Pattern Recognition of International Crises using Hidden Markov Models

Philip A. Schrodt
Department of Political Science
University of Kansas
Lawrence, KS 66045 USA
phone: 913-864-3523 fax: 913-864-5700
p-schrodt@ukans.edu

Chapter contributed to
Diana Richards, ed.
Non-linear Models and Methods in Political Science
Ann Arbor: University of Michigan Press (forthcoming 1998)

June 1997

This research was funded in part by the National Science Foundation through grant SBR-9410023. The Behavioral Correlates of War data utilized in this paper were originally collected by Russell J. Leng and were made available by the Inter-university Consortium for Political and Social Research. Neither the original collector nor the Consortium bear any responsibility for the analyses or interpretations presented here. The computer source code and WEIS-coded Middle East event data set used in this paper can be downloaded from the Kansas Event Data System web site: http://www.ukans.edu/~keds. An earlier version of this paper was presented at the Synergy in Early Warning Conference, Centre for Refugee Studies, York University, March 1997.

Abstract

Event data are one of the most widely used indicators in quantitative international relations research. To date, most of the models using event data have constructed numerical indicators based on the characteristics of the events measured in isolation and then aggregated. An alternative approach is to use quantitative pattern recognition techniques to compare an existing sequence of behaviors to a set of similar historical cases. This has much in common with human reasoning by historical analogy while providing the advantages of systematic and replicable analysis possible using machine-coded event data and statistical models. This chapter uses "hidden Markov models"—a recently developed sequence-comparison technique widely used in computational speech recognition—to measure similarities among international crises. The models are first estimated using the Behavioral Correlates of War data set of historical crises, then applied to an event data set covering political behavior in the contemporary Middle East for the period April 1979 through February 1997. A split-sample test of the hidden Markov models perfectly differentiates crises involving war from those not involving war in the cases used to estimate the models. The models also provide a high level of discrimination in a set of test cases not used in the estimated, and most of the erroneously-classified cases have plausible distinguishing features. The difference between the war and nonwar models also correlates significantly with a scaled measure of conflict in the contemporary Middle East. This suggests that hidden Markov models could be used to develop conflict measures based on event similarities to historical conflicts rather than on aggregated event scores.

Introduction

Event sequences are a key element in human reasoning about international events. Human analysts "understand" an international situation when they recognize sequences of political activity corresponding to those observed in the past. Empirical and anecdotal evidence point to the likelihood that humans have available in long-term associative memory a set of "templates" for common sequences of actions that can occur in the international system (and in social situations generally). When part of a sequence is matched, the analyst predicts that the remainder of the sequence will be carried out *ceteris paribus,* though often the analyst will make a prediction for the express purpose of insuring that the remainder of the sequence is *not* carried out. Sequences can be successfully matched by human analysts in the presence of noise and incomplete information, and can also be used to infer events that are not directly observed but which are necessary prerequisites for events that have been observed.

The use analogy or "precedent-based reasoning" has been advocated as a key cognitive mechanism in the analysis of international politics by Alker (1987), Mefford (1985, 1991) and others, and is substantially different from the statistical, dynamic and rational choice paradigms that characterize most contemporary quantitative models of international behavior. Khong (1992) and Vertzberger (1990) review the general arguments in the cognitive psychology literature on use of analogy in political reasoning; May (1973) and Neustadt & May (1986) discuss it from a more pragmatic and policy-oriented perspective. As Khong observes:

> Simply stated, ... analogies are cognitive devices that "help" policymakers perform six diagnostic tasks central to political decision-making. Analogies (1) help define the nature of the situation confronting the policymaker; (2) help assess the stakes, and (3) provide prescriptions. They help evaluate alternative options by (4) predicting their chances of success, (5) evaluating their moral rightness and (6) warning about the dangers associated with options. (pg. 10)

The ubiquitousness of analogical reasoning is supported by a plethora of experimental studies in cognitive psychology in addition to the case studies from the foreign policy literature.[1]

For a human decision-maker, analogical reasoning is a form of bounded rationality because "associative recall" is an easy task for the human brain. In particular, associative recall is substantially easier for the human brain than sequential or deductive reasoning. Most experimental evidence suggests that human memory is organized associatively so that when one item is recalled, this naturally activates links to other items that have features in common, and these are more likely to be recalled as well (Anderson 1983; Kohonen 1984)

For example, few readers of this volume would have difficulty answering the question "Name two major conservative political leaders from the Western United States in the post-WWII period": An answer probably comes to mind, "from nowhere", in about a second.[2] Most readers can also determine the answer to the question "What are the prime factors of 9,699,690?" but working this out takes considerably more time and effort, and is anything but automatic. This occurs despite the fact that the information required to solve the second problem is substantially less than that required to solve the first. Associative recall is fast and easy; deductive reasoning is slow and hard. Millions of people spend hours of leisure time watching the television shows

[1] This is in distinct contrast to experimental work on the decision-making mechanisms postulated by most "rational choice" theories current in political science, where experimental support is almost nonexistent. Instead, hundreds of tests show that human decision-making, whether casual or expert, is dominated by techniques largely incompatible with rational choice assumptions (see for example Kahneman, Slovic & Tversky 1982; Hogarth & Reder 1987; Green & Shapiro 1994). This discordance between theory and experiment is extraordinarily inconvenient given the substantial intellectual investment made by political scientists in the rational choice paradigm, but—to the extent that political science intends to remain a positive rather than an abstract discipline—the problem will have to be confronted sooner or later.

[2] A more abstract example I've used in classes is "Name a state shaped like a kidney bean." Rarely does more than a second or two pass before someone comes up with an answer, and it is always "New Jersey." It is highly unlikely that the students have ever thought about this question before, so the answer could not result from memorization.

Solutions coming "from nowhere" means that the processing is *subcognitive*—the brain is working on the solution without one being consciously aware of how that processing is done. Sequential processing such as that involved in solving arithmetic problems, in contrast, is conscious—we are [sometimes painfully] aware of each discrete step in the reasoning. Far from being a quasi-mystical experience, subcognitive (or "intuitive") processing is quite common: For example a fluent speaker constructs grammatically correct sentences subcognitively, whereas a beginning speaker must do this sequentially through the conscious application of memorized rules.

"Jeopardy" and "Wheel of Fortune"—both games of associative recall—but the mathematics olympiads do not attract such an audience.

When one attempts to solve these two problems on a digital computer, however, the comparative advantage shifts because computer memory is typically organized sequentially without regard to content.[3] On a computer, the second problem can be solved with an elementary set of operations—in fact it is one of the first problems typically assigned to beginning programmers—whereas the first problem is virtually impossible unless one has a database already set up to answer it. Change the first problem slightly—"Name two Native American political leaders from the Western United States" or "Name two conservative political leaders from Western Canada"—and an entirely different database would be required by the computer. In contrast, a slight change in the second problem—"Find the highest common denominator of 9,699,690 and 418, 209"—would require only slight changes in the program used to solve it.

Because analogies are so prevalent in human political reasoning, it would be helpful to have some computational method of determining them . That, in turn, requires determining some means of ascertaining the general characteristics of a set of sequences. In Schrodt (1991), I posed this problem in the following manner:

> In human pattern recognition, we have a general idea of what a category of event sequences look like—the archetypal war, the archetypal coup, and so forth—and probably match to these ideals rather than to clusters of sequences. In a sense, ideal sequences are the centroid of a cluster of sequences, but that centroid is a sequence rather than a point. If a method could be found for constructing such a sequence, the cluster could be represented by the single ideal sequence, which would substantially reduce computing time and provide some theoretical insights as to the distinguishing characteristics of a cluster. (pg. 186)

[3] Some experimental exceptions to this exist, particularly in multi-processor systems designed for specific natural language processing tasks. Not coincidentally, the first widely-used computer language to provide content-addressable arrays—PERL—is also very popular for natural language tasks. PERL, however, simply simulates content-addressable memory by using search algorithms in a physical machine memory that is sequential.

Neural networks are the most common technique for implementing associative recall in digital computers, and show many of the same characteristics as human recall, such as an insensitivity to noise and missing values. While these methods work quite effectively on many cross-sectional problems, they are less effective in time series (see Weigand & Gershenfeld 1994).

The problem of generalizing sequences is particularly salient to the analysis of international political behavior in the late 20th century because, due to current changes in the international system, many contemporary situations do not have exact historical analogs. Yet human analysts are clearly capable of making analogies based on some characteristics of those behaviors. For example, because of its unusual historical circumstances, the situation in Zaire in 1997 had a number of unique characteristics, but during the crisis analysts pieced together sufficient similarities to a variety of earlier crises in Africa and elsewhere to come to the correct conclusion that Zaire had entered a period of rapid political change. The key to this analysis, however, was the ability to use general analogies: if one insisted on an analogy to a single case—which a human analyst would almost never do, but a computer might—then the Zairian case would be nearly impossible to analyze using analogies.

If a generalized event sequence is something concrete and objectively describable, as opposed to a warm fuzzy associative-recall feeling of "I'm sure I've seen this before...", it should be possible to find models and algorithms that can characterize those sequences. Such is the motivation of this paper, which demonstrates the use of a sequence recognition technique—hidden Markov models—for differentiating crises in the Behavior Correlates of War (BCOW: Leng 1987) event data set, then applies those models to a contemporary event data set on the Middle East. I demonstrate that these models are usually sufficient to discriminate BCOW crises that involved war from those that did not using the same split-sample design employed in Schrodt (1990, 1991). Models based on the BCOW data are then used to study interactions in three dyads in the Levant—Israel>Palestinians, Israel>Lebanon and Syria>Lebanon—using a WEIS-coded event data set covering April 1979 to February 1997. Despite the very substantial differences between the BCOW and Levant data sets in terms of coding procedures, historical time period, and underlying political behavior, the models that were estimated on the BCOW data show highly significant correlations with the level of conflict found in the Levant data, indicating that the hidden Markov models are successfully generalizing at least some of the characteristics of that behavior.

