Early Warning of Conflict in Southern Lebanon using Hidden Markov Models

Philip A. Schrodt
Department of Political Science
University of Kansas
Lawrence, KS 66045 USA
phone: 785-864-3523 fax: 785-864-5700
p-schrodt@ukans.edu

Paper presented at the annual meeting of the American Political Science Association,
Washington, DC

August 1997

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 WEIS-coded Middle East event data set and the estimation programs used in this paper can be downloaded from the Kansas Event Data System web site: http://www.ukans.edu/~keds. My thanks to Larry Bartels, Scott Bennett, Deborah Gerner, Joshua Goldstein, Walter Mebane, Mohan Penubarti, and Jas Sekhon for helpful comments on earlier presentations of the hidden Markov approach.

Abstract

This paper extends earlier work on the application of hidden Markov models (HMMs) to the problem of forecasting international conflict. HMMs are a sequence comparison method widely used in computerized speech recognition as a computationally efficient method of generalizing a set of sequences observed in a noisy environment. The technique is easily be adapted to work with \international event data. The paper provides a theoretical "micro-foundation" for the use of sequence comparison in conflict early-warning based on coadaptation of organizational standard operating procedures.

The left-right (LR) HMM used in speech recognition is first extended to a left-right-left (LRL) model that allows a crisis to escalate and de-escalate. This model is tested for its ability to correctly discriminate between BCOW crisis that involve and do not involve war. The LRL model provides slightly more accurate classification than the LR model. The interpretation of the hidden states in the LRL models, however, is more ambiguous than in the LR model.

The HMM is then applied to the problem of forecasting the outbreak of armed violence between Israel and Arab forces in south Lebanon during the period 1979 to 1997 (excluding 1982-1985). An HMM is estimated using six cases of "tit-for-tat" escalation, then fitted to the entire time period. The model identifies about half of the TFT conflicts—including all of the training cases—that occur in the full sequence, with only one false positive. This result suggests that HMMs could be used in an event-based monitoring system. However, the fit of the model is very sensitive to the number of days in a sequence when no events occurred, and consequently the fit measure is ineffective as an early warning indicator.

Nonetheless, in a subset of models, the maximum likelihood estimate of the sequence of hidden Markov states provides a robust early warning indicator with a three to six month lead. These models are valid in a split-sample test, and the patterns of cross-correlation of the individual states of the model are consistent with theoretical expectations. While this approach clearly needs further validation, it appears promising.

The paper concludes with observations on the extent to which the HMM approach can be generalized to other categories of conflict, some suggestions on how the method of estimation can be improved, and the implications that sequence-based forecasting techniques have for theories of the causes of conflict.

Introduction

The problem of developing early warning indicators of political conflict has been an important focus of quantitative international relations research from almost the beginning of the "scientific" approach. The pioneering arms race modelling work of Lewis Richardson, for example, was motivated in part by Richardson's assumption that unstable arms races were an important precursor to war; Choucri & Robinson (1979), Singer & Wallace (1979) and Hopple, Andriole & Freedy (1984) provide additional examples of early quantitative studies on this problem.

Following a large and generally unsuccessful effort in the late 1970s to develop early warning indicators using event data (see Laurance 1990, Schrodt 1994), early warning research shifted its focus to other techniques. Most notable among these were the expected utility models of Bueno de Mesquita and his colleagues (Bueno de Mesquita 1980; Bueno de Mesquita, Newman & Rabushka 1996); another substantial effort involved computational models derived from artificial intelligence methods (Cimbala 1984; Hudson 1991) and systems dynamics (Hughes 1984; Ward 1985). With this shift in techniques, the collection of contemporaneous event data sets—particularly those readily available in public archives—slowed and eventually stopped except for a few individual efforts.

Interest in event-based early warning began to revive about a decade ago, when the NSF's Data Development in International Relations project sponsored several new event data collections (Merritt, Muncaster and Zinnes 1993). These efforts gained greater momentum in the policy community with the end of the Cold War, which vastly complicated the monitoring tasks of governments and international organizations who were interested in conflict mediation. The earliest manifestation of this trend in the realm of event-based research were the efforts by Gurr, Harff and others (Gurr & Harff 1996) to study the precursors and accelerators of ethnoconflict and state breakdown. The still-classified "State Failure Project", nominally sponsored by the office of the Vice-President of the United States, was a very large-scale quantitative effort using several hundred variables in linear and nonlinear models to identify cross-sectional precursors to state breakdowns. In March 1997, a conference on early warning in Toronto attracted over a hundred representatives of academic, government, IGO, and NGO organizations interested in early warning, and the "Middle East Prediction Project", coordinated by Stephen Weber and Janice Gross Stein, is systematically assessing forecasting techniques in the Middle East.[1]

Despite all of this attention, early warning remains a difficult problem, whether done with quantitative or qualitative methods. For example, notwithstanding funding in the billions of dollars, access to a wide variety of information sources and a clear focus on a single opponent, Western intelligence agencies failed to anticipate both the timing and characteristics of the collapse of the Warsaw Pact. Early warning is almost nonexistent in low-priority areas such as Somalia, Rwanda, and Sierra Leone. In some of these cases, as I will argue below, early warning may be impossible for theoretical reasons. However, there are other cases where advances in communications and analytical techniques should make possible the development of indicators that would not have been feasible when quantitative early warning research began thirty years ago.

The objective of this paper is to develop such a model. Its fundamental premise is that a significant subset of international behaviors consist of regularized sequences of events that are repeated—in a noisy fashion—over time. These sequences can be used to predict subsequent behavior in a manner similar to the cognitive processes used by many human political analysts and decision-makers. By developing a computational early warning algorithm, both the successes and failures of the technique can be studied statistically, and the process can be refined incrementally. The models can also be employed in systematic early-warning efforts. The analysis focuses on tit-for-tat violence between Israel and Arab military forces in southern Lebanon for the period 1979-1997, excluding the 1982-1985 period when Israeli forces occupied parts of Lebanon north of the Litani River. The technique employed is the hidden Markov model (HMM).

[1] For further information on this project, see http://www.arizona.edu/~spiro/mideast.html

Micro-Foundations:Behavioral

In keeping with recent concerns within the political methodology community that formal models should have micro-foundations—a "story" as to why human behavior might be expected to follow the patterns assumed by a model—some theoretical justification of the use of sequences is appropriate. The sequence analysis approach has a long history in political science—at the most fundamental level, it is simply a systematic rendition of the "case study" or "lessons of history" technique that has been used by decision-makers since time immemorial (see May 1973, Mefford 1985, Neustadt & May 1986, Vertzberger 1990, Khong 1992) . History is considered relevant to decision-makers because they assume that when a particular set of events and circumstances observed in the past is observed again, the resulting events from that prior case can also be expected to apply in the new case, all other things being equal.

This simple observation is both reinforced and attenuated by the fact that it is reflexive—the methods that decision-makers use to interpret the past have an impact on how they create the future. If decision-makers act consistently on the "lessons of history", then history will in fact have lessons.

By itself, however, belief in the importance of historical examples is insufficient to create empirical regularities because of "Van Crevald's Law"[2]: A conspicuously successful strategic innovation is unlikely to succeed twice precisely because it was successful the first time. More generally, work of the Santa Fe Institute on the so-called the "El Farol Problem" (see Casti 1997) has demonstrated that systems of adaptive utility maximizers generally do not exhibit regularized behavior *because* they look at history. In computer simulations, such agents tend to show quasi-chaotic behavior that is *not* predictable. If the political world consists solely of rational

[2] "...war consists in large part of an interplay of double-crosses [and] is, therefore, not linear but paradoxical. The same action will not always lead to the same result. The opposite, indeed, is closer to the truth. Given an opponent who is capable of learning, a very real danger exists that an action will not succeed twice *because* it has succeeded once." (Van Creveld 1991:316; italics in original).

adaptive agents, there is little point in trying to make predictions based on past behaviors.[3] There are undoubtedly some forms of international behavior (for example international exchange-rate behavior) for which this is true.

But it is not be true in all cases. Situations of international conflict usually involve organizational behavior rather than individual behavior, and for a variety of reasons both theoretical and practical, organizations are substantially less likely to engage in rapidly adaptive behavior than are individuals. Mature organizations instead are likely to rely on rule-based standard operating procedures (SOPs) that are designed to insure that a specific set of stimuli will invoke a specific response (Cyert and March 1963, Allison 1971). A classical Weberian bureaucracy, unlike the adaptive maximizer of complexity theory, is virtually designed to assure the success of a sequence analysis approach.

The SOPs are themselves adaptive—they are designed to effectively solve problems and many are acquired through historical experience. But in a situation of the protracted interaction, two organizations with SOPs are *coadaptive*: each responds in part to the environment created by the other.[4] In most circumstances, this eventually brings their SOPs into a Nash equilibrium within the space of possible SOPs where neither can change strategies unilaterally without a loss of utility. This is more likely to occur when the same organizations have been interacting over a period of time, and when the payoff environment has been relatively stable. This is found, notably, in the situation of protracted conflicts and enduring rivalries. These are situations characterized by exactly the competitive SOP "lock-in" that I've outlined above—antagonists fight, on repeated occasions, over the same issues, often over the same territory, and without resolution.

[3] Predictions could still be made on the basis of other characteristics of the system—for example the effects that economic or technological changes have on the utility functions of the actors, and even predictions about the *range* of strategic outcomes. But in the absence of a completely specified model and complete information, there is little point in trying to make point predictions in a chaotic system.

[4] A detailed discussion of the concept of coadaptation is beyond the scope of this paper, but general discussions from a natural science perspective can be found in Maynard-Smith (1982) and Kauffman (1993); Anderson, Arrows and Pines (1988) discuss a number of social science applications, and Schrodt (1993) applies the concept to the issue of international regimes.

To summarize, sequence-based prediction will not work in all circumstances, but it will work in a significant number of cases. In addition, those instances where it will not work—rapid and complex adaptation—are frequently situations where other methods are not going to work either. This relevance of event sequences may also explain in part why study of history remains popular with politicians and diplomats despite our best efforts to divert them to the study of game theory and statistics.

Micro-Foundations: Analytical

The empirical problem I am studying in this paper is predicting political change based on an *irregular nominal time series with a stochastic component*. Because this type of data is quite different than that found in most political science studies (which for time series analysis usually employs methods derivative of econometrics), some definitions are in order:

irregular In contrast to most econometric time series, the observations in an event data series occur at irregular intervals: many days may pass between events, and multiple events can occur in a single day.

nominal In this analysis I will be using the 22 discrete categories of activity recorded in the 2-digit codes of the World Events Interaction Survey scheme (WEIS; McClelland 1976), plus a "non-event" category. While these categories are roughly ordinal on a conflict-cooperation dimension—and most statistical studies of event data use interval-level scales such as Goldstein's (1992)—the original events are not scaled. Furthermore, the news reports from which event data are generated consist of categories of events codified by natural language; they do not deal with interval-level measures. This contrasts with most econometric data, such as prices, interest rates, unemployment rates, proportions, volumes, populations and so forth, where the underlying metric is an interval or ratio variable.

time series Despite their irregularity of incidence, each event in an event data set can be assigned to a point in time. In event data generated from newswire sources, the timing of an event is typically accurate to a day; coding errors (particularly with machine coding) probably introduce another day or two of uncertainty in the data of some events. Times series concepts such as autocorrelation and cross-correlation are therefore relevant to event data: For example the occurrence of an event at time t will affect the probable distribution of events at times t+k.

stochastic Event data contain a variety of sources of random error; I have discussed these in substantial detail in Schrodt (1994). From the standpoint of a predicting sequences, these stochastic components involve at least the following:

background noise An event occurs that appears to be a precursor but is due to causes independent of future events.

editorial error Events occur but are not reported by Reuters or alternatively on a slow news day, Reuters may report events that it normally would not report. I will usually refer to this by the statistical term "censoring" even though in the Levant it is usually not due to political censoring.[5]

coding error The code assigned to a story by the machine-coding system is different than what should have been assigned according to the WEIS system. This is typically due to sentences that have unusual grammatical constructions; these are relatively rare and can usually be avoided with an appropriate filter.

schematic error The WEIS coding system combines two sets of behavior that have distinct natural-language representations—and which should remain distinct for the purposes of prediction—into a single category. WEIS may also do the opposite—separate two sets of behavior that could be combined—as in WEIS's notoriously overlapping "Warning" and "Threaten" categories.

