	Lead minus lag (sequence analysis method)

MUC	Message Understanding Conference (a DARPA-sponsored research project on NLP)
NEXIS™	An electronic data service
NGO	Non-governmental organization (political actor)
NLP	natural language processing (a subfield in computer science and linguistics)
NSF	National Science Foundation (United States government agency)
PANDA	Protocol for the Analysis of Nonviolent Direct Action (event data set)
PLO	Palestine Liberation Organization (political actor)
PRC	People’s Republic of China (political actor)
SAFED	Southern Africa Event Data (event data set)
SHERFACS	Frank Sherman’s extension of the FACS data set (event data set)
SVO	subject-verb-object (standard syntactic order of English sentences)
U.N.	United Nations (political actor)
UPI	United Press International (news service)
U.S.	United States (political actor)
U.S.S.R.	Union of Soviet Socialist Republics (political actor)
VRA	Virtual Research Associates (consulting company)
WEIS	World Events Interaction Survey (event data set)

Actor Abbreviations Used in KEDS Project

ISR	Israel
IRN	Iran
IRQ	Iran
JOR	Jordan
LEB	Lebanon
KUW	Kuwait
PAL	Palestinians
SAU	Saudi Arabia
SYR	Syria
UAE	United Arab Emirates
UAR	Egypt
USA	United States of America
USR	Union of Soviet Socialist Republics