Hidden Markov models

Techniques for comparing two sequences of discrete events—nominally-coded variables occurring over time—are poorly developed compared to the huge literature involving the study of interval-coded time series. Nonetheless, several methods are available, and the problem has received considerable attention in the past three decades because it is important in the problems of studying genetic sequences in DNA, and computer applications involving human speech recognition. Both of these problems have potentially large economic payoffs, which tends to correlate with the expenditure of research efforts. Until fairly recently, one of the most common techniques was the Levenshtein metric (see Kruskal 1983; Sankoff & Kruskall 1983); Schrodt (1991) uses this in a study of the BCOW crises. Other non-linear methods such as neural networks, genetic algorithms, and locating common subsets within the sequences (Bennett & Schrodt 1987; Schrodt 1990) have also been used.

Hidden Markov models (HMM) are a recently developed technique that is now widely used in the classification of noisy sequences into a set of discrete categories (or, equivalently, computing the probability that a given sequence was generated by a known model). While the most common applications of HMMs are found in speech recognition and comparing protein sequences, a recent search of the World Wide Web found applications in fields as divergent as modelling the control of cellular phone networks, computer recognition of American Sign Language and (of course) the timing of trading in financial markets. The standard reference on HMMs is Rabiner (1989), which contains a thorough discussion of the estimation techniques used with the models as well as setting forth a standard notation that is used in virtually all contemporary articles on the subject.

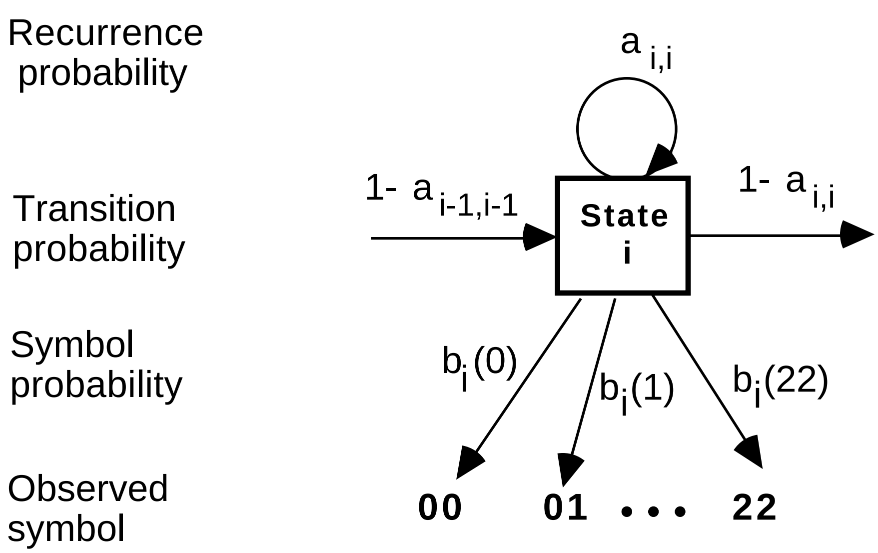
An HMM is a variation on the well-known Markov chain model, one of the most widely studied stochastic models of discrete events (Bartholomew 1975). Like a conventional Markov chain, a HMM consists of a set of discrete states and a matrix $A = \{a_{ij}\}$ of *transition probabilities* for going between those states. In addition, however, every state has a vector of *observed symbol probabilities*, $B = \{b_j(k)\}$ that corresponds to the probability that the system will produce a symbol of type k when it is in state j. The states of the HMM cannot be directly observed and can only be inferred from the observed symbols, hence the adjective "hidden".[4]

While the theory of HMM allows any type of transition matrix, the model that I will be testing is called a "left-right model" because it imposes the constraint that the system can only move in one direction, though it can remain in the existing state. The transition matrix is therefore of the form

$$\begin{pmatrix} a_{11} & 1-a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & 1-a_{22} & \ldots & 0 \\ 0 & 0 & a_{33} & \ldots & 0 \\ \ldots & & & & \ldots \\ 0 & 0 & 0 & \ldots & 1-a_{n-1,n-1} \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}$$

and the individual elements of the model look like those in Figure 1. This model is widely used in speech recognition because the pronunciation of a word moves in a single direction: parts of a word may be spoken slowly or quickly but in normal speech the ordering of those parts is never modified.

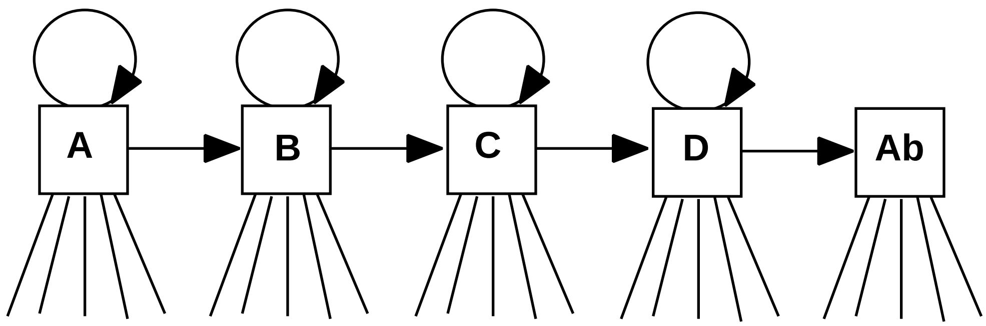
Figure 1. An element of a left-right hidden Markov model



[4] This is in contrast to most applications of Markov models in international politics where the states correspond directly to observable behaviors (see Schrodt 1985 for a review) .

A series of these individual elements form an HMM such as the 5-state model illustrated in Figure 2. Because of the left-right restriction, the final state of the chain is an "absorbing state" that has no exit probability and recurs with a probability of 1. The left-right restriction also means the transition matrix is completely determined by the "recurrence" probabilities a_{ii}.

Figure 2. A left-right hidden Markov Model



The Whitson and Meyers implementation—which is designed for experimenting with speech recognition systems—also includes a vector of symbol probabilities for each transition between states. This is relevant in the speech recognition problem because the shift from one part of a word to another is frequently signaled by a distinct change in sound. Transitions could also be important in political event sequences—for example the outbreak of hostilities changes the character of a crisis—although in political event data generated from a source such as Reuters, such a change is only rarely signaled by a single event.

In empirical applications, the transition matrix and symbol probabilities of an HMM are estimated using an iterative maximum likelihood technique called the Baum-Welch algorithm. This procedure takes a set of observed sequences (for example the word "seven" as pronounced by twenty different speakers, or a set of dyadic interactions from the BCOW crisis set) and finds values for the matrices A and B that locally maximize the probability of observing those sequences. The Baum-Welch algorithm is a nonlinear numerical technique and Rabiner (1989:265) notes "the algorithm leads to a local maxima only and, in most problems of interest, the optimization surface is very complex and has many local maxima."

Once a set of models has been estimated, it can be used to classify an unknown sequence by computing the maximum probability that each of the models generated the observed sequence. This is done using a dynamic programming algorithm that requires on the order of N^2T calculations, where N is the number of states in the model and T is the length of the sequence.[5] Once the probability of the sequence matching each of the models is known, the model with the highest probability is chosen as that which best represents the sequence. Matching a sequence of symbols such as those found in daily data on a six-month crisis coded with using the 22-category World Events Interaction Survey scheme (WEIS; McClelland 1976), generates probabilities on the order of $10^{-(T+1)}$—which is *extremely* small, even if the sequence was in fact generated by one of the models[6]—but the only important comparison is the *relative* fit of the various models. The measure of fit usually reported is the log of the likelihood; this statistic is labeled α (alpha).

For example, in a typical speech-recognition application such as the recognition of bank account numbers, a system would have HMMs for the numerals "zero" through "nine". When a speaker pronounces a single digit, the system converts this into a set of discrete sound categories (typically based on frequency), then computes the probability of that sequence being generated by each of the ten HMMs corresponding to the ten digits spoken in English. The HMM that has the highest likelihood—for example the HMM corresponding to the numeral "three"—gives the best estimate of the number that was spoken.[7]

The application of the HMM to the problem of generalizing the characteristics of international event sequences is straightforward. The symbol set consists of the event codes

5 Exhaustive enumeration of all of the ways that a model could generate a sequence, in contrast, would require on the order of $2TN^T$ calculations, which is prohibitively large for sequences of any practical length (Rabiner 1989: 262).

6 Assume that each state has ten associated WEIS categories that are equally probable: $b_i(k)=0.10$. Leaving aside the transition probabilities, each additional symbol will reduce the probability of the complete sequence by a factor of 10^{-1}. The transition probabilities, and the fact that the WEIS codes are not equiprobable, further reduce this probability.

An insurmountable disadvantage of this computation is that one cannot meaningfully compare the fit of two sequences to a single HMM unless the sequences are equal in length. In other words, it is possible to compare a sequence to a series of models, but one cannot compare several arbitrary sequences to a single model.

7 If none of the probabilities are higher than some threshold, the system could request that the speaker repeat the digit or transfer the call to a human operator.

taken from an event data set such as WEIS or BCOW. The states of the model are unobserved, but have a close theoretical analog in the concept of crisis "phase" that has been explicitly coded in data sets such as the Butterworth international dispute resolution dataset (Butterworth 1976), CASCON (Bloomfield & Moulton 1989, 1997) and SHERFACS (Sherman & Neack 1993), and in work on preventive diplomacy such as Lund (1996).[8] For example, Lund (1996:38-39) outlines a series of crisis phases ranging from "durable peace" to "war" and emphasizes the importance of an "unstable peace" phase. In the HMM, these different phases would be distinguished by different distributions of observed WEIS events. A "stable peace" would have a preponderance of cooperative events in the WEIS **01-10** range; the escalation phase of the crisis would be characterized by events in the **11-17** range (accusations, protests, denials, and threats), and a phase of active hostilities would show events in the **18-22** range. The length of time that a crisis spends in a particular phase would be proportional to the magnitude of the recurrence probability a_{ii}.