With considerable loss of specificity, I will usually refer to all of these stochastic components as "noise."

In addition to these problems, the process of developing a predictive model must also deal with specification error and standard errors in estimating the parameters. In short, sequence analysis is just a conventional, if rather messy, statistical modeling problem.

[5] Plenty of governmental censorship of journalists goes on, but the area is sufficiently well monitored that a story censored in Israel normally will be reported from Lebanon and vice versa. Overt censorship rarely succeeds for more than a day or two, though these delayed reports are one of the sources of timing errors in the data.

The conjunction of all of these characteristics means that most conventional time series techniques are completely inappropriate for the analysis of the disaggregated event data stream. The almost universal response to this in statistical analyses is to generate a regular, interval-level time series by aggregating the data at fixed intervals (typically a month or a year) using a scaled value assigned to each category (e.g. Goldstein's (1992) recent WEIS scale, or the earlier Azar-Sloan (1975) scale for the COPDAB data set). Standard interval-level time series methods can then be used on the aggregated data.

The advantage of this approach is that a wide variety of methods are readily available. The clear disadvantage is that the process of reducing behavior to a single dimension through scaling loses a great deal of information and introduces a large number of free parameters. For example in principle (although almost never in practice), a month characterized by a large amount of conflict in the first two weeks (negative numbers on most scales), followed by a large amount of reconciliation in the last two weeks (positive numbers) could aggregate to value close to zero, which is the same value that would occur in a month where nothing happened.

A second, more subtle, problem occurs with aggregation: it removes the analysis a step further from the cognitive and organizational processes that are generating the events. While decision-makers may do some aggregation—one of the most commonly used metaphors in political analysis is indicating whether a situation is "heating up" or "cooling down"—detailed political responses are usually triggered by specific sets or sequences of events, not by the crossing of some numerical threshold.

In political activity, unlike economic activity, both the stimuli and responses are likely to be discrete, not continuous. Prices of stocks or the levels of interest rates, for example, move in predictable adjustments and when they fail to move continuously across that range (as in an investigation of NASDAQ trading a couple years ago), suspicions are triggered. Furthermore, small changes in the price will almost always result in proportionally small changes of supply and demand.

Political events, in contrast, move in jumps that are predicated on the prior state of the system. The fall of a single rocket following a period of peace will trigger a major response, whereas the fall of a single rocket during a period of war usually will go unnoticed. A model that can maintain the event data in its disaggregated form is, *ceteris paribus*, more likely to be successful in predicting actual behavior.

The Weakness of Linear Modeling

A recent paper by Hinich (1997) also makes some interesting observations on the limited utility of standard linear forecasting models—for example the Box-Jenkins paradigm—in forecasting political behavior. Hinich observes that the stochastic linear model provides very poor predictions (particularly in the long term) if the system is highly autoregressive and, in a later discussion, noted that linear estimates are predicated on the stochastic disturbance terms of the process being independent. In a tightly-interlinked and historically-sensitive system such as the Levant, behavior is autoregressive, but the errors are not independent, and consequently the utility of linear models is severely compromised.

Political behavior in the Levant is autoregressive in the sense that the effects of a disturbance such as the Lebanese civil war, Israel's invasion of Lebanon, the outbreak of the Palestinian intifada or the assassination of Israeli Prime Minister Rabin will be apparent in the event data series for a period of years. Or, to get into a *really* autoregressive series, it is not coincidental that some of the territory involved in the current Israeli-Arab conflict is dotted with fortifications remaining from the Crusades—a millennium in the past—or that the border between Maronite and Druze control in the Lebanese civil war was just a few kilometers south of the famous Nahr al-Kalb, a canyon containing inscriptions from armies dating back to the Babylonians and Egyptian pharaoh Ramses II.

In addition, organizational SOPs of organizations cause error disturbances to be correlated. Any random outbreak of violence in the Arab-Israeli conflict sets off a ritualized set of accusations, protests, attempts at mediation and appeals for restraint by regional and international actors: the pattern depends only on who initiated the violence. In fact, if one removes the proper names, much of the Reuters record for this period is difficult to classify by date: Many a traveler in the region has had the unsettling experience of reading a newspaper for several pages only to notice, after finally encountering some glaring anachronism[6], that the newspaper is three weeks, or three months, or even three years, out of date.

Because of these characteristics, the linear modelling approach is not going to work very well. Those same characteristics, however, improve the likelihood that a sequence comparison approach will work. Additional *prima facie* evidence of this is found in the failure of the earlier DARPA efforts: Linear prediction techniques (unlike nonlinear methods) were well-developed at the time of the DARPA work, and the computer power available at the time was sufficient for that task. If event prediction could have been solved using linear methods, that probably would have been discovered a quarter-century ago.

Hidden Markov models[7]

Techniques for comparing two sequences of discrete events 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 two decades because it is important in the study of genetic sequences 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.

[6] The glaring anachronism is frequently an item of *popular* culture, which is less static in this region than political conflict.

[7] This section is taken from Schrodt (1998) with minor modifications.

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). As with a conventional Markov chain, a HMM consists of a set of discrete states and a matrix $\mathbf{A} = \{a_{ij}\}$ of *transition probabilities* for going between those states. In addition, however, every state has a vector of *observed symbol probabilities*, $\mathbf{B} = \{b_j(k)\}$ 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".[8]

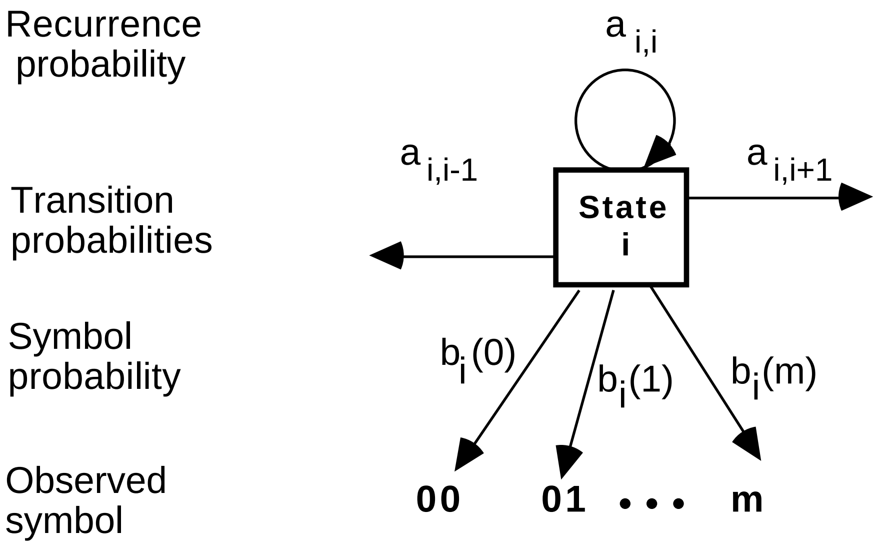
While the theory of HMM allows any type of transition matrix, the model that I will be testing allows transitions only to the previous state and the next state (as well as remaining in the current state). This is an extension of the unidirectional "left-right" (LR) model that is widely used in speech recognition and analyzed in Schrodt (1998); it allows the possibility that a crisis can de-escalate into a lower state as well as moving forward to the next state. The transition matrix **A** is therefore of the form

[8] 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) .

$$\begin{pmatrix} a_{11} & 1-a_{11} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & a_{23} & 0 & \dots & 0 \\ 0 & a_{32} & a_{33} & a_{34} & \dots & 0 \\ \dots & & & & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & a_{n-1,n} \\ 0 & 0 & 0 & 0 & \dots & a_{nn} \end{pmatrix}$$

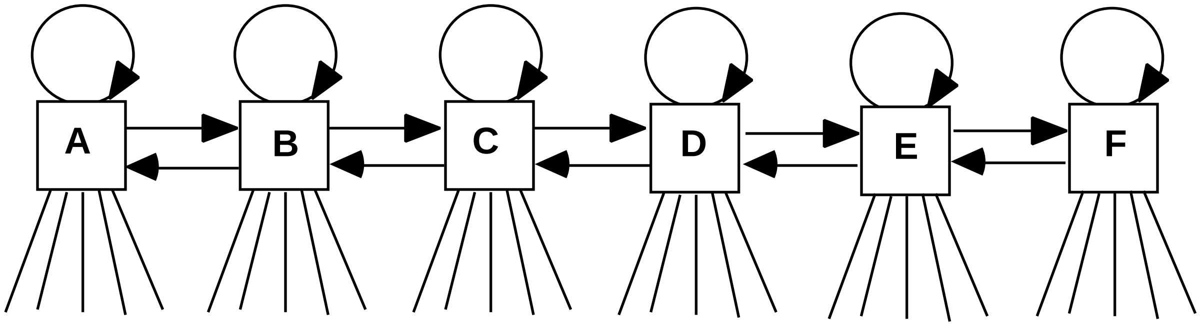
and the individual elements of the model look like those in Figure 1. I will refer to this as a "left-right-left" (LRL) model.

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



A series of these individual elements form an HMM such as the 6-state model illustrated in Figure 2. In contrast to the LR model, every state is accessible from every other state. Consistent with the fact that I'm using the model to study escalation behavior—that is, crises that have a clear "beginning" and "end"—sequences are assumed to start in State A.

Figure 2. A left-right-left (LRL) hidden Markov Model



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.[9] 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 an 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.[10] 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. Finally, a technique called the "Viterbi algorithm" can

[9] Rabiner (pg. 253) notes that the Baum-Welch algorithm is equivalent to the more familiar "expectation-modification" (EM) approach of Dempster, Laird and Rubin.

[10] 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: 262).

be used to estimate the most likely set of hidden states that the system was in, given an observed set of symbols and a set of transition and observation probabilities.[11]

Matching a sequence of symbols such as those found in daily data on a three-month crisis coded with using the 22-category WEIS scheme 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[12]—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 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 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.[13]

The application an HMM to the problem of generalizing the characteristics of international event sequences is straightforward. The symbol set consists of the event codes taken from an event data set such as WEIS or BCOW. The states of the model are unobserved, but have a close

[11] When the hidden state sequence is of interest, parameter estimation of an HMM can also be interpreted as a type of inductive clustering. The "states" of the HMM correspond to various clusters of behavior that are described by the symbol observation vectors **B**, and the sequence generated by the Viterbi algorithm gives a cluster assignment to each event in the sequence. It is also important to note that the Markov character of the model applies to the hidden states, not to the individual events. For example the Viterbi algorithm computes the maximum likelihood sequence of the sequence as a whole, not just to consecutive pairs of events.

[12] 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 type of 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.

[13] 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.

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).[14] 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 such as the Levenshtein metric or neural networks. First, if N<<M, the structure of the model is relatively simple. For example an LRL model with N states and M symbols has 2(N-1) + N(M+1) 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; both of these are present in international event sequences.

[14] 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 scaled aggregative methods that 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. 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. HMMs estimated from international event data tend to include one or two "background" states that correspond closely to the distribution of events generated by a particular source (e.g. Reuters/WEIS) when no crisis is occurring. 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 a clear probabilistic interpretation of the model that uses familiar structures and concepts such as 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.

Data and Early Warning Criteria

The event data used in this study were machine-coded using the WEIS system from Reuters lead sentences obtained from the NEXIS data service for the period April 1979 through May 1997 using the Kansas Event Data System (KEDS) program (Gerner et al. 1994; Schrodt, Davis & Weddle 1994).[15] 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. A **00** nonevent was added for each day in which no events were recorded in either direction in the dyad. Multiple events occurring in the same day are kept in the sequence.

The focus of the early warning analysis is tit-for-tat (TFT) military conflict between Israel and various Arab military organizations in southern Lebanon. Prior to 1982, this usually involved Palestine Liberation Organization (PLO) military forces; after 1985, it usually involved the Amal or Hizballah militias. This region has seen substantial military contention from almost the beginning of the Zionist presence in mandatory Palestine—for example the oft-targeted Israeli town of Kiryat Shimona is named in memory of eight settlers who died in one such clash in 1920. There is also ample reason to believe that organizational SOPs govern behavior on both sides: Israel, the PLO and the Shi'a militias all have extensive political and command infrastructures.

[15] 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 and a sentence was not coded if it contained six or more verbs or no actor was found prior to the verb (sentences meeting these criteria have a greater-than-average likelihood of being incorrectly coded by KEDS). This produced a total of 3,497 ISR>LEB and LEB>ISR events for the entire April 1979 to June 1997 period.

While KEDS is capable of distinguishing between different actors who are likely to engage in uses of force in Lebanon—for example Amal versus Hizballah—I did not do so in this study. Forces *allied* with Israel—notably Israel's client militia, the "South Lebanon Army"—are coded as Israeli rather than Lebanese. This coding probably underestimates activity prior to 1982, when some uses of force in southern Lebanon were coded as Palestinian rather than Lebanese.

With one major exception—the transition of anti-Israel forces in southern Lebanon from Palestinian to Shi'a—the actors have remained the same and consequently organizational co-adaptation is likely to have occurred over time. The analysis skips over the 1982-1985 period during which the military opposition shifted from the PLO to the Shi'a forces and coadaptation was occurring between Israel and its new opposition in the region.[16]

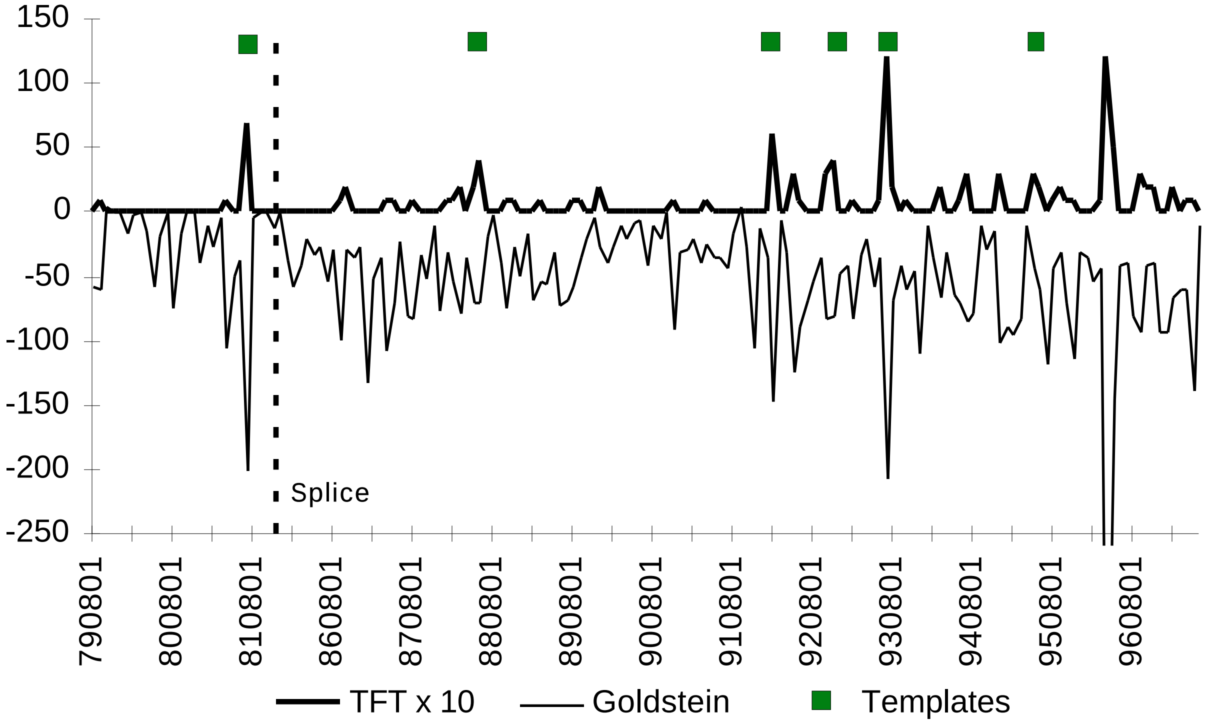
Two different predictive targets are being used: the number of TFT incidents, and the Goldstein-scaled score of the ISR>LEB conflict.[17] A TFT conflict is defined as a use of force (WEIS 22) by one party (either Israel or Lebanon) followed by a reciprocal use of force by the other within two days. These events are aggregated by month. Figure 3 shows the time series for these two sets of data.

[16] The calculation of the cross-correlation does not include 1982-1985, although the sequences fitted to the HMM include information from this period when that is necessary to complete a 100-event sequence (i.e. the Jan.86 to Mar.86 subsequences include some events from 1985).

If the 1982-1985 period is included in the assessment of predictive power, the results are considerably weaker, although to the extent that I looked at them, they are generally consistent with the results of the spliced model (for example the background and template models track each other closely and show the opposite of the expected correlations with the indicators). While the focus of military conflict is southern Lebanon, some of the Israeli retaliation occurs well outside of this area—air attacks on militia camps near Beirut are fairly common and are included as TFT events. Attacks by the Arab forces operating from Lebanon, whose air power has been confined to the occasional motorized hang glider, are exclusively on Israeli and SLA forces operating in southern Lebanon and attacks into the Hula valley and western Galilee (notably Kiryat Shimona and environs).

[17] The Goldstein scores for ISR>LEB and LEB>ISR are highly correlated, with r=0.82 (N=171), so only one of these dyads is analyzed.

Figure 3. Time series of the TFT and Goldstein scores



The success of the prediction will be assessed with cross-correlation—the correlation of W_{t-k} with X_t, where W is the warning indicator and X is the behavior to be predicted.[18] Most of the assessment will be done with cross-correlograms such as Figure 5 [below]: high correlations at negative values of the lag imply that X correlates with *earlier* values of W (i.e. W is an early-warning indicator); high correlations at positive values of the lag imply that X correlates with *later* values.[19] A custom program is used to compute the cross-correlations appropriately

[18] An aside: I am somewhat puzzled that cross-correlation is not used more commonly in political science time series research. No information is lost in cross-correlation compared to more complex methods such as Box-Jenkins, Granger analysis, VAR and spectral analysis, as each of these techniques have an identical set of sufficient statistics: the autocorrelation functions and cross-correlation functions of the variables under study. Granted, the more complex techniques allow multiple lags, but given the high levels of autocorrelation typically found in social science data, the collinearity resulting from multiple lags will inflate the standard errors of the parameter estimates and often confuse an analysis as much as they clarify it. Cross-correlation is simple to interpret and seems like a good place to start on the analysis of most political time series.

[19] Positive "lags" are not early warning, but frequently are useful for diagnostic purposes.

despite the splice in the data set.[20] The resulting sample size is around 160 and the critical values of r for a two-tailed significance test are

p=0.10:	0.131	p=0.05:	0.155	p=0.01:	0.203

Note that the empirical analysis employed here violates virtually all of the assumptions of the standard significance test so these levels should be considered illustrative only.

Estimation Algorithm

The HMM parameters were estimated by extensively modifying the source code written by Meyers & Whitson (1995). Their C++ code implements an LR hidden Markov model and the corresponding Baum-Welch maximum likelihood training algorithm. I translated this code from the Solaris C++ environment to a Macintosh CodeWarrior ANSI C environment, 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 WEIS sequences. The source code for this program is available at the KEDS web site: http://www.ukans.edu/~keds. I then extended the code to handle the LRL model, and implemented the Viterbi algorithm described in Rabiner (1989) in order to estimate the most likely state sequence.[21]

The resulting program is very fast—estimation of the HMM matrices using six 100-event sequences with a 45-symbol set and 64 Monte-Carlo iterations of the initial matrix took about 45 seconds on a Power Macintosh 7100/80, and the computation of the probability of a sequence

[20] If you are attempting to replicate this at home using a garden-variety statistical package, you'll find that the sample size is sufficiently large that a cross-correlation which ignores the splice gives much of the same results. Then again, no one is likely to replicate much of this study without some knowledge of programming...

[21] The Meyers & Whitson code is clean, well-documented, and survived my translation to run correctly the first time. I would assume that either my C code or their C++ code would port easily to a DOS/Windows or OS/2 environment for those so inclined. In the process of extending the model to the LRL form, I rewrote the estimation equations to correspond exactly to those in Rabiner—the Meyers & Whitson implementation differed slightly from Rabiner's equations, presumably because their models estimate a separate vector for "transition symbols." These new procedures produce estimates similar to those of Meyers & Whitson when all probabilities to previous states are forced to zero. The one part of the Rabiner system that I've not implemented is the vector π of initial state probabilities.

The complete program used in this analysis has not been posted at the KEDS web site because it contains a rat's nest of poorly documented *#if ... #endif* blocks that allow all of the various analyses reported in this paper to be done within a single program. With that caveat, the code is available on request.

being generated by a particular HMM is nearly instantaneous. The program requires about 1 Mb of memory for a system using 45 codes, 6 states and 100-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. Consistent with the CASCON and SHERFACS approaches, the models I estimated used 6 states.

Results

LRL versus LR models

I repeated the experiments in Schrodt (1998) for discriminating between nonwar and war BCOW crises (translated into WEIS codes) using both the LRL model and a "circular" LRL model where the system can move from State A to State F and vice versa. As noted in Table 1, the accuracy of the LRL models is slightly greater than the LR model. Except in the nonwar split-sample test, the incorrectly classified cases were the same outliers found in Schrodt (1998). In separate tests not reported here, I found that the difference between the alphas for the nonwar and war LRL models correlates very highly with the Goldstein conflict scores for ISR>PAL, ISR>LEB and SYR>LEB; this behavior is consistent with the LR model.