The HMM has several advantages over alternative models for sequence comparison. First, if N<<M, the structure of the model is relatively simple. For example a left-right model with N states and M symbols has 2(N-1) + N*M parameters compared to the M(M+2) parameters of a Levenshtein metric. HMMs can be estimated very quickly, in contrast to neural networks and genetic algorithms. While the resulting matrices are only a local solution—there is no guarantee that a matrix computed from a different random starting point might be quite different—local maximization is also true of most other techniques for analyzing sequences, and the computational efficiency of the Baum-Welch algorithm allows estimates to be made from a number of different starting points to increase the likelihood of finding a global maximum. The HMM model, being stochastic rather than deterministic, is specifically designed to deal with noisy output and with indeterminate time (see Allan 1980); both of these are present in international event sequences.

8 Sherman & Neack (1993) provide a review of the evolution of these data sets. Schrodt & Gerner (1997) demonstrate that distinct political phases—defined statistically using clusters of behavior—are found in event data sets covering the Middle East.

An important advantage of the HMM, particularly in terms of its possible acceptability in the policy community, is that it can be *trained by example*: a model that characterizes a set of sequences can be constructed without reference to the underlying rules used to code those sequences. This contrasts with the interval-level aggregative methods using event data scales such as those proposed by Azar & Sloan (1975) or Goldstein (1992). These scales, while of considerable utility, assign weights to individual events in isolation and make no distinction, for example, between an accusation that follows a violent event and an accusation during a meeting.[9] The HMM, in contrast, dispenses with the aggregation and scaling altogether—using only the original, disaggregated events—and models the relationship between events by using different symbol observation probabilities in different states.

In contrast to most existing work with event data—which usually deals with events aggregated by months or even years—the HMM requires no temporal aggregation. This is particularly important for early warning problems, where critical periods in the development of a crisis may occur over a week or even a day. Finally, indeterminate time means that the HMM is relatively insensitive to the delineation of the start of a sequence, which was frankly the biggest problem I had in my earlier work on this problem. It is simple to prefix an HMM with a "background" state that simply gives the distribution of events generated by a particular source (e.g. Reuters/WEIS) when no crisis is occurring and this occurs in the models estimated below. A model can simply cycle in this state until something important happens and the chain moves into later states characteristic of crisis behavior.

There is a clear interpretation to each of the parameters of the A and B matrices, which allows them to be interpreted substantively; this contrasts with techniques such as neural networks that have a very diffuse parameter structure. More generally, there is clear probabilistic interpretation of the model that uses familiar structures and concepts such as

[9] Mindful of these problems, Leng's BCOW coding scheme makes such distinctions, employing an elaborate set of codes and cross-references that place an event in the context of the crisis as a whole. Unfortunately, the sheer complexity of this coding makes the data difficult to analyze using conventional techniques, and as a consequence the information available in the BCOW data has probably not been fully exploited.

probability vectors, maximum likelihood estimates and the like. Finally—and not insignificantly—the technique has already been developed and is an active research topic in a number of different fields. The breadth of those applications also indicates that the method is relatively robust. While there is always a danger in applying the *technique du jour* to whatever data on political behavior happen to be laying around, the HMM appears unusually well suited to the problems of generalizing and classifying international event data sequences, a task for which there are at present no particularly satisfactory solutions.

Testing the Model

As is typical with machine learning protocols, the HMM will be evaluated using split-sample testing. Because the knowledge structures of many machine learning systems are quite large, they will frequently achieve 100% discrimination among their test cases,[10] and can be nontrivially tested only on data other than those on which they were trained. In a sense, this is a distinction between learning and memorization: If a system can only parrot back the discriminations found in its training set, this only demonstrates that the knowledge structure is sufficient to "memorize" those differences, not that general principles have been learned. In this respect machine learning studies apply a more difficult standard of empirical accuracy than that used in most statistical research, where all of the available data are typically used to estimate the parameters.

Data

The hidden Markov models were first estimated using the BCOW sequences studied in Schrodt (1990; 1991). The BCOW events were re-coded into WEIS categories according to the translation table listed in the Appendix. The four subsets of crises listed in Table 1 were

[10] The exception occurs when two cases have different classifications but have identical values for all of the classifying variables. In such situations insufficient information exists in the data set to make the discrimination. Compared to many machine learning systems, the left-right HMM involves relatively few parameters and will not necessarily achieve 100% discrimination—an example of this occurs below—but the split-sample protocol is still justified as a conservative means of testing the model.

analyzed.[11] The short names (e.g. *pastry*) correspond to the BCOW file identifiers. "Training" sequences were used to estimate the HMM matrices for the war and nonwar sequences; the system was tested with the remaining "test" sequences.

In contrast to the design in Schrodt (1990, 1991)—which distinguished with separate codes whether events were occurring between the principal actors in the conflict, the principals and outside actors, and so forth—this study looked at simple directed-dyadic sequences involving the principal actors ("Side A" and "Side B") identified in the BCOW data set. This was done to provide comparability with a general event stream such as one generated by Reuters, where the "sides" of a conflict are not necessarily evident. The HMMs are therefore trying to model the general characteristics of "dyads involved in a crisis" rather than making distinctions based on the role of various actors.

In order to record the passage of time in the various crises, days where nothing occurred were assigned a **00** non-event code; this is by far the most common "event" in the sequences. Sequences were coded from the beginning date to the ending date of the crisis as reported in the BCOW data set. When the BCOW data set reported multiple events on a single day, all of these were included. This is consistent with the structure of the hidden Markov model because the events observed on a particular day could occur as multiple observations from a single state of the model. In contrast, some of the other methods I've worked with (for example parallel event sequences and the Levenshtein metric) assume a strict temporal ordering. In those models, the fact that some days have multiple events while other days contain zero or one events complicates the estimation of the model. Dyads containing fewer than 20 BCOW events were not included in the analysis. Dyadic sequences typically contained about 30 to 70 actual events, though in a few

[11] The BCOW crises not included in the Schrodt (1990, 1991) studies were generally those whose length in events is very long (e.g. Suez or the Cuban Missile Crisis); or those that I could not easily classify into war or nonwar categories (e.g. Trieste). The HMM method is less sensitive to the length of a crisis than were the earlier methods I studied, so it should be possible to analyze the longer crises in a later test.

cases there were over 200 events. When the nonevent days were added, most of the sequences contained between 200 and 300 events.[12]

TABLE 1. Data Sets Analyzed

BCOW file	Crisis	Date
Crises without war, training set		
fashod	Fashoda Crisis	1898-1899
1stmor	First Moroccan Crisis	1904-1906
bosnia	Bosnian Crisis	1908-1909
2ndmor	Second Moroccan Crisis (Agadir)	1911
rhine	Rhineland Crisis	1936
Crises without war, test set		
pastry	Pastry War Crisis	1838-1839
brprt	British-Portuguese Crisis	1889-1890
anschl	Anschluss Crisis	1937-1938
munich	Munich Crisis	1938
berair	Berlin Blockade	1948-1949
Crises involving war, training set		
schles	Schleswig-Holstein War	1863-1864
spam	Spanish-American War	1897-1898
centam	Second Central American War	1906-1907
chaco	Chaco Dispute and War	1927-1932 (see note)
italet	Italo-Ethiopian War	1935-1936
Crises involving war, test set		
balkan	Balkan Wars	1912-1913
palest	Palestine War	1947-1948
kash1	First Kashmir War	1947-1949
kash2	Second Kashmir War	1964-1966
bangla	Bangladesh War	1971

Note: The .chaco data covers a number of military actions leading to the outbreak of war, but not the continuous military conflict from September 1932 to June 1935.

[12] The shortest sequences used were those in the *pastry* crisis—around 80 events—and the longest sequences were in *chaco*—around 1000.

The Levant data were machine-coded using the WEIS system from Reuters lead sentences obtained from the NEXIS data service for the period April 1979 through February 1997. These data were coded using the Kansas Event Data System (KEDS) machine-coding program (Gerner et al. 1994; Schrodt, Davis & Weddle 1994).[13] KEDS does some simple linguistic parsing of the news reports—for instance, it identifies the political actors, recognizes compound nouns and compound verb phrases, and determines the references of pronouns—and then employs a large set of verb patterns to determine the appropriate event code. Schrodt & Gerner (1994), Huxtable & Pevehouse (1996) and Bond et al. (1996) discuss extensively the reliability and validity of event data generated using Reuters and KEDS. The sequences that were tested were filtered of any of the WEIS codes that did not occur in the translated BCOW data (see Appendix) and a **00** nonevent was added for each day in which no events were recorded. As in the BCOW sequences, multiple events occurring in the same day are kept in the sequence.

Estimation Algorithm

The HMM was implemented by slightly modifying the source code written by Meyers & Whitson (1995). Their C++ code implements a left-right hidden Markov model and the corresponding Baum-Welch maximum likelihood training algorithm using the algorithms described by Rabiner (1989). I translated this code from the Solaris C++ environment to a Macintosh

[13] The NEXIS search command used to locate stories to be coded was

(ISRAEL! OR PLO OR PALEST! OR LEBAN! OR JORDAN! OR SYRIA! OR EGYPT!)
AND NOT (SOCCER! OR SPORT! OR OLYMPIC! OR TENNIS OR BASKETBALL)

Only the lead sentences were coded; this produced a total of 83,196 events.

In contrast to the data I have used in earlier papers (e.g. Schrodt & Gerner 1994, Schrodt & Gerner 1995), this data set was generated under the control of a "complexity filter" that did not code sentences if

- the sentence contained six or more verbs or
- no actor was found prior to the verb.