Table 1. Number of BCOW cases correctly classified by models

	LR Model	LRL Model	Circular Model
Nonwar crises, split sample (N=16)	10	11	14
War crises, split sample (N=17)	15	15	15
Nonwar crises, full sample (N=31)	30	30	30
War crises, full sample (N=26)	23	24	24

Beyond correctly classifying a few additional cases, however, the LRL models did not show any clear advantages in discrimination over the LR model; this was contrary to my expectations. In particular, the classification distance—measured by the difference in alphas for the war and nonwar HMMs—was not necessarily higher for the LRL model, either for individual cases or in total. This differs systematically, however: the total discriminating distances for the nonwar cases are Circular > LRL > LR whereas in the war cases they are LR > LRL > Circular. This is presumably because the war crises have a clearer progression of events—peace to war to peace—whereas the nonwar crises may go through several cycles of escalation and de-escalation.[22]

A second difference between the LR and LRL models is that the variation in the maximal HMM found by the Monte Carlo procedure estimates is much greater. The LR models in Schrodt (1998) show a fairly consistent structure with high recurrence probabilities in five or six of the states of a 6-state model. The LRL models, in contrast, display a much wider variety of parameter combinations. For example, a common pattern in the transition probabilities is to have two adjacent states with very low recurrence probabilities but a high probability of going to the other state: in other words a pattern such as

$$\begin{pmatrix} 0.86 & 0.14 & 0 & 0 & \ldots & 0 \\ 0.28 & 0.01 & 0.71 & 0 & \ldots & 0 \\ 0 & 0.91 & 0.01 & 0.08 & \ldots & 0 \\ \ldots & & & & \ldots & \ldots \end{pmatrix}$$

In this case, the second two states are effectively acting as a single state with a high recurrence probability, but the two states rapidly oscillate in a BCBCBCBCB... pattern. The existence of these patterns also implies that fewer than six states may be required.[23]

In order to further explore the distribution of the estimates of models, I computed the mean and standard deviation of the parameter estimates on 2048 Monte Carlo experiments with the LR

[22] In a couple of cases, the Circular model estimated on the war cases ended up with zero estimates for some transition probabilities, thus forcing the model to be LR once it got into a certain set of states. This did not occur in the nonwar cases, at least in the HMMs I examined.

[23] Alternatively, these oscillating states may be accurately reflecting a true feature of the data: tit-for-tat behavior. The example above is a simplified version of States D and E in the P77 model discussed below; the actual recurrence probabilities are 0.0034 and 0.0002. If one looks at the D and E vectors, there are 14 symbols with observation probabilities $b_{kj}>0.01$. Twelve of these—corresponding to WEIS **02, 03, 06, 12, 21** and **22**—occur in symmetric pairs (e.g. **06** and **28**) for ISR>LEB and LEB>ISR, and in ten cases the differences between the b_{kj} and $b_{k(j+22)}$ have opposite signs. The remaining case is **22**/**44**, where both differences are positive. Finally, symbols **18** and **39** almost form a pair with the same sign, which can be interpreted as Israel "demonstrates" and LEB "threatens", possibly reflecting an actor-dependent difference in the wording used in Reuters reports. All of these patterns are consistent with the LRL model capturing closely linked reciprocal or tit-for-tat behavior—quite possibly reported in a single story—in the event data stream.

and LRL models. This revealed several interesting characteristics. First, in the LRL model, the mean prior-state, recurrence and next-state probabilities are nearly equal in States B, C, D and E (the averages are 0.31, 0.34 and 0.35 respectively); in States A and F the recurrence probability averages 0.54. In the LR model, the mean recurrence probability for States B, C, D and E is 0.66—suspiciously close to exactly 2/3—though for State A it is 0.92. The standard deviations mirror this: they are consistently around 0.25 for the LRL model and 0.22 for the LR. This implies that the variance of the LR estimates are substantially smaller in proportion to the mean probabilities, a ratio of about 3 for the recurrence probability of the LR compared to the 1.4 for the LRL, but those variances are still very high.[24]

The **B** matrices of symbol observation probabilities do not show the equiprobable behavior of the transition matrices, but in most cases the standard deviations are less than the mean values. The exception to this is the nonevent **00** in both models, and the force event **22/44** in the LRL model. In general, the standard deviations of the symbol probabilities are higher for later states (D, E, and F) than for earlier states, and the standard deviations tend to be less in the early states of the LR model than in the LRL model. These characteristics are consistent with the behaviors one would expect from the models, but the magnitude of these differences is relatively small. In short, except for the low variance of the recurrence probability in State A of the LR model, one cannot really argue for one form of the model over the other based on the distribution of the parameter estimates.

A second difference between the LR and LRL models is that the first state does not necessarily correspond to the background frequency of events. In the case of the BCOW crises, this is probably due to the fact that the sequences begin with some triggering sequence of events setting off the crisis, then it frequently settles back into a quiescent period (or periods) before rapid escalation occurs. Because the LRL model, unlike the LR model, can go back to an earlier

[24] Also, of course, this is reversed for the LR transition probability, which is just a linear function of the recurrence probability and thus has the same variance. Because the probabilities in the LR models are distributed across two states, whereas in the LR model they are distributed across three, it is difficult to compare the variances.

state, State A can be used for escalation (in other words, have relatively high symbol probabilities in the WEIS **11** to **22** range) while later states can be used for the background, where the **00** nonevent is most probable.

The upshot of this analysis is that the LRL model is somewhat more accurate, and it is definitely more flexible, but it does not provide a dramatic improvement over the LR model. The remainder of this analysis will be done with the LRL model, which seems to represent a compromise between the restrictions of the LR model and the excessive generality of the Circular model (in particular, the early states of the LRL model are more likely to correspond to the escalation phases of a crisis, whereas in the Circular model any set of states could correspond to the escalation). However, in many applications the LR model might be sufficient.

Early Warning using the Fit of an HMM

The next set of calculations was designed to determine whether the HMMs could be used to develop an early warning model by using analogies. I first identified six months in the ISR-LEB data series that involved TFT conflict; this was defined as a month that included 5 or more cases where there was a use of force (WEIS 22) by one party followed by a use of force by the other party within two days. These are the "templates" for the behaviors I am trying to identify and predict. The template months are

July 1981 [7]	May 1988 [6]	February 1992 [6]
October 1992 [7]	July 1993 [12]	May 1995 [5]

where [n] gives the number of TFT events in the following two months. These choices of templates are deliberately somewhat arbitrary, as the objective of this exercise is "learning by example." In addition to the template model, I also computed a "background" HMM that consisted of the 100 events prior to each of ten randomly chosen dates.[25] The background model is likely to be necessary because the fit of the HMM is very sensitive to the number of

[25] Literally random: the Excel random number generator was used to produce these.

non-events (see Schrodt 1998) and I anticipated that only the *difference* between the fit of the template and background HMMs would give meaningful results.

In contrast to the BCOW test, which examined directed behavior, the activities of both parts of the dyad were included in the model. This is required to identify TFT behavior, because the events in only one directed dyad are insufficient to distinguish between unilateral behavior and reciprocal behavior. This was done by recoding the LEB>ISR events with codes **23** through **44**, corresponding to the original WEIS codes **01** to **22**. If no event occurred with either dyad, the **00** nonevent was assigned to the day. The resulting model contains 45 event codes (2*22 + 1).

The early warning sequence for each template was the 100 events *prior* to the first day of the template month.[26] This is again a bit sloppy, as the actual outbreak of TFT violence does not necessarily occur early in the month. However, the objective of this exercise is early warning and I am trying to model the period *leading up to* the initiation of TFT violence, not the violence itself. This gets around the obvious criticism of the BCOW tests: it is easy to distinguish crises that involve wars from those that do not if you've got the entire sequence. In this test, we *do not* have the TFT sequence in the templates, only the events leading up to it (although these often involve uses of violence by one side or the other, just not a TFT sequence as I've defined it).

The alpha series for the fit of each month in the time series is generated by taking the 100 events prior to the end of the month and calculating the probability that this sequence was generated by the model. According to the underlying theory of HMMs, we should see a correlation between the fit of the template HMM—or the difference between this fit and the background model—and the TFT series. Figure 4 shows an example of the difference of the two series and, for comparison, the Goldstein scale.

[26] Why 100?—because I have ten fingers... The length of the warning sequence is a free parameter and other values might work better, depending on the application. I did some experiments early in the research with sequences of 50 and 200 events in addition to the 100 event length; the results were roughly comparable but 100 appeared to produce somewhat better cross-correlations. Given the vagaries of timing in this region—for example the effects of de facto unilateral cease-fires during various religious holidays—it is unlikely that the model will be very sensitive to the length of the sequence.

Figure 4. Time series of difference between template and background alphas with Goldstein scores

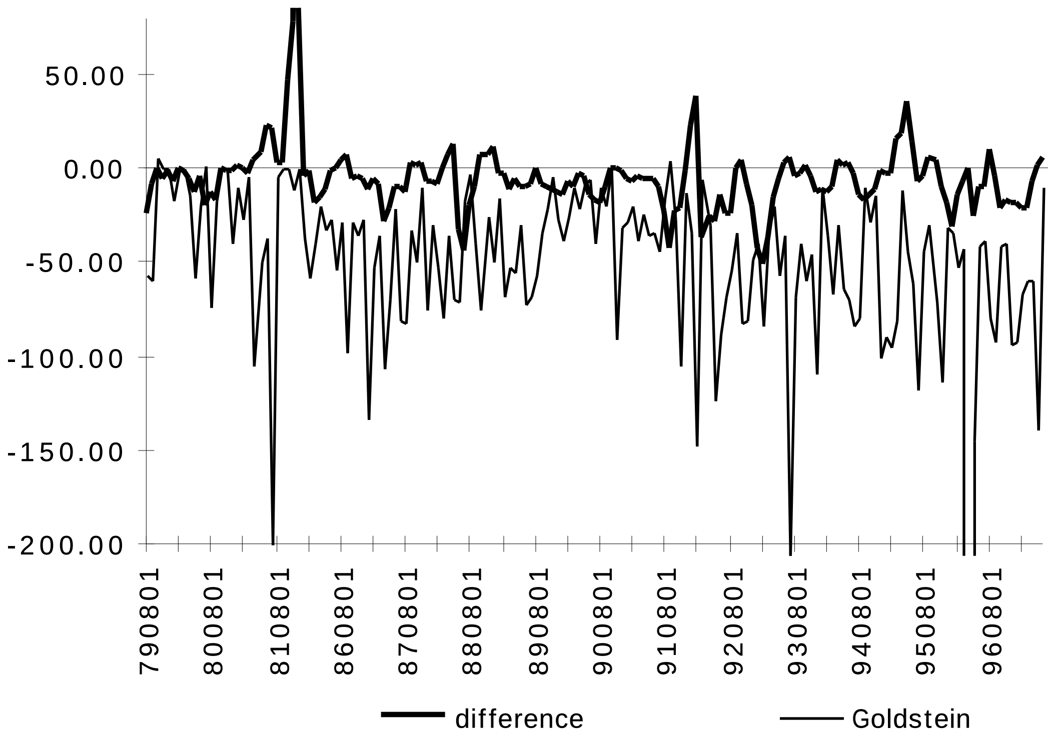


Figure 5. Cross-correlation of TFT with background and template model alphas

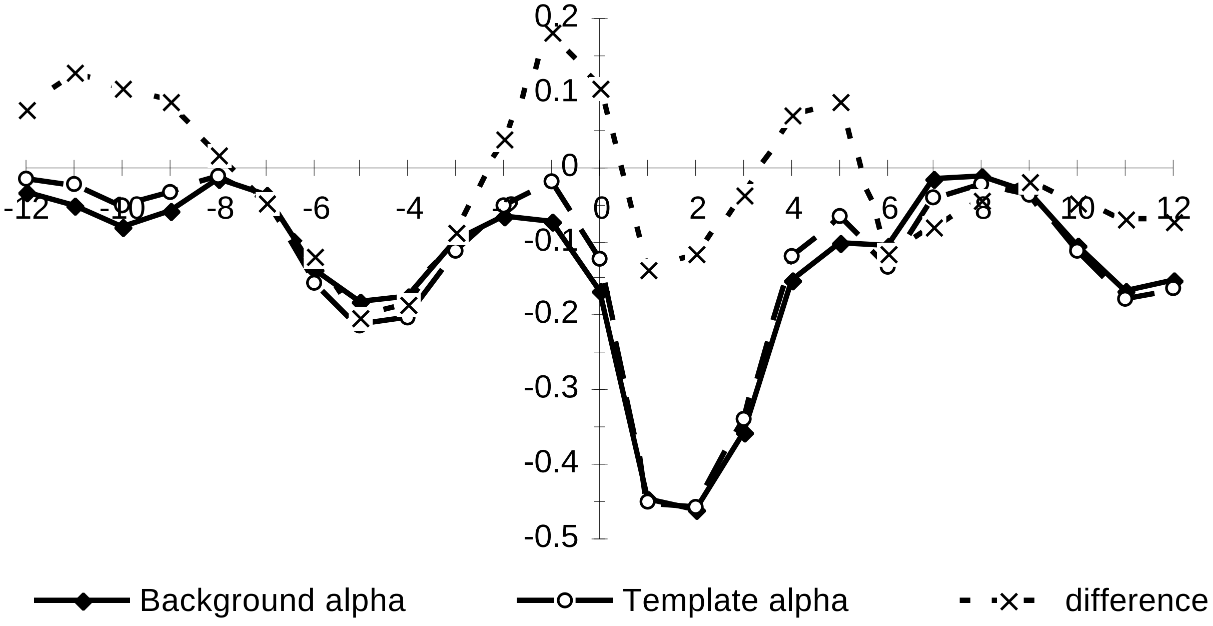
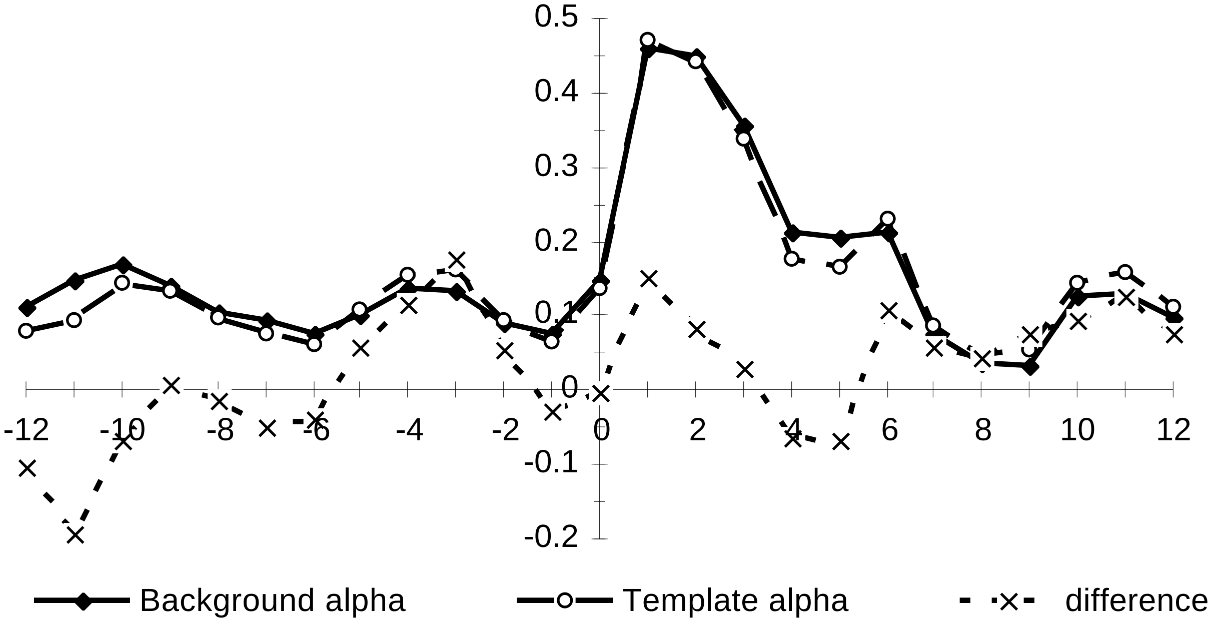


Figure 6. Cross-correlation of Goldstein scores with background and template model alphas

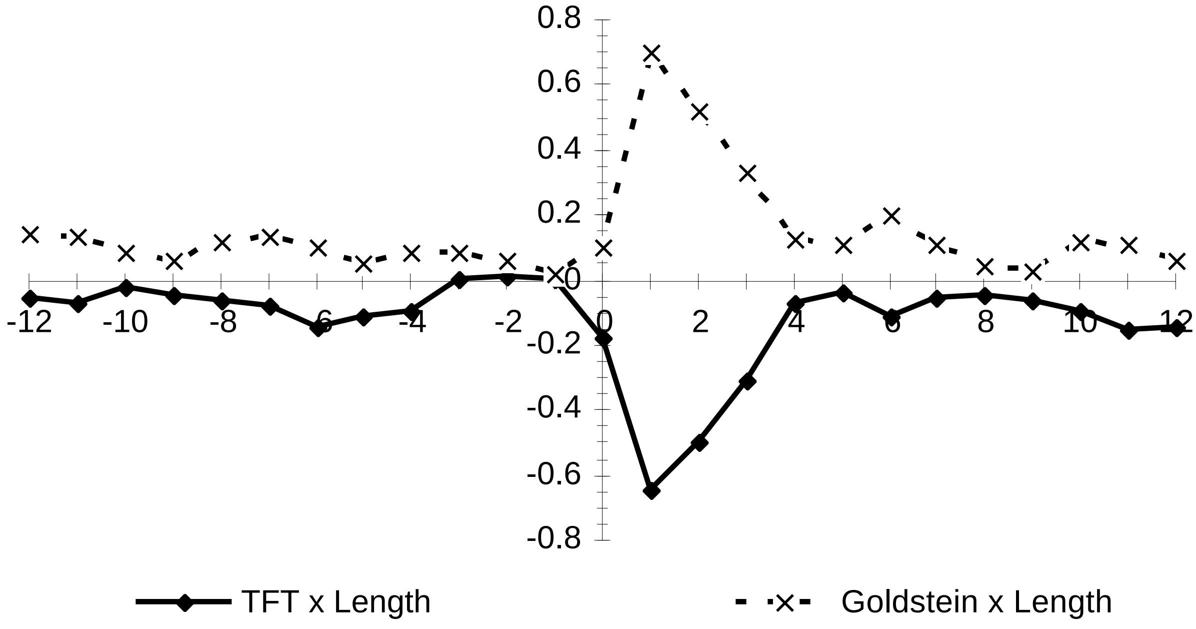


Figures 5 and 6 show the cross-correlation of the three measures—background alpha, template alpha, and difference—with the TFT and Goldstein measures. At first glance, these appear very promising as indicators—there is the expected high correlation at +1 and +2 months (when the 100-event sequence is likely to coincide closely with an actual TFT sequence) and a tantalizing early warning cross-correlation centered at about -4 months. However, the researcher's enthusiasm is quickly dampened on noticing that the cross-correlations of background and template models are almost indistinguishable. It is further dampened on noticing that the impressive cross-correlation patterns have the *wrong sign*!—if the theory is correct, one should see *positive* correlations with the TFT measure and *negative* correlations with the Goldstein scale, yet the opposite, quite conspicuously, occurs.

The researcher pauses, takes a deep breath, and—prospects of publication receding rapidly—contemplates changing his professional affiliation to Communications Studies.

The reason for both of these anomalous results is apparent in Figure 7, which shows a third variable—the length in *days* (as distinct from events) of each monthly sequence—cross-correlated with both measures. This is very similar to the cross-correlation curves in Figures 5 and 6, and accounts for both the sign of those curves and the fact that they coincide. In general, the alphas for both models decrease as the number of true events increases (and hence the length of the sequence in days decreases).[27] High negative values of the Goldstein score, and high positive values of the TFT score coincide with periods of high activity, hence the direction of the correlation. The impact of events versus nonevents so dominates the calculation of the alphas in these models (and with this data set) that it almost completely determines the fit.

Figure 7. Cross-correlation of TFT and Goldstein with length of sequence in days



One redeeming feature comes out of this otherwise useless calculation: In Figure 5 the *difference* between the background and template alphas shows a relatively high correlation (in the

[27] In a time-series plot, the background and template alpha curves are virtually indistinguishable.

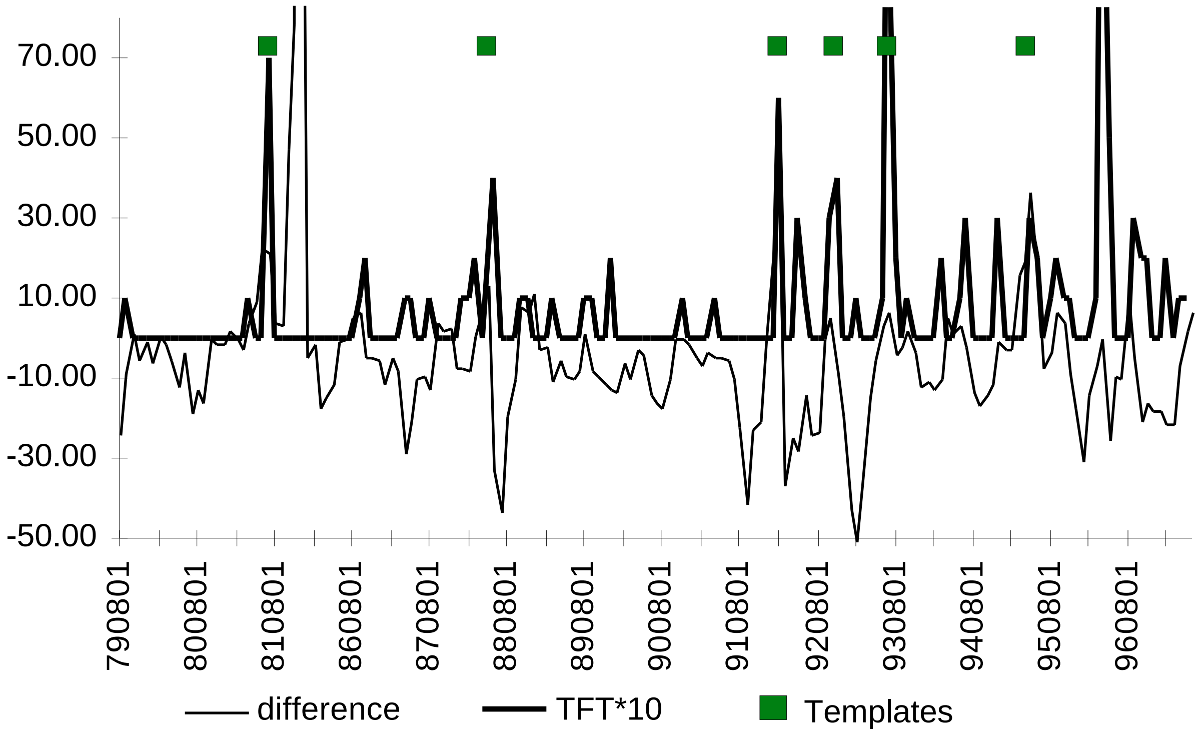
correct direction...) with the TFT series.[28] As expected, this peaks at a lag of -1 month. Several of the TFT sequences extend across two or more months, whereas the templates are based on sequences that terminate at the end of the month before the TFT sequence.