Sentences that met these criteria had a greater-than-average likelihood of being incorrectly coded by KEDS, thus by using the filter should result in somewhat less noisy data.

From spot-checking some of the more densely reported dyads (e.g. ISR>PAL and ISR>LEB), this new data set generally results in Goldstein scores that are smaller in magnitude. The bivariate regressions for these two dyads are

ISR>PAL	G96 = 0.73 G95 - 2.75	r = 0.93	N = 192
ISR>LEB	G96 = 0.71 G95 - 0.66	r = 0.88	N = 192

where G96 are the Goldstein scores for the data set used in this paper and G95 are the scores for the data set used in Schrodt & Gerner (1997). The overall patterns in the series are generally very similar between the two data sets.

CodeWarrior ANSI C environment,[14] in the process combining Meyers and Whitson's separate driver programs for training and testing into a single program, and modifying the input format to handle the BCOW and WEIS sequences. The source code for this program is available at the KEDS web site: http://www.ukans.edu/~keds.

The resulting program is very fast—estimation of the HMM matrices for about a dozen sequences using the Baum-Welch algorithm required less than a minute on a Power Macintosh 7100/80, and the computation of the probability of a sequence being generated by a particular HMM is nearly instantaneous. The program requires about 1.5 Mb of memory for a system using 23 codes, 12 states and 1000-event sequences. The largest arrays required by the program are proportional to (M+T)*N, where M is the number of possible event codes, T is the maximum sequence length and N is the number of states, so it would obviously be possible to substantially increase the complexity of the HMM beyond that studied in this paper without running into memory constraints on a contemporary personal computer.

Consistent with the CASCON and SHERFACS approaches, the models I estimated used 6 states. Some additional experiments were done using a 12-state model and this produced much the same results.[15] Adding additional states to the models would strain neither memory nor computing time but, as noted below, a small number of states seems to be sufficient for the BCOW crises. Because the Baum-Welch algorithm is a numerical estimation method that is dependent on the initial values assigned to the probabilities, I ran at least 512 experiments with the matrices initialized to different random sets of probabilities, and then selected the model that

[14] The choice of C over C++ was purely personal—I'm currently more comfortable working in the former language. The Meyers and Whitson code is clean, well-documented, and survived my translation to run correctly the first time. I would assume that either the C or C++ code would port easily to a DOS/Windows or OS/2 environment for those so inclined. The code posted on the web page does not implement the multiple initial matrices but this less-documented program is available from the author.

[15] The 12-state models resulted in about a 4% improvement in the total likelihood in both the war and nonwar training cases. The classification accuracy is generally similar to that of the 6-state model—including the cases which were misclassified—with 3 errors in the war test cases and 6 in the nonwar. Curiously, only 6 of the states in the nonwar model and 7 of the states in the war model have high (>0.85) recurrence probabilities (including the absorbing state), indicating that most of the remaining states do not contribute substantially to the likelihood of the model. While the original 6-state configuration was chosen to mirror the Butterworth and CASCON schema, it seems to be close to optimal on the basis of the empirical tests as well.

had the highest total probability for the cases in the training set. A spot-check of the best-fitting results generated by separate runs of 128 experiments showed an extremely high correlation (r>0.99) between the alpha probabilities computed for each of the training cases, so the algorithm appears to be finding a global maximum in terms of these.[16] There is less convergence between the probabilities in the A and B matrices, though these are generally similar. This is presumably due to the fact that various combinations of recurrence probabilities and observed symbol probabilities can produce almost identical likelihoods for the training sequences.

Results

Discriminating BCOW War and Nonwar Crises

The HMMs estimated from the nonwar and war BCOW crises (translated into WEIS codes) are reported in Table 2 and Figure 3; Table 2 also reports the events in the transition vectors that have relatively high probabilities. The matrices are quite plausible, as are the differences between them; both models generated large recurrence probabilities on all six states. Both of the models successfully match all of their training cases —in other words, all of the nonwar training cases show a higher likelihood of fitting the nonwar model than the war model, and vice versa for the war training cases. The HMM thus meets the minimal requirements of any machine-learning approach: it can successfully classify its training cases. Because the set of 83 parameters used in the model (5 recurrence probabilities and 6 vectors of 13 symbol probabilities) are substantially smaller than the several thousand events in the training sets, it is unlikely that this fit is tautological.

The nonwar matrix begins with a series of cooperative events in state A. As conjectured, the distribution of the probabilities in this vector is close to that of the vector of marginal probabilities of events in the training set: the two vectors correlate with r=0.95 for all true events,

[16] The difference between the best and worst fit among the experiments was around 3% of the value of sum of the probabilities: this difference is about 100 in the nonwar set and 200 in the war set. The min_delta_psum parameter in the program controls when the algorithm stops optimizing because the change in probabilities is too small: this was originally set at 0.01 but I increased it to 1.0 without any apparent degradation of the ability of the algorithm to find an optimum. The higher value results in a considerably faster program: the estimation using 512 experiments on the 6-state model requires about an hour on a Macintosh 7100/80.

and r=0.98 when the nonevent is included. The model then passes the time with nonevents in state B before escalating into conflictual events in state C. The transition between states B and C is likely to be either a consult, promise or request. State D is generates another sequence of nonevents, and then state E is dominated by just three event types: promise (probability 0.81), approve (probability 0.10) and agree (probability 0.08). State E rather conspicuously appears to represent the "dispute resolution" phase of the crisis. The absorbing state settles back into a mix of cooperative and conflictual (but nonviolent) events.

The war matrix shows a very different pattern. State A primarily generates nonevents, again closely reflecting the marginal probabilities of events in the training set: the correlation is r=0.82 for the true events and r=0.9995 when the nonevent is included.[17] State B involves a mix of mediating (consult, promise and request; total probability 0.37) and confrontational (accuse, demonstrate, seize and force; total probability 0.30) events. In state C, force has the highest probability. In contrast to the nonwar model, nonevents have high probabilities in the transition vector, indicating that the shift between states is signaled by a change in the distribution of events rather than a single triggering events. States D and E are dominated by nonevents and a mixture of conciliatory and confrontational events, and the absorbing state once is more dominated by force events. My guess is that states D and E are most likely the result of situations where the BCOW data include a period of peace negotiations following the cessation of hostilities, whereas the absorbing state is used to model the cases where hostilities continue until virtually the end of the data (specifically the Schleswig-Holstein War and Italo-Ethiopian War). The presence of force events in the transition vectors of states D and E is consistent with this interpretation and the recurrence probability on state E is so high (0.9946; for state D it is 0.9858) that it could virtually serve as an absorbing state itself.

The results of the split-sample testing are reported in Table 3, which gives the α log-likelihood values for the fit of various dyadic sequences using the HMMs estimated on the

[17] The ridiculously high value of r that results from inclusion of the nonevents is obviously due to the extremely skewed frequency distribution.

training cases. The war model classifies somewhat more accurately than the nonwar model, but both models do quite well and the cases that are incorrectly classified are concentrated in a set of plausible exceptions rather than distributed randomly.

Table 2a. Hidden Markov recurrence probabilities and event matrices: Nonwar Crises

	A	B	C	D	E	Abs
recurrence probability	0.96	0.98	0.96	0.99	0.64	1.00
Event						
00 none	0.58	0.97	0.33	0.97	0.00	0.85
01 comment	0.02	0.00	0.02	0.00	0.00	0.00
02 consult	0.07	0.003	0.04	0.00	0.00	0.04
04 approve	0.04	0.003	0.07	0.006	0.10	0.20
05 promise	0.14	0.006	0.17	0.003	0.81	0.04
06 grant	0.00	0.00	0.005	0.00	0.00	0.00
07 reward	0.002	0.00	0.00	0.00	0.00	0.00
08 agree	0.005	0.00	0.005	0.00	0.08	0.005
09 request	0.07	0.002	0.14	0.004	0.017	0.03
12 accuse	0.04	0.007	0.08	0.006	0.00	0.01
17 threaten	0.002	0.00	0.005	0.00	0.00	0.00
18 demons	0.01	0.004	0.11	0.01	0.00	0.004
19 reduce rel.	0.00	0.00	0.005	0.00	0.00	0.002
21 seize	0.005	0.003	0.005	0.00	0.00	0.00
22 force	0.00	0.002	0.005	0.001	0.00	0.002
transition events	03 (.23)	03 (.30) 05 (.33) 09 (.21)	03 (.20) 09 (.37)	00 (.57) 18 (.26)	00 (.36) 04 (.20) 05 (.22)	NA

Table 2b. Hidden Markov recurrence probabilities and event matrices: War Crises

	A	B	C	D	E	Abs
recurrence probability	0.99	0.97	0.95	0.99	0.99	1.00
Event						
00 none	0.94	0.29	0.40	0.70	0.89	0.08
01 comment	0.002	0.01	0.03	0.02	0.01	0.00
02 consult	0.002	0.00	0.00	0.00	0.00	0.00
04 approve	0.004	0.14	0.00	0.01	0.01	0.07
05 promise	0.003	0.03	0.00	0.01	0.00	0.00
06 grant	0.01	0.13	0.07	0.06	0.01	0.00
08 agree	0.00	0.003	0.01	0.01	0.01	0.00
09 request	0.01	0.10	0.07	0.02	0.01	0.00
12 accuse	0.01	0.09	0.01	0.03	0.003	0.00
17 threaten	0.00	0.006	0.00	0.003	0.00	0.00
18 demons	0.005	0.15	0.04	0.09	0.001	0.21
19 reduce rel.	0.00	0.01	0.02	0.01	0.01	0.00
21 seize	0.002	0.03	0.02	0.004	0.02	0.07
22 force	0.01	0.03	0.33	0.04	0.03	0.58
transition events	00 (.71)	00 (.39) 21 (.17)	00 (.46) 08 (.16)	00 (.44) 22 (.25)	08 (.30) 19 (.20) 22 (.26)	NA