Figure 8 shows the alpha-difference (Δ) and TFT series. Using a threshold of Δ>2.0 and a lag time of [-2,-1,0] for the TFT events, only one false positive occurs—just prior to the 1982 splice—and generally months where the alpha-difference is greater than 2.0 occur contemporaneously with TFT months. All of the templates are identified correctly. There are large number of false negatives: Only about half of the TFT points are associated with Δ>2.0 points, and interestingly the model misses the major incidence of TFT violence involving Hizballah rocket attacks and the Israeli "Operation Grapes of Wrath" in the spring of 1996.

Figure 8 suggests that while the difference between the background and template alphas cannot be used for early warning, they still can be used for *monitoring* an event-data stream for a specific type of behavior that has been defined by a set of analogies. Thus, for example, if a human analyst identified a certain pattern of behavior that she thought was a good early warning indicator, an HMM-based system could monitor a set of event sequences (e.g., those produced by a machine-coding system processing the Reuters newsfeed) and alert the analyst when that sequence was observed. Similarly, if an analyst wanted to evaluate whether a specific type of event sequence could be used as an early warning indicator, it would be easy to search a set of event data to determine other instances of the sequence. HMMs are only one of several different ways to do this, but may well prove more robust and computationally efficient than the alternative techniques.

[28] This is not true for the cross-correlation with the Goldstein series in Figure 6: its pattern is consistent with a null model of zero correlation. This is a distinct contrast to the correlation between the Goldstein scores and the difference of the nonwar and war BCOW models, which are quite large.

Figure 8. Time series of difference between template and background alphas with the TFT scores



Early warning using hidden states

But wait, there's more!

There is an additional indicator derived from the HMM that might be useful for early warning: the hidden state of the system. As noted earlier, the Viterbi algorithm allows one to compute the sequence of hidden states that has the maximum likelihood for a given model and sequence of observations. If the theory underlying the use of the HMM is correct, we should see a system spending more time in the early states of the template model as it begins to approach a TFT event. The proportion of time spent in those early states could then be used as an early warning indicator.

In order to determine whether this would work, I used a two stage process:

1. Estimate an HMM using Monte-Carlo methods (64 experiments)

2. Repeat [1] a large number of times (e.g. 128 or 256) and select the HMM that maximizes the total cross-correlation at lags -2, -3 and -4 between the TFT measure and $Q_{BC,}$ the proportion of time the system spends in states B and C

In other words, this technique searches across a large number of estimated models to find one with the desired behavior. The search phase in [2] is necessary for two reasons: First, there are a large number of local maxima in the estimation even when Monte-Carlo experiments are used. Second, even if some state or states can serve as a leading indicator, there is no guarantee in an LRL model that these will be states B and C. Consequently I need to systematically search for those models where states B and C serve this role.

Figure 9 shows the cross-correlations for two such models, which I've labeled according to their total cross-correlation r's at lags -2, -3 and -4 and the Q_{BC} statistic. These two models provide exactly the early-warning indicator I have been seeking, although curiously the cross-correlation of Q52 peaks at a lag of -5 even though it was selected for earlier lags. The alpha curve, on the other hand, looks identical to that in Figure 5—even after selecting the model for the cross-correlation of Q_{BC}, alpha responds only to the number of nonevents in the sequence.

It is important to note that the cross-correlation patterns seen in Figure 9 are *not* typical—only a very small number of models show this behavior, and most have Q_{BC} cross-correlations near zero. I plotted the distribution of these cross-correlations over 512 estimated models and this distribution shows a high "ridge" around r = 0.0 at all lags and leads except -1 to +3, with the distribution falling off sharply outside the range $-0.1 < r < 0.1$.[29] Nonetheless, there is a clear "dip" in that ridge in the -3 to -6 lag range, suggesting that even globally a small but disproportionate number of models provide early warning.

At this point, the obvious question arises as to whether this is a real model , or merely a computer-assisted exercise of "beat the significance test." For starters, I would note that models with high Q_{BC} cross-correlations emerge quite consistently from this technique—in other words,

[29] This graph is posted at the web site—it is quite informative in color but hopelessly confusing in black&white.

I am presenting typical results of a search across 128 or 256 models, not the best results that I achieved over months of computation. The early warning models are rare, but not very rare.

To provide a stronger test, I estimated some models using a split sample design. I divided the data set in half at July 1990, then found the HMM that maximized the Q_{BC} x TFT cross-correlation for the data prior to July 1990. I then calculated the cross-correlations for only the second half of the data ($t \geq$ July 1990). In the split-sample, the sample size is around 70 so the illustrative critical values of r for the utterly inappropriate significance tests are

p=0.10	0.198	p=0.05	0.235	p=0.01	0.306

Most of the same templates were used as before,[30] so the estimated model includes information from the second half of the data set, but the selection criteria on the model do not.

The results of this exercise are shown in Figure 10 for this model, which I've labeled P77. Consistent with the search algorithm finding true characteristics of political behavior in this region, the early-warning cross-correlation found in the first half of the data set is also found in the second half. In addition, the cross-correlations in the leading period are quite random. The model also provides a weaker early warning for the Goldstein scale, again consistent with expectations.

Figures 11 through 13 provide additional evidence that the P77 model is operating as we would expect from the underlying theory. Figures 11 and 12 show the cross-correlation pattern by the individual hidden state. As expected, the cross-correlations show a clear pattern of progressively later lag times, with two exceptions: State B actually lags behind State A—so the order of these has been reversed in Figure 12—and State F shows no cross-correlation at all (as noted below, States D & E are a coupled pair and hence are combined).

[30] Due to a minor bug in the program, the first template (July 1981) was replaced by the 100-event sequence ending in May 1997. Because May 1997 precedes TFT behavior in June 1997, it is still a legitimate template. If only all program bugs were this innocuous...

Figure 9. Cross-correlation of TFT with for Q_{BC} and alpha for Models Q52 and Q61

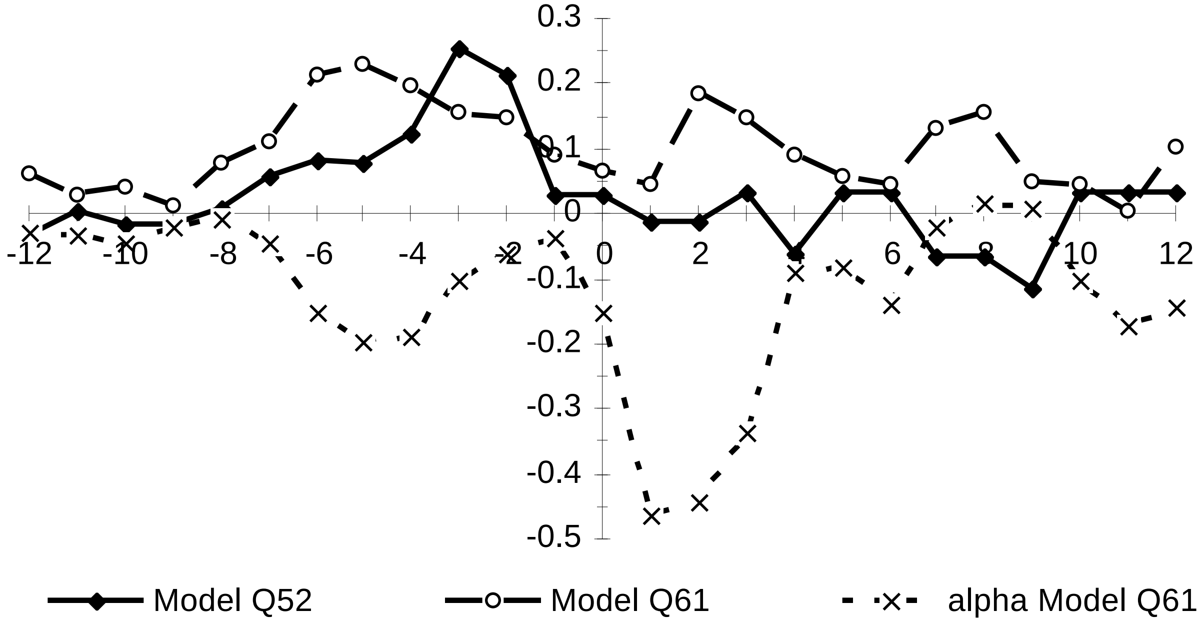
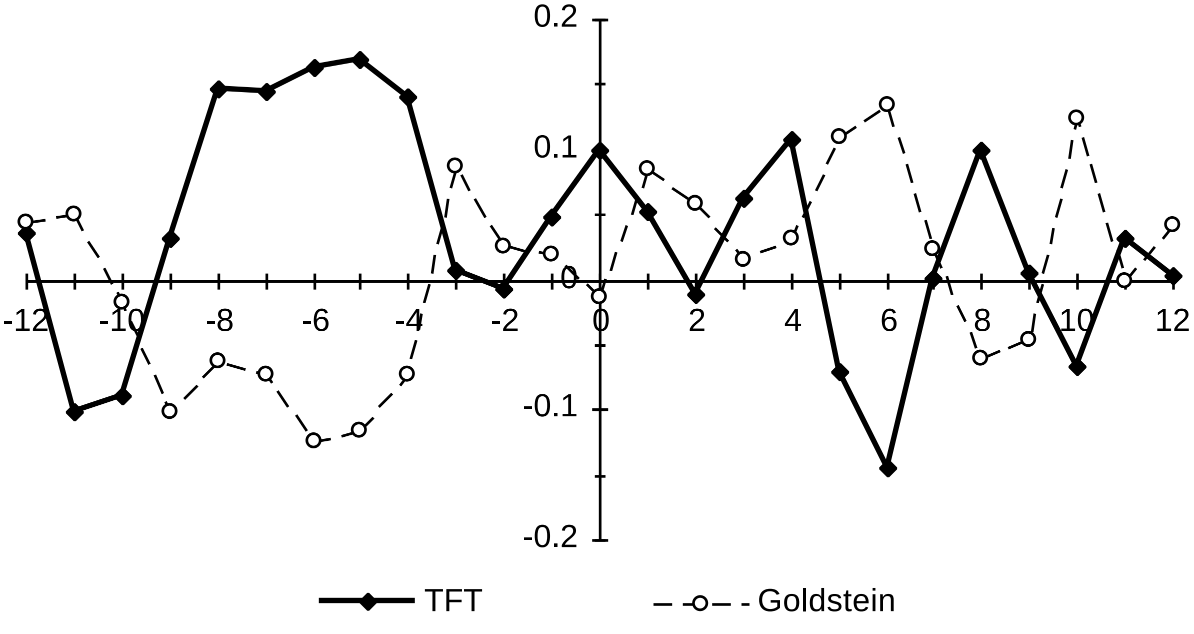


Figure 10. Q_{BC} cross-correlation for the P77 model



I repeated the split-sample tests in a random set of data that have a similar marginal distribution of events but no auto-correlation.[31] The search algorithm was able to find models that produced high cross-correlations at lags -2, -3 and -4 between the first half of the TFT series and the Q_{BC} statistic computed on the models generated from the random data, albeit at a slightly lower level (0.60 to 0.65 in the random data versus 0.72 to 0.77 in the ISR-LEB data). However, none of the other characteristics found in the split-sample tests on the real data are found: These models do not produce high cross-correlations at lags -2, -3 and -4 in the second half of the data, and there is no pattern to the correlations of the states other than those for which the models were explicitly selected.

Figures 12 and 13 show the structure of the HMM. Figure 12 shows the transition probabilities, which are characterized by high recurrence probabilities in States A and F, a very tight coupling between States D & E, and a looser coupling between States B and C. States B and C have relatively high recurrence probabilities, but are more likely to go between each other than to states A or D (though it is unclear how this relates to the fact that the cross-correlation of Q_A peaks between the cross-correlation peaks of Q_B and Q_C).

Figure 13 combines the symbol probabilities for both halves of the dyad—for example the "22" categories is the sum of the **22** category (ISR>LEB) and the **44** category (LEB>ISR); the **00** probabilities have been truncated. State A has broad range of cooperative and conflictual observation probabilities that may be a measure of an escalation phase before the outbreak of TFT conflict. The State D/E combination seems to involve a lot of negotiation, with relatively high probabilities in the WEIS **03** (consult), **06** (promise) and **12** (accuse) categories. True event probabilities in the B and C vectors are concentrated in the verbal conflict categories (WEIS **11** to **14**) without compensating consultations and promises, which may be why those states function as early warning indicators.

[31] The marginal distribution follows the ISR>PAL data—a program from Schrodt (1998) was used to generate this—but there are no theoretical reason to expect this would not generalize. As before, six templates were used to estimate the model. The TFT sequence from the ISR-LEB series was used in the cross-correlation test.

So, is the Q_{BC} early warning indicator sensitive to actual features in the data or is the cross-correlation pattern just luck? Arguing for the chance interpretation is the fact that models producing early warning are exceptional rather than typical. However, HMM parameter estimates are so underdetermined, both in terms of the large number of local maxima in the estimation procedure, and the structure of the parameters, that estimated models will always exhibit a variety of behaviors. Arguing for the reality of the model is the fact that several characteristics of the P77 estimates are consistent with the underlying logic model of precursors:

- It works in a split-sample test;
- The cross-correlation of the different states are consistent with their order in the model;
- The observation probabilities of the various states are distinct and plausible;
- These characteristics do not occur in a set of random data.;

Figure 11. Cross-correlation with TFT by states in the P77 Model

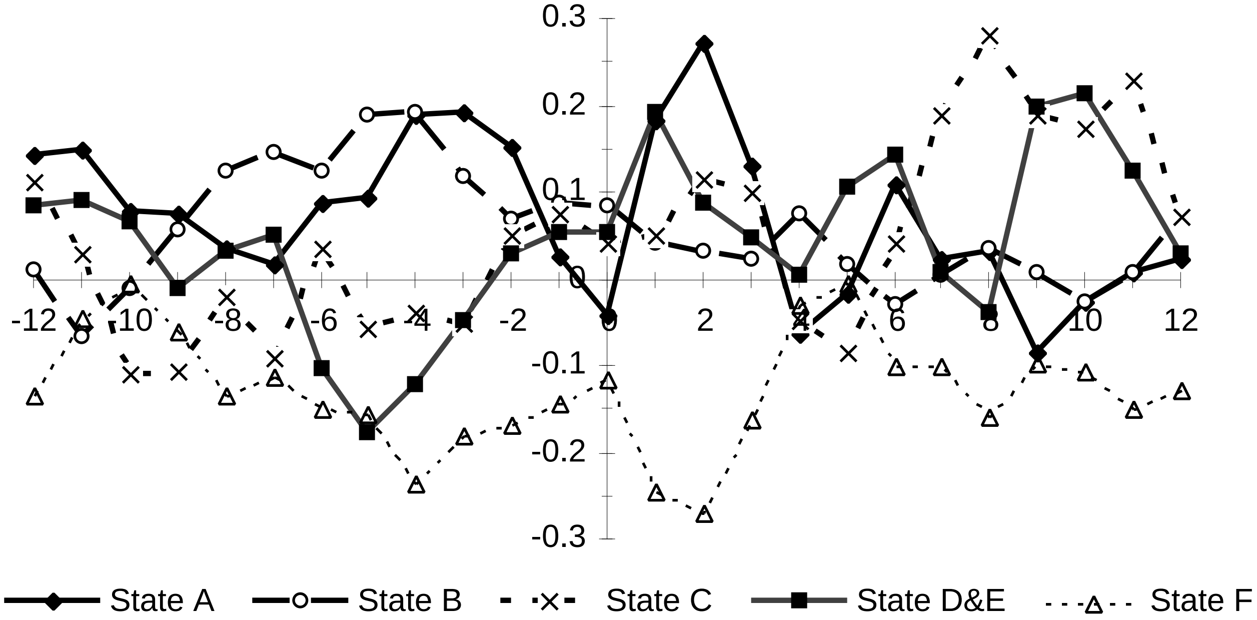


Figure 12. Positive cross-correlation with TFT by states in the P77 Model

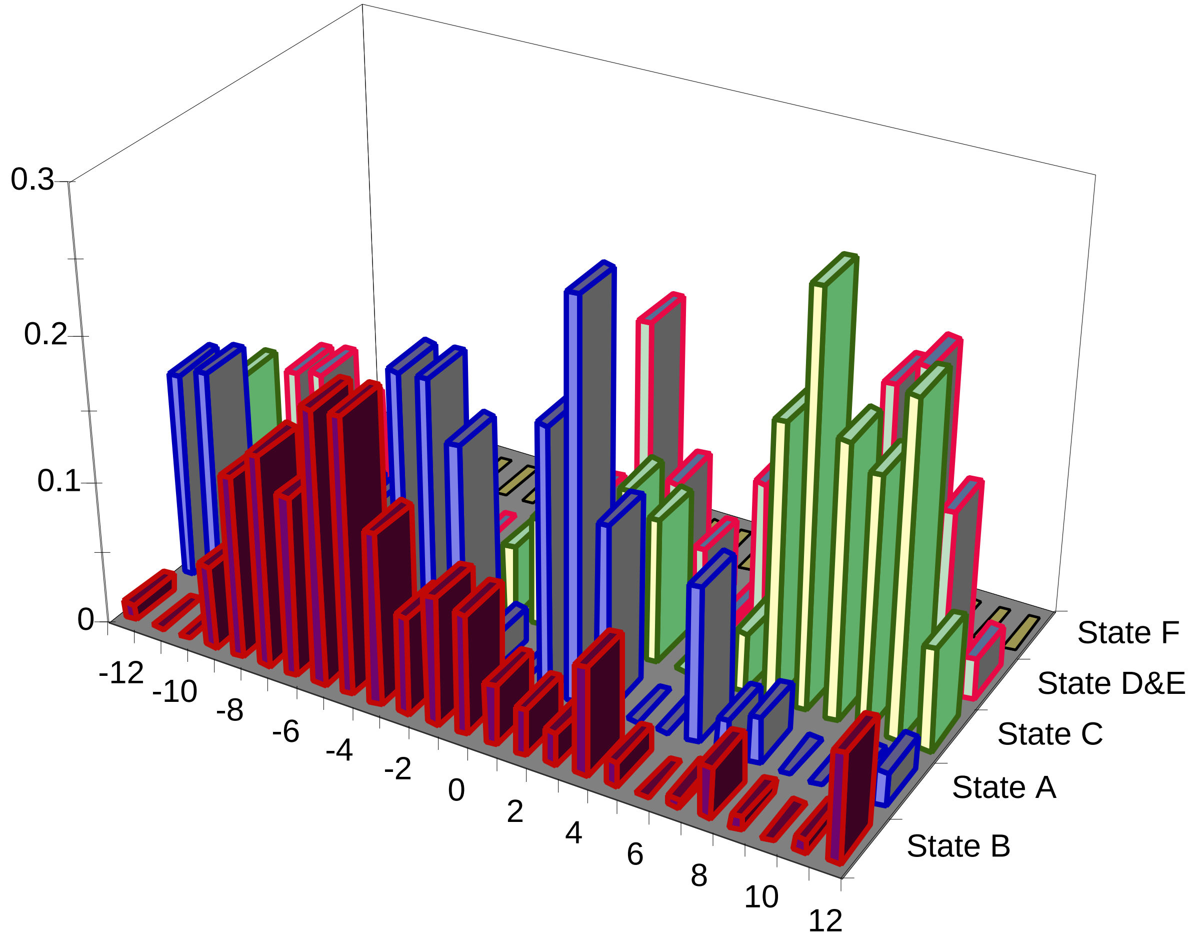


Figure 12. Transition probabilities in the P77 model

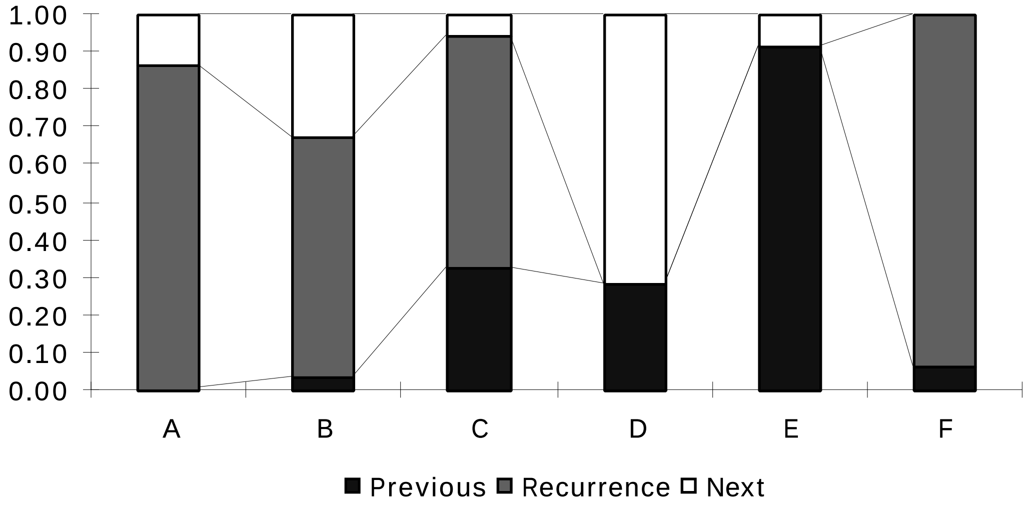
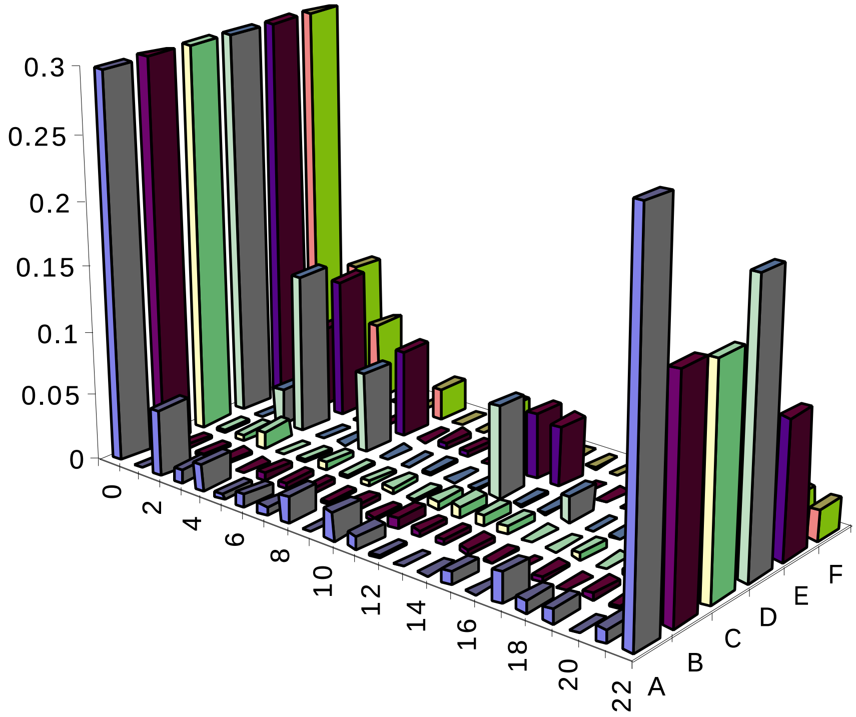


Figure 13. Symbol Probabilities in the P77 Model



Conclusion

This study of HMMs as models TFT conflict in southern Lebanon has produced tantalizing, but hardly conclusive, results. In this concluding section, I will address three issues. First, to what extent are HMMs likely to be effective as a general early warning method? Second, how could the estimation procedure be improved? Finally, what theoretical insights does this exercise provide about conflict processes?