Figure 3a. HMM Event Probabilities: Nonwar crises

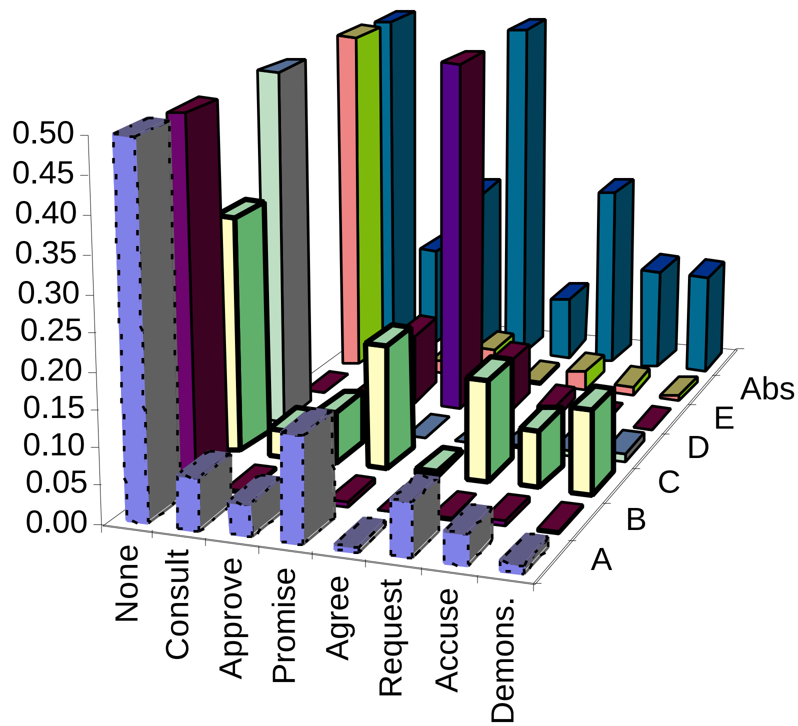


Figure 3b. HMM Event Probabilities: War crises

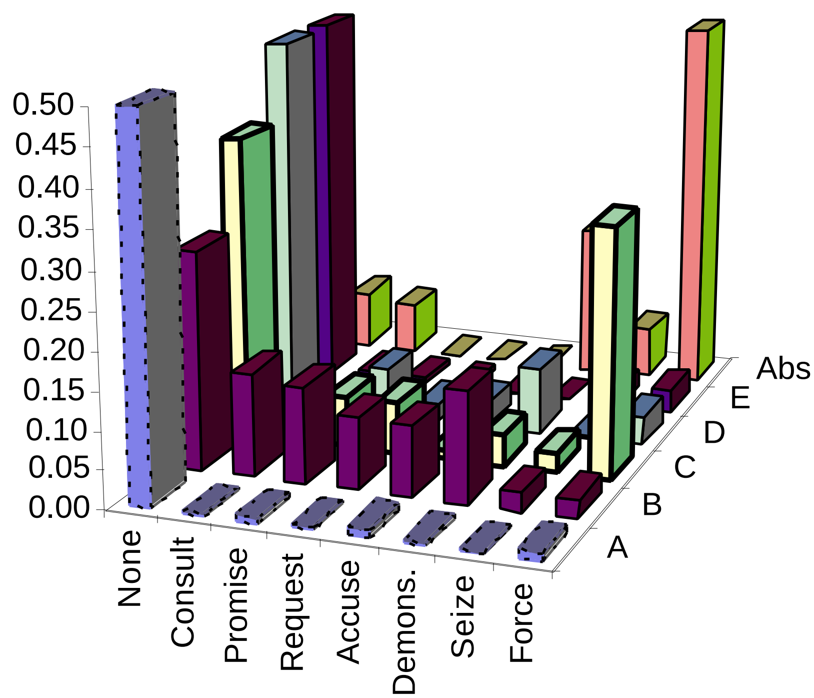


Table 3: Alpha values for the test cases

Nonwar Test Cases

BCOW crisis file	Dyad	Log-likelihoods nonwar HMM	war HMM	Correct?
.pastry	MEX > FRN	-104.2095	-119.8629	Y
	FRN > MEX	-109.8434	-119.6688	Y
.brprt	UK > POR	-164.1453	-164.2271	Y
	POR > UK	-181.8453	-176.2579	N
.anschl	AUS > GER	-167.3658	-184.7649	Y
	GER > AUS	-188.242	-221.9629	Y
.munich	CZE > GER	-393.079	-411.0417	Y
	GER > CZE	-376.0795	-355.9724	N
	UK > GER	-253.7782	-263.6895	Y
	GER > UK	-171.3611	-200.1183	Y
	FRN > GER	-222.8409	-211.1711	N
.berair	UK > USR	-244.2776	-240.3056	N
(Berlin	USR > UK	-167.5521	-165.2587	N
airlift)	USA > USR	-465.0612	-472.7058	Y
	USR > USA	-294.8895	-296.4974	Y
	USR > GER	-260.5101	-173.012	N

War Test Cases

BCOW crisis file	Dyad	Log-likelihoods nonwar HMM	war HMM	Correct?
.balkan	BUL > TUR	-199.4287	-154.2102	Y
	TUR > BUL	-134.231	-116.043	Y
	MTN > TUR	-135.3081	-122.4961	Y
	BKL > TUR	-154.9236	-170.3853	N
	TUR > BKL	-127.5491	-143.9149	N
	BUL > SER	-131.8183	-115.1773	Y
.palest	EGY > ISR	-179.272	-135.0227	Y
	ARL > ISR	-312.3664	-211.1503	Y
	ISR > ARL	-275.2968	-198.1442	Y
.kash1	IND > PAK	-610.1478	-556.1742	Y
	PAK > IND	-479.0293	-470.0874	Y
.kash2	IND > PAK	-588.8899	-443.0561	Y
	PAK > IND	-519.3982	-403.8226	Y
.bangla	IND > PAK	-500.4738	-376.3052	Y
	PAK > IND	-488.6324	-420.9545	Y
	BNG > PAK	-236.5325	-219.4431	Y
	PAK > BNG	-336.4198	-253.9302	Y

*BNG = Bangladesh; BKL = Balkan League; MTN = Montenegro; ARL = Arab League

All but two of the test dyads in the war set show a higher likelihood of being generated by the war model than by the nonwar model; the two cases where this is not true involve the Balkan League/Turkey dyad, a sequence that contains only a single use of force. For the war crises, 10 of the 16 test dyads have a higher probability of fitting the nonwar HMM than the war HMM, and half of the incorrect classifications occur in just one of the crises—the Berlin airlift. That crisis probably generates outliers because of the atypical number of reduce and seize events: there are 14 (0.69%) and 21 (1.03%) of these in the 2040 events in the .berair file. This proportion is much closer to that found in the war training set (0.53% and 1.10% of 6645 events) than in the nonwar training set (0.15% and 0.11% of 4590 events), so from the standpoint of the training sets, this crisis looks more like a war. The Munich crisis GER>CZE dyad concludes with a number of force events; arguably these events could be considered close to a war, particularly from the standpoint of Czechoslovakia.

Using the BCOW models to measure conflict in the Middle East

The second set of calculations was designed to determine whether the HMMs could be used to reveal anything about a contemporary political situation. The nonwar and war HMM models were first re-estimated using both the training and test cases.[18] Figures 4, 6 and 7 show the log-likelihood fit of the two models to three of the densest dyads from the Reuters-based Levant data set: ISR>PAL, SYR>LEB and ISR>LEB. The two lines below the X-axis are the alpha log-likelihoods; the line near the X-axis is the $\alpha_{war} - \alpha_{nonwar}$ difference. The WEIS sequences used to generate the fit were generated by taking the 100 events prior to the end of each month. This sequence typically covers about two months, though it is shorter in times of intense activity. Because all of the sequences are the same length, their values can be compared over time.

[18] In contrast to the earlier results, these models do not classify all of the training cases correctly: on the validation test, .berair USR > GER is incorrectly classified in the nonwar set; .balkan BKL > TUR, .balkan TUR > BLK, and .chaco PAR > BOL are incorrectly classified in the war set. All of these cases except .chaco were also problematic in the earlier tests. These erroneous distances are between 5% and 50% of magnitude of the distances in the correctly classified cases, so most of the errors are near misses.

Before discussing the results, it should be noted that this is a fairly audacious exercise because it is comparing two sets of data that have *nothing* in common other than the underlying political activity. The BCOW data deal with a set of crises that occurred as much as a century and a half before the Levant data set; and these were human-coded using a complex coding scheme from an assortment of historical documents. In contrast, the Levant dataset was machine-coded using simple source-event-target coding from a single source, Reuters. The political events recorded in the two data sets are themselves quite different, at least in my translation—in particular the translated BCOW is missing entirely some of the most frequent WEIS event categories in the Levant data: the accusations, denials and counter-accusations in WEIS categories 10 to 17. Finally, the only linkage between the two sets of behavior is found in the relatively tenuous HMM matrices.

The first thing that is conspicuous in the figures is that the nonwar and war alpha curves track each other very closely. This probably reflects the effects of the presence or absence of nonevents; these are much prevalent in the BCOW dyads than in these politically-active Levantine dyads. Periods with a high intensity of activity—for example the Palestinian *intifada* and various Syrian and Israeli interventions in Lebanon—consistently show much lower alpha values than periods of low activity. This reduction in alpha is probably due in large part to the fact that actual events (as distinct from nonevents) have a low probability (see Table 3) in most of the states of both HMMs.[19]

For contrast, Figure 5 shows the alpha curves for a set of random simulated data that has the same marginal event probabilities as the ISR>PAL data set but no autocorrelation.[20] Three

[19] This may also be due in part to the crudeness of the BCOW to WEIS translation. For example BCOW contains a "continuous military conflict" code that I translated into a single WEIS force event. In fact, such codes presumably indicate multiple consecutive days of WEIS force events. Such sequences are common during the interventions in Lebanon and during the *intifada* but would have no BCOW counterparts given my translation rules.