Generality

My theoretical justification focused on the use of sequence-analysis methods in predicting a class of protracted conflicts characterized by co-adapted SOPs. From the practical standpoint of designing systems for early warning, the amount of conflict generated by protracted conflicts is not inconsequential—southern Lebanon is just one of a variety of cases—and merely being able to anticipate these cases would be a substantial improvement over the status quo.[32]

However, the HMMs were also effective in identifying, though not necessarily predicting, other conflicts found in the BCOW cases, and in generalizing the BCOW crises to measure conflict in the contemporary Levant. My sense is that this ability of HMMs to classify sequences of behavior will make them useful in other forms of early warning beyond the case of protracted conflict. Lebow (1981), Leng (1993) and others have suggested a number of common patterns in conflict escalation, and if these can be systematically characterized by event data sequences, they could be used as early warning indicators.[33] Models might also be made more robust by predicating them on the values of other variables—for example the presence of ethnolinguistic divisions, income inequality or the level of industrial development in an area—that may not be apparent from the events alone.

On the other hand, the coadaptation argument suggests that there are a couple of categories of conflict where sequence-analysis will *not* work for early warning (a proposition that could be tested). One category are situations where the conflict involves new organizations confronting each other for the first time. For example, I would be surprised if sequence analysis (or any other dynamic model) could predict the initial phases of the U.S.-Iranian hostage crisis, the initial phases of the Soviet intervention in Afghanistan or the UN intervention in Somalia. Second, sequence analysis is going to be less effective in dealing with situations where there has been significant

[32] If this anticipation resulted in effective action to head off the violence, it would eventually invalidate the model as well. The likelihood of encountering this "problem" seems remote...

[33] A remaining problem in the development of a practical monitoring or early warning system involves the tradeoff of Type I and Type II errors. At the Toronto early warning conference, I heard both of the following sentiments expressed (by different individuals in different organizations)

- "If the system gives me any false alarms, it will have no credibility" (low tolerance for Type I errors);
- "I don't care how many false alarms the system gives; just make sure it gets the real crises" (low tolerance for Type II errors).

Clearly a single system cannot satisfy both of these audiences. It should be possible to create systems with *differing* levels of sensitivity. A system that provides a simple "heads up" alert can afford to generate more false alarms than a system that provides a "start shipping $30-million of emergency food aid" alert, to say nothing of a "Send the Marines" alert.

strategic innovation, such as the 1967 and 1973 Middle East wars (the innovation occurring on the part of Israel in 1967 and Egypt and Syria in 1973).[34] These situations are extremely difficult for humans to anticipate—that's the whole idea!—and may be formally chaotic in the sense of systems dynamics.

From the perspective of developing a global early warning system, the problem is not just developing one or two indicators or models but rather developing a number of them. We are unlikely to be able to develop, with physics-like reductionism, a single theory to human conflict behavior because of the very substantial information processing capabilities of humans. Humans can be motivated to kill each other—and are, on regular occasions—for quite a wide variety of reasons. The protracted conflicts in southern Lebanon are somewhat similar, but hardly identical, to those involving Israel and the Palestinians—many of the same actors are involved, although not the same issues—but both are quite different from the protracted ethnic conflict in Rwanda and Burundi. Yet all these are protracted. This suggests that as an initial step, one would want to develop a number of contextually specific models based on analogies. Because the HMM is an inductive algorithm, this is easy to do once the appropriate event data have been collected.

Estimation

From the standpoint of estimation, the most troublesome aspect of the HMM approach is the high variance of the parameter estimates. This is apparently an inescapable characteristic of the technique: Baum-Welch estimation is a nonlinear method and there are no conditions that one can impose to identify the parameters.

That said, there are obviously more systematic ways to search for a global maximum (or at least a set of high local maxima) than the Monte Carlo method employed here: the structure of the problem is almost begging for the use of a genetic algorithm (GA).[35] In addition, Rabiner

[34] 1967 and 1973 are both examples of strategic military innovations but the same arguments apply to diplomatic innovations such as Camp David and Oslo.

[35] My thanks to Walter Mebane and Jas Sekhon for suggesting this. Programming a GA to operate on the HMM is a straightforward task—probably a couple hundred lines of code—but this paper is already past due...

(1989:273-274) indicates that in speech-recognition problems, the maximization is particularly sensitive to the initial values of the symbol observation probabilities in the **B** matrix, although not the transition probabilities in the **A** matrix. In an LRL model, however, the **A** matrix may also be sensitive to the initial parameter estimates—for example it would be helpful to force State A to be the background state. A GA may be considerably more efficient at finding optimal starting points than the Monte Carlo method.

Even with these modifications, it seems likely that a practical early warning model will require a certain amount of fine-tuning. For example, the 6-state model is probably roughly the correct size, but increasing or decreasing the number of states might improve the fit. Another arbitrary parameter that could be modified is the 100-event sequence length—why not 64 events or 128 events? Such tweaking apparently is quite common in the development of speech-recognition software, and would be recommended in the development of any early-warning system. On the other hand, a thin line separates "fine tuning" and "over-training", and ultimately the effectiveness of a model can only be assessed on data on which it has not been trained, whether in a predictive mode or with additional comparable cases.

Theory

On the theoretical level, the first thing that we gain from this approach is a reproducible method of evaluating whether sequence-based precursors exist. In particular, the HMM does not involve the human hindsight bias that plagues the evaluation of early warning indicators using qualitative historical comparison. If one takes the WEIS machine-coding dictionaries and the quantitative definition of a TFT event as a given, only four free parameters separate the early warning indicator from the Reuters text: the choice of templates, the sub-sequence length, the number of states in the system and the number of Monte Carlo experiments used in the estimation. All other parameters are determined from the data. The fact that machine coding removes the effects of human hindsight bias from the event coding further increases the possibility that the early warning indicators are real rather than determined by idiosyncratic coding and scaling decisions.

This in turn may also allow us to successfully distinguish actual protracted conflicts—conflicts resulting from coadaptive SOPs—from conflicts that are merely repetitive and result from the tails of the Poisson distribution. Protracted conflicts have precursors; Poisson conflicts do not. Again, the sequence-analysis approach—the indeterminacy of the HMMs notwithstanding—has the advantages of transparency (a term that I use deliberately instead of "objectivity") and reproducibility. The estimated HMM parameters should also provide some insight into what is important in a precursor and what is not. Additional theoretical guidance on this issue can be found in the event data, early warning and preventive diplomacy literatures.

There is also a level of analysis issue involved here. The behavioralist enterprise has tended to operate at a high level of generality. Its indicators—usually based on a realist conceptions of conflict and cooperation—are assumed to be more or less universal across cases and time. Human political analysts, in contrast, tend to want very specific information: not just the country where violence is occurring, not just the village, but which street in the village.[36] Unfortunately, these subtle nuances of individual cases of conflict are least useful in a generalized system for the prediction of international conflict. Event data and event sequence provide a middle-level between the two approaches—they are more specific than the highly aggregated indicators used, for example, by the COW research, but they typically do not go to the level of coding who called who a pig.

Once a number of contextually specific models had been developed and verified, then the next stage of theory development would be finding common characteristics of those models (again, the extant theoretical literature provides plenty of guidance on this issue). In addition, some of the contextual differences might be linked to exogenous static variables that could classify which models apply in which circumstances; for example, the State Failure Project uses static variables almost exclusively. But one needs first to demonstrate first that these conflicts are predictable in

[36] Really: this sort of information may be required to determine whether an outbreak of conflict had political content and hence might be a precursor to a wider escalation or was just a family feud and unlikely to escalate.

a contextually-specific sense before trying to generalize the models. There are generalizations to be made from studying apples and oranges, but fewer to be made from studying apples and bowling balls, or apples, Apple Records, and Apple Computers.

Bibliography

Allison, G. T. 1971. *The Essence of Decision.* Boston: Little, Brown.

Anderson, P.W., K.J. Arrow and D. Pines, eds. 1988. *The Economy as an Evolving Complex System.* New York: Addison Wesley.

Azar, E. E., and T. 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.

Bueno de Mesquita, B., D. Newman and A. Rabushka. 1996. *Red Flag over Hong Kong.* Chatham, NJ: Chatham House Publishers.

Bueno de Mesquita, B. 1981. *The War Trap.* New Haven: Yale University Press.

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

Casti, J. L. 1997. *Would-Be Worlds.* New York: Wiley.

Choucri, N. and T. W. Robinson, eds. 1979. *Forecasting in International Relations: Theory, Methods, Problems, Prospects.* San Francisco: W.H. Freeman.

Cimbala, S. 1987. *Artificial Intelligence and National Security.* Lexington, MA: Lexington Books.

Cyert, R. M. and J. G. March. 1963. *A Behavioral Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall.

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.

Gurr, T. R. and B. Harff. 1996. *Early Warning of Communal Conflict and Humanitarian Crisis.* Tokyo: United Nations University Press, Monograph Series on Governance and Conflict Resolution.

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

Hopple, G. W., S. J. Andriole, and A. Freedy, eds. 1984. *National Security Crisis Forecasting and Management.* Boulder: Westview.

Hinich, M.. 1997. Forecasting Time Series. Paper presented at the 14th Summer Conference on Political Methodology. Columbus, Ohio.

Hudson, V., ed. 1991. *Artificial Intelligence and International Politics*. Boulder: Westview

Hughes, B. B. 1984. *World Futures: A Critical Analysis of Alternatives.* Baltimore: Johns Hopkins.

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

Kauffman, S. A. 1993. *The Origins of Order.* Oxford: Oxford University Press.

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

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.

Laurance, E. J. 1990. "Events Data and Policy Analysis." *Policy Sciences* 23:111-132.

Lebow, R. N. 1981. *Between Peace and War: The Nature of International Crises.* Baltimore: Johns Hopkins.

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.

Maynard-Smith, J. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge 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.

Merritt, R. L., R. G. Muncaster, and D. A. Zinnes. 1993. *International Event-Data Developments: DDIR Phase II.* Ann Arbor: University of Michigan Press.

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. The Role of Stochastic Models in International Relations Research. In *Theories, Models and Simulation in International Relations*, ed. M. D. Ward. Boulder: 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. 1993. Rules and Co-Adaptation in Foreign Policy Behavior. Paper presented at the International Studies Association, Acapulco.

Schrodt, Philip A. 1994. Event Data in Foreign Policy Analysis. in L. Neack, J. A.K. Hey, and P. J. Haney. *Foreign Policy Analysis: Continuity and Change.* New York: Prentice-Hall, pp. 145-166.

Schrodt, P.A. 1998. Pattern Recognition of International Crises using Hidden Markov Models. in *Non-linear Models and Methods in Political Science*. ed., D. Richards. Ann Arbor: University of Michigan Press (forthcoming; the paper be downloaded from http://wizard.ucr.edu/polmeth/working_papers97/schro97.html)

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. Ann Arbor: University of Michigan Press.

Singer, J. D. and Wallace M.D. 1979. *To Augur Well: Early Warning Indicators in World Politics*. Beverly Hills: Sage.

Van Creveld, M. 1991. *Technology and War*. New York: Free Press.

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

Ward, M. D., ed. 1985. *Theories, Models and Simulations in International Relations*. Boulder: Westview Press.