[20] The marginal probabilities are:
00:0.38; **01**:0.005; **02**:0.05; **03**:0.10; **04**:0.01; **05**:0.005; **06**:0.02; **07**:0.01; **08**:0.04; **09**:0.02; **10**:0.01; **11**:0.03; **12**:0.02; **13**:0.01; **14**:0.005; **15**:0.01; **16**:0.005; **17**:0.01; **18**:0.01; **19**:0.03; **20**:0.01; **21**:0.04; **22**:0.19. Multiple events are included in a single a day according to the probability

Prob(n events | not a **00** event) = $(0.5)^{n-1}$

features are evident in this figure. First, as one would expect, the two curves are basically just noise—due to the 100-event sequence length, they are significantly autocorrelated at a lag of one month but beyond one month the autocorrelation pattern is consistent with white noise. Second, the war and nonwar alpha curves themselves are highly correlated (r = 0.80; p<.001). Finally, the alpha value for the war model is consistently higher than the value for the nonwar model, which is to be expected because around 20% of the events in this sequence are force events.

Figure 8 and Table 4 compare the difference in the HMM alpha log-likelihoods with the Goldstein-scaled time series that we have been using for the last several years;[21] those data cover August 1979 to October 1996. Figure 8 shows a relatively close correspondence between the alpha-difference and the Goldstein score for Israel > Palestinian behaviors during most of the period. The correlation is a highly significant 0.30 for the entire period and 0.52 for the period before the Oslo agreements (September 1993). As noted in Table 4, the correlations between the Goldstein score and the difference between the HMM probabilities is less dramatic for the other two dyads but they are still significant.

This probability generates multiple events at a level that is actually a bit higher than the distribution found in the actual data.

[21] See Schrodt & Gerner (1994). We converted the individual WEIS events to a monthly net cooperation score using the numerical scale in Goldstein (1992) and totaling these numerical values for each of the directed dyads for each month. The Goldstein score has been divided by 4 to bring the two measures into scale with each other.

Figure 4. Alphas for Israel > Palestinians

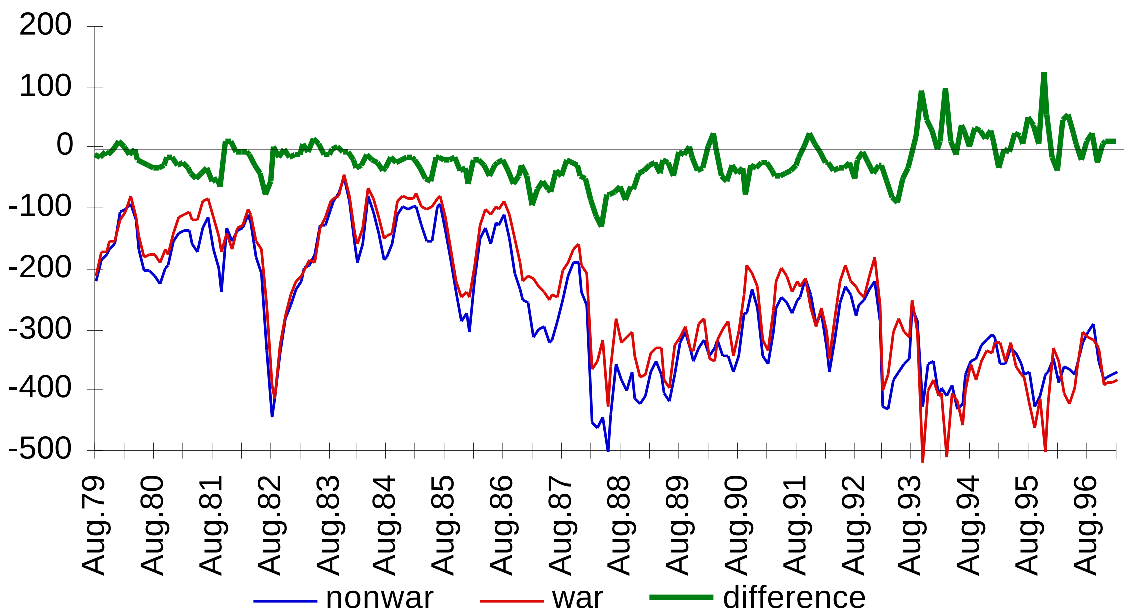


Figure 5. Alphas for random sequences

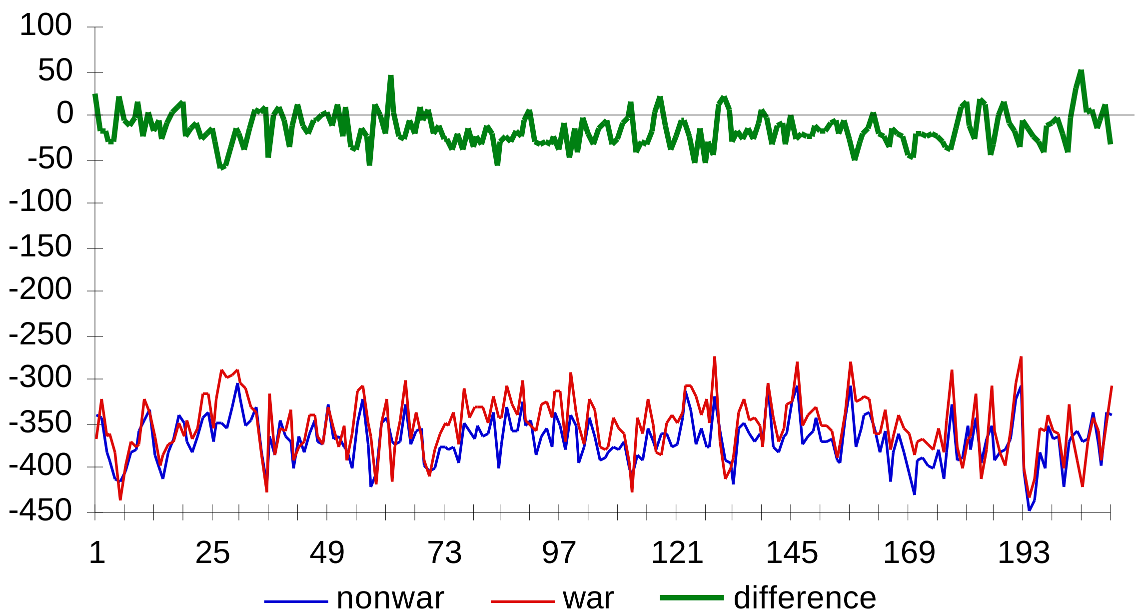


Figure 6. Alphas for Syria > Lebanon

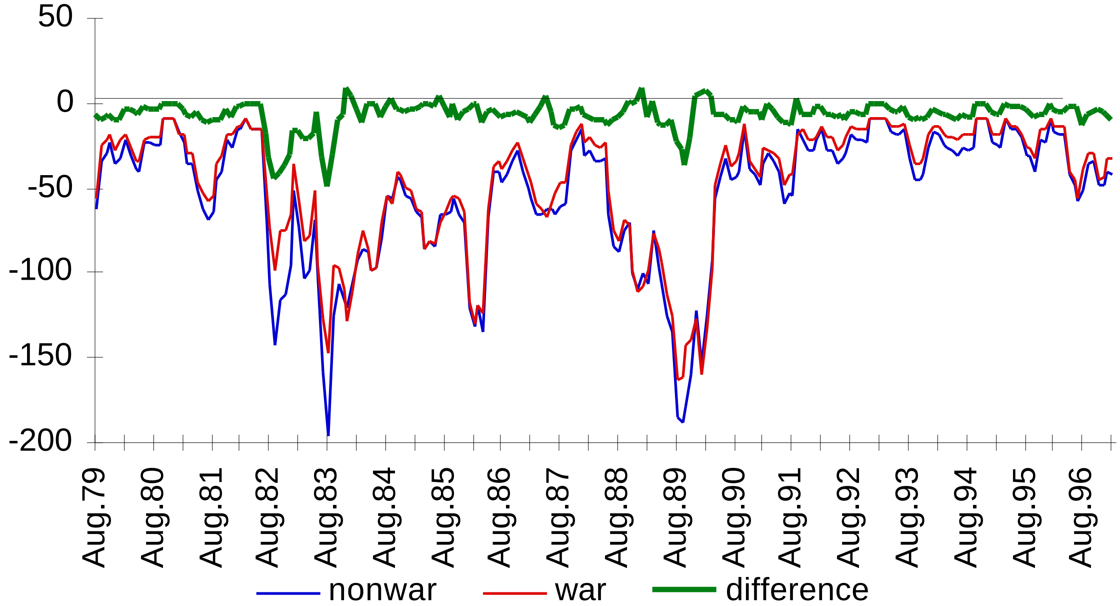


Figure 7. Alphas for Israel >Lebanon

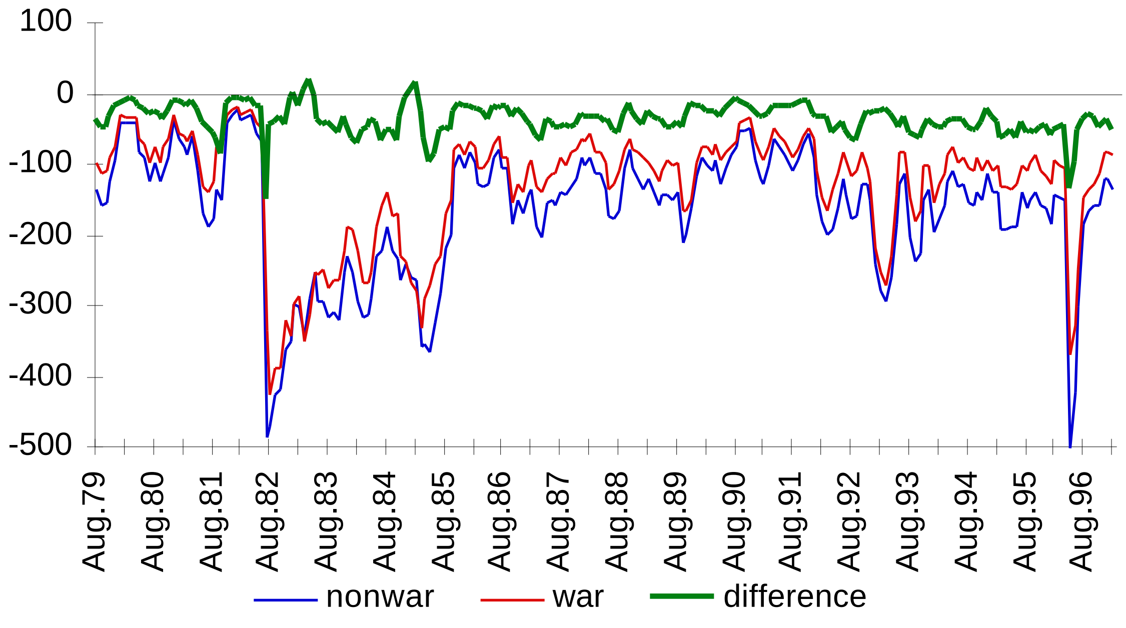


Figure 7. Comparison of Goldstein and HMM difference scores for Israel > Palestinians

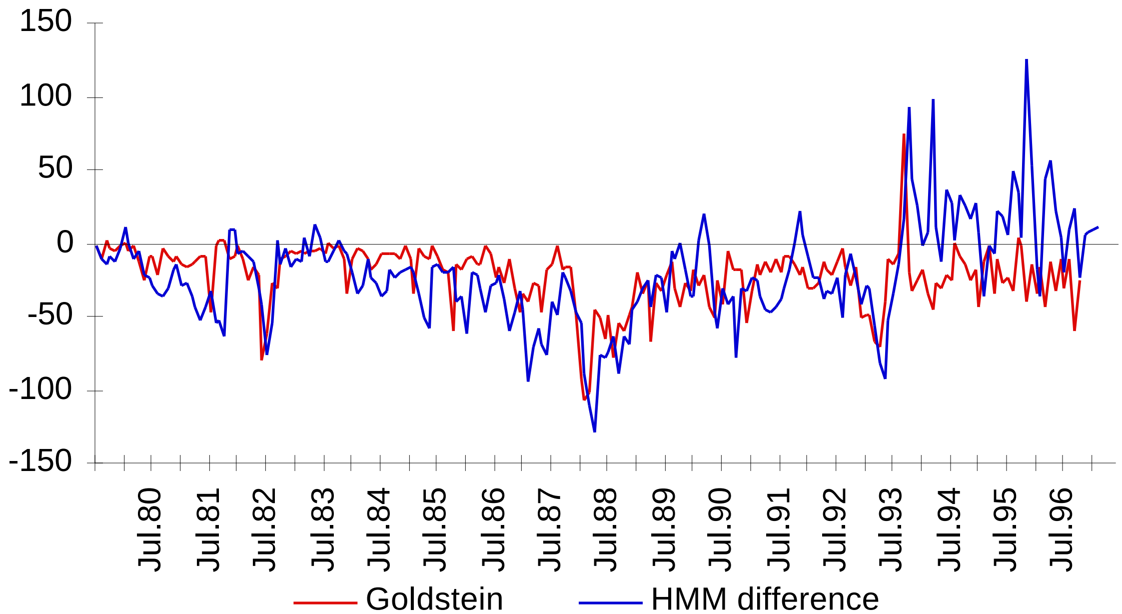


Table 4: Correlation between Goldstein scores and HMM difference

Dyad	N	r	t	prob.
Israel>Palestinians	207	0.30	4.42	<0.001
Israel>Palestinians, pre-Oslo	170	0.52	7.82	<0.001
Israel>Lebanon	207	0.15	2.24	0.026
Syria>Lebanon	207	0.20	2.89	0.004

While the alpha-difference and Goldstein scores in Figure 8 generally track each other, particularly on major events such as the invasion of Lebanon and the *intifada*, there are a couple of interesting distinctions. First, the alpha-difference is somewhat more sensitive in measuring the level of conflict (in the sense of moving away from the nonwar model) than is the Goldstein score: for example this is conspicuous in the period prior to the summer of 1981 where there was considerable conflict between Israel and PLO militias then residing in southern Lebanon. Second, the alpha-difference is much more sensitive to periods of negotiations than is the Goldstein score. This is most evident in the post-Oslo period but can also be seen in a positive peaks in October-December 1991 corresponding to the beginning of the Madrid negotiations; the positive point in that November-December 1981 corresponds to the cease-fire between the PLO (in southern Lebanon) and Israel that was brokered by the United States; and the peak in March-June 1983 appears to correspond to a series of prisoner-exchange negotiations brokered by Austria.[22]

The dramatic difference between the two scores in the post-Oslo period is probably due to a difference in the measures. The Goldstein scale is generally a cooperation-to-conflict continuum, where high positive values correspond to active cooperation. The "nonwar" sequences from BCOW, in contrast, represent militarized crises that are resolved just short of war. Relations between the Palestinians and Israel during the post-Oslo period are clearly closer to the latter situation—a continuous crisis punctuated by violent incidents—than they are to the active cooperation implied by positive values on the Goldstein scale. Hence the Oslo period provides a distinctly closer match to the nonwar HMM than to the war HMM despite the fact that it continues to be characterized by substantial levels of disagreement and occasional major outbreaks of violence.

[22] This last peak may be *too* sensitive—during the period of these negotiations there was continued Israeli-Palestinian conflict in Lebanon, the West Bank and Gaza, and the Reuters narrative does not support an interpretation of markedly improved relations.

Conclusion

The hidden Markov model is only one step in developing systematic techniques that deal with international events as sequences. The strength of the approach lies in its inductive nature. There are clearly simpler rules for distinguishing BCOW war and nonwar crises: looking for codes involving military conflict is the most obvious. But to construct those simpler rules, one must first know the characteristic that distinguishes the sequences: in a sense, one must already know the answer. An inductive method such as the Baum-Welch algorithm does not need to know the answer; it can *find* the answer. The system did not know, *a priori*, the importance of the WEIS codes designating military conflict: it discovered them. If a nonlinear model can discover those distinctions, it may be capable of discovering things that are not so obvious. In this concluding section, I will address several possible extensions of this technique, with particular attention to possible applications to crisis early warning.

The single most important extension of the work in this chapter would be to generalize the left-right model to one that can revert to the previous state in addition to going on to the next. There are no technical problems in doing this; it is simply a matter of computer programming. This would provide an HMM that was more consistent with the concept of crisis phase: a crisis can temporarily de-escalate into an earlier phase as well as escalating into the next phase. Such a model might provide a better differentiation of crisis states, and the left-right-left configuration makes every state accessible from every other state, including the "background" vector found in state A.[23]

Second, BCOW data were used because they provided a strong test of the ability of an HMM to generalize about types of political behavior. One would probably obtain cleaner models by working from a single contemporary source—Reuters and WEIS—rather than jumping across time, sources and coding schemes to obtain exemplars. The WEIS-coded crisis data set

[23] It may or may not be useful to eliminate the absorbing state by linking the final state back to the initial state. An absorbing state makes sense in BCOW—where the coding rules provide a definitive end to the crisis—but less sense in real-time monitoring, where the end of a crisis simply means going back to the background state.

being collected by Goldstein and Pevehouse (Goldstein 1997) that covers about a dozen contemporary crises—including the Arab-Israeli conflict, Iran-Iraq, Chechnya, the former Yugoslavia, and the Great Lakes of Africa—is an obvious source for this.

Finally, the war/nonwar crisis distinction used in this study is quite crude. A more sophisticated alternative would be to use Leng's (1993) typology of bargaining strategies—bullying, reciprocating, appeasement, stonewalling, and trial-and-error—to differentiate between dyadic political activities. The probabilities of a dyad fitting each of several different models would then place it in an N-dimensional vector space. This is a straightforward generalization of the Goldstein and Azar-Sloan scales, which place behaviors on a single conflict-cooperation dimension. To the extent that movement in this space—for example going from a conciliatory to a bullying bargaining strategy—is a precursor to later changes in the political environment, this would be useful for early warning.[24]

Furthermore, the ability of the HMM to determine models *by example*—in other words, to inductively determine the matrix from a set of cases rather than the analyst having to anticipate, deductively, the relative importance of various WEIS categories in the modes of behavior he or she wishes to study—simplifies the construction of metrics that go beyond those found in the classic conflict-cooperation continuum. Those novel metrics may, in turn, prove more useful in dealing with early warning in new political situations that may be important in the 21st century—for example state breakdowns and widespread ethnic conflict—and which do not fit neatly into the Westphalian behaviors assumed in the existing event data scales.

Alternatively, this technique could be used to simply monitor the likelihood of specific crisis precursors, without attempting to aggregate these probabilities into a single quantitative measure or a location in a vector space. In comparison with earlier techniques for the analysis of event data—which frequently required a great deal of statistical sophistication and "tweaking" of the resulting models—the HMM is sufficiently robust that it could be estimated by an analyst with little or no

[24] Schrodt and Gerner (1997) demonstrate a version of this vector-based approach to early warning by using the Goldstein-scaled behavior of various dyads in the Levant

knowledge of the underlying mathematical methods. In this scenario, the output of a monitoring system would be a list of probable matching sequences and their likelihoods. If the problem of comparability among sequences of different lengths could be worked out, an automated system (using machine-coded event data) could provide a real-time alert whenever the probability of a dyadic behavior matching one of the precursor models exceeded some threshold. This technique is substantially closer to the style of political analysis used in most policy settings, and therefore might be more acceptable than earlier event data efforts that relied on simple quantitative indicators without providing specific historical referents.

None of this is to suggest that the use of precedent and analogies is a panacea. Political forecasting will always be a difficult task, and the literature dealing with the use of precedent in political reasoning focuses at least as much on how analogies can be misused as how they are successfully used.[25] Yet political analysis, unlike weather forecasting or billiards, is a reflexive endeavor: Political behavior is determined in part by how individuals analyze politics. The most common flaws cited in the human use of historical analogy are the undue influence of superficial similarities, the failure to consider the role of background conditions, and a tendency to search only a limited set of candidate examples. These same flaws are likely to be shared by HMMs, so at worst these models may provide a good indicator of possible precedents that human political actors could be considering. At best, a more sophisticated system—perhaps combining HMMs with other techniques—could be developed that specifically avoids some of the problems known to occur in human political pattern recognition.

[25] Khong (1992) and Vertzberger (1990) tend to focus on failures; Neustadt & May (1986) provide a combination of successes and failures. Because foreign policy failures (such as the Bay of Pigs invasion and the Vietnam War) tend to be studied more intensely than successes (such as the forty-year stability of the Cold War borders in Germany and Korea), the effectiveness of precedent-based reasoning may be underestimated in the foreign policy literature.

Bibliography

Alker, H. R. 1987. Fairy Tales, Tragedies and World Histories: Towards Interpretive Story Grammars as Possibilist World Models. *Behaviormetrika* 21:1-28.

Allan, P. 1980. Diplomatic Time and Climate: A Formal Model. *Journal of Peace Science* 4:133-150.

Anderson, J. R. 1983. *The Architecture of Cognition.* Cambridge: Harvard University Press.

Azar, E. E., and Thomas Sloan. 1975. *Dimensions of Interaction*. Pittsburgh: University Center for International Studies, University of Pittsburgh.

Bartholomew, D. J. 1971. *Stochastic Models for Social Processes.* New York: Wiley.

Bennett, S. and P. A. Schrodt. 1987. Linear Event Patterns in WEIS Data. Paper presented at American Political Science Association, Chicago.

Bloomfield, L. P., and A. Moulton. 1989. *CASCON III: Computer-Aided System for Analysis of Local Conflicts.* Cambridge: MIT Center for International Studies.

Bloomfield, L. P. and A. Moulton. 1997. *Managing International Conflict.* New York: St. Martin's Press.

Butterworth, R. L.. 1976. *Managing Interstate Conflict,1945-74: Data with Synopses.* Pittsburgh: University of Pittsburgh University Center for International Studies.

Gerner, D. J., P. A. Schrodt, R. A. Francisco, and J. L. Weddle. 1994. The Machine Coding of Events from Regional and International Sources. *International Studies Quarterly* 38:91-119.

Gochman, C. S. and R. J. Leng. 1983. Realpolitik and the Road to War: An Analysis of Attributes and Behavior. *International Studies Quarterly.* 27: 97-120.

Goldstein, J. S. 1992. A Conflict-Cooperation Scale for WEIS Events Data. *Journal of Conflict Resolution* 36: 369-385.

Goldstein, J. 1997. Testing Theories of International Cooperation in Regional Conflicts. Project funded by the National Science Foundation.

Green, D. P. and I. Shapiro. 1994. *Pathologies of Rational Choice Theory*. New Haven: Yale University Press.

Hogarth, R. M. and M. W. Reder, eds. 1987. *Rational Choice: The Contrast between Economics and Psychology*. Chicago: University of Chicago Press.

Huxtable, P. A. and J. C. Pevehouse. 1996. Potential Validity Problems in Events Data Collection. *International Studies Notes* 21: 8-19.

Kahneman, D., P. Slovic and A. Tversky. 1982. *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Khong, Y. F. 1992. *Analogies at War*. Princeton: Princeton University Press.

Kohonen, T. 1984. *Self-Organization and Associative Memory.* New York: Springer-Verlag.

Kruskal, J. B. 1983. An Overview of Sequence Comparison. In *Time Warps, String Edits and Macromolecules,* ed. D. Sankoff and J B. Kruskal. New York: Addison-Wesley.

Leng, R. J. 1987. *Behavioral Correlates of War, 1816-1975*. (ICPSR 8606). Ann Arbor: Inter-university Consortium for Political and Social Research.

Leng, R. J. 1993. *Interstate Crisis Behavior, 1816-1980.* New York: Cambridge University Press.

Lund, M. S. 1996. *Preventing Violent Conflicts: A Strategy for Preventive Diplomacy.* Washington, D.C.: United States Institute for Peace.

McClelland, C. A. 1976. *World Event/Interaction Survey Codebook.* (ICPSR 5211). Ann Arbor: Inter-University Consortium for Political and Social Research.

May, E. 1973. *"Lessons" of the Past: The Use and Misuse of History in American Foreign Policy.* New York: Oxford University Press.

Mefford, D. 1985. Formulating Foreign Policy on the Basis of Historical Programming. In *Dynamic Models of International Conflict*, ed. U. Luterbacher and M. D. Ward. Boulder: Lynne Rienner Publishing.

Mefford, D. 1991. Steps Toward Artificial Intelligence: Rule-Based, Case-Based and Explanation-Based Models of Politics. In *Artificial Intelligence and International Politics*, ed. V. Hudson. Boulder: Westview.

Myers, R. and J. Whitson. 1995. HIDDEN MARKOV MODEL for automatic speech recognition (C++ source code). http://www.itl.atr.co.jp/comp.speech/Section6/Recognition/myers.hmm.html

Neustadt, R. E. and E. R. May. 1986. *Thinking in Time: The Uses of History for Decision Makers.* New York: Free Press.

Rabiner, L. R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings. of the IEEE* 77,2:257-286

Sankoff, D. and J. B. Kruskal, eds. 1983. *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison.* New York: Addison-Wesley.

Schrodt, P. A. 1985. Stochastic Models in International Relations Research. In *Theories, Models and Simulation in International Relations,* ed. M. Ward. Boulder, CO: Westview.

Schrodt, P. A. 1990. Parallel Event Sequences in International Crises, 1835-1940. *Political Behavior* 12: 97-123.

Schrodt, P. A. 1991. Pattern Recognition in International Event Sequences: A Machine Learning Approach. In *Artificial Intelligence and International Politics,* ed. V. Hudson. Boulder: Westview.

Schrodt, P. A. and D. J. Gerner. 1994 . Validity assessment of a machine-coded event data set for the Middle East, 1982-1992. *American Journal of Political Science* 38: 825-854.

Schrodt, P. A., S. G. Davis and J. L. Weddle. 1994. Political Science: KEDS—A Program for the Machine Coding of Event Data. *Social Science Computer Review* 12: 561-588.

Schrodt, P. A., and D. J. Gerner. 1997. Empirical Indicators of Crisis Phase in the Middle East, 1982-1995. *Journal of Conflict Resolution* 41:529-552.

Sherman, F. L., and L. Neack. 1993. Imagining the Possibilities: The Prospects of Isolating the Genome of International Conflict from the SHERFACS Dataset. In *International Event-Data Developments: DDIR Phase II*, ed. R. L. Merritt, R. G. Muncaster, and D. A. Zinnes, 87-112. Ann Arbor: University of Michigan Press.

Vertzberger, Y.I. 1990. *The World in their Minds: Information Processing, Cognition and Perception in Foreign Policy Decision Making.* Stanford: Stanford University Press.

Weigand, A. S. and N. A. Gershenfeld. 1994. *Time Series Prediction: Forecasting the Future and Understanding the Past.* Reading, MA: Addison Wesley

Appendix: WEIS equivalents of BCOW codes

The following rules were used to convert the BCOW events to WEIS-coded events:

Physical actions

11212	07	21121	07	12131	06	23121	08
11719	22	21133	18	12183	19	23131	08
11121	07	21143	19	12173	21	23151	19
11131	08	21211	01	12373	06	23163	21
11333	17	21233	21	12719	02	23171	01
11353	18	21311	07	12223	22	23301	06
11413	01	21333	01	12232	03	23141	19
11313	18	31121	08	12243	19	23211	01
11363	22	31132	06	32111	01	23223	21
11443	22	31133	17	32132	21	23231	01
11433	22			32141	01	23251	01
11423	21	12111	03	32142	21	23261	01
11453	18	12121	03	32163	21	33111	06
11513	22	12521	08	32153	21	33131	06
11523	22	12511	08	32143	21	23719	02
11533	22	12361	01	32151	01		
11553	22	12142	10	32161	01	14113	22
11521	22	12152	06	32173	01	14123	22
11663	01	12223	19	32611	01	14143	22
11673	21	12342	12			14151	03
11633	22	12362	05	13111	03	14153	21
11643	22	12161	19	13121	03	14213	18
11621	01	12631	03	13131	03	14223	18
11653	21	12641	21	13211	03	14251	04
21141	06	12533	19	13551	08	14263	21
21111	07	12363	19	23111	06	14719	02

Verbal Actions

col. 26 code	col. 29 code	WEIS code
1	1	04
1	2	02
1	3	12
2	any	05
3	any	09

This coding system does not generate WEIS events in the following categories:

06, 07, 10, 11, 13, 14, 15, 16, 20

2-Digit WEIS Categories

01	Yield	11	Reject	20	Expel
02	Comment	12	Accuse	21	Seize
03	Consult	13	Protest	22	Force
04	Approve	14	Deny		
05	Promise	15	Demand		
06	Grant	16	Warn		
07	Reward	17	Threaten		
08	Agree	18	Demonstrate		
09	Request	19	Reduce Relationship		
10	Propose				