Chapter Six
Hidden Markov Models

Note: This is a draft chapter from a text tentatively titled *The Analysis of International Event Data* (Philip A. Schrodt and Deborah J. Gerner). Please do not quote without permission. The chapter makes some references to material earlier in the book but otherwise should be readable. The material is largely from the following two articles:

Philip A. Schrodt. 1999. "Pattern Recognition of International Crises using Hidden Markov Models." in Diana Richards (ed.) *Political Complexity: Nonlinear Models of Politics.* Ann Arbor: University of Michigan Press.

Philip A. Schrodt. 1999. "Early Warning of Conflict in Southern Lebanon using Hidden Markov Models." Pp. 131-162 in Harvey Starr, ed. *The Understanding and Management of Global Violence: New Approaches to Theory and Research on Protracted Conflict.* New York: St. Martin's Press.

P.A.S. 19 Oct 99

6.1. Introduction

Both of the sequence analysis techniques introduced in Chapter 5 involve very large numbers of parameters. Given the complexity of human associative memory, and the large number of examples of political behavior that a human analyst can draw upon, this may be appropriate. However, by the more parsimonious standards typical of statistical modelling in political science, the models are very messy.

This chapter discusses a simpler sequence recognition technique, hidden Markov models. Following the design used in Chapter 5, we first demonstrate that these models are usually sufficient to discriminate BCOW crises that involved war from those that did not. Models based on the BCOW data are then used to study interactions in three dyads in the Levant—Israel>Palestinians, Israel>Lebanon and Syria>Lebanon—using a WEIS-coded event

data set covering April 1979 to February 1997. Despite the very substantial differences between the BCOW and Levant data sets in terms of coding procedures, historical time period, and underlying political behavior, the models that were estimated on the BCOW data show highly significant correlations with the level of conflict found in the Levant data, indicating that the hidden Markov models are successfully generalizing at least some of the characteristics of that behavior. Finally, the hidden Markov models are used to derive an early-warning indicator for tit-for-tat behavior in southern Lebanon.

6.2. Hidden Markov models

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

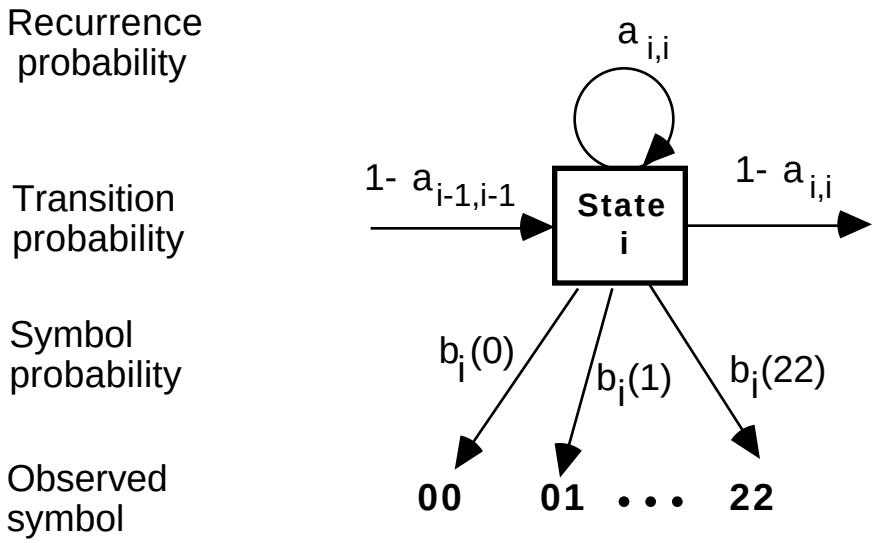
An HMM is a variation on the well-known Markov chain model, one of the most widely studied stochastic models of discrete events (Bartholomew 1975). Like a conventional Markov chain, a HMM consists of a set of discrete states and a matrix $A = \{a_{ij}\}$ of *transition probabilities* for going between those states. In addition, however, every state has a vector of *observed symbol probabilities*, $B = \{b_j(k)\}$ that corresponds to the probability that the system

will produce a symbol of type k when it is in state j. The states of the HMM cannot be directly observed and can only be inferred from the observed symbols, hence the adjective "hidden".[1]

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

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

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



[1] 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) .

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

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

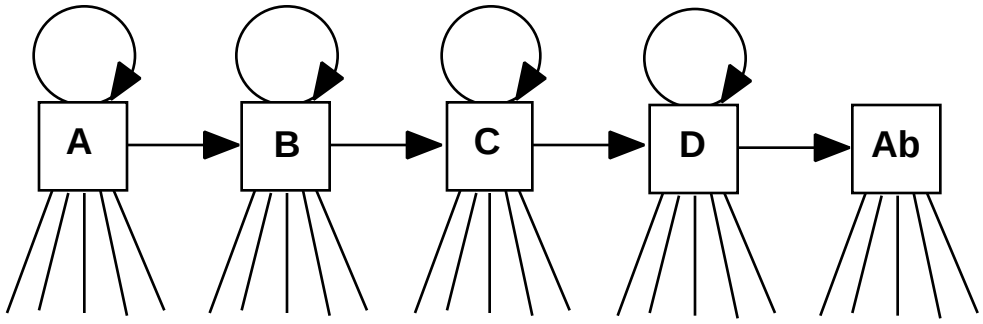


Figure 2. A left-right hidden Markov Model

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

In empirical applications, the transition matrix and symbol probabilities of an HMM are estimated using an iterative maximum likelihood technique called the Baum-Welch algorithm. This procedure takes a set of observed sequences (for example the word "seven" as pronounced by twenty different speakers, or a set of dyadic interactions from the BCOW crisis set) and finds values for the matrices A and B that locally maximize the probability of observing those

sequences. The Baum-Welch algorithm is a nonlinear numerical technique and Rabiner (1989:265) notes "the algorithm leads to a local maxima only and, in most problems of interest, the optimization surface is very complex and has many local maxima."

Once a set of models has been estimated, it can be used to classify an unknown sequence by computing the maximum probability that each of the models generated the observed sequence. This is done using 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.[2] Once the probability of the sequence matching each of the models is known, the model with the highest probability is chosen as that which best represents the sequence. Matching a sequence of symbols such as those found in daily data on a six-month crisis coded with using the 22-category World Events Interaction Survey scheme (WEIS; McClelland 1976), generates probabilities on the order of $10^{-(T+1)}$—which is *extremely* small, even if the sequence was in fact generated by one of the models[3]—but the only important comparison is the *relative* fit of the various models. The measure of fit usually reported is the log of the likelihood; this statistic is labeled α (alpha).

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

---

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

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

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

has the highest likelihood—for example the HMM corresponding to the numeral "three"—gives the best estimate of the number that was spoken.[4]

The application of the HMM to the problem of generalizing the characteristics of international event sequences is straightforward. The symbol set consists of the event codes taken from an event data set such as WEIS or BCOW. The states of the model are unobserved, but have a close theoretical analog in the concept of crisis "phase" discussed in Chapter 3. In the HMM, different phases would be distinguished by different distributions of observed WEIS events. A "stable peace" would have a preponderance of cooperative events in the WEIS **01-10** range; the escalation phase of the crisis would be characterized by events in the **11-17** range (accusations, protests, denials, and threats), and a phase of active hostilities would show events in the **18-22** range. The length of time that a crisis spends in a particular phase would be proportional to the magnitude of the recurrence probability a_{ii}.

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

[4] 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.

An important advantage of the HMM, particularly in terms of its possible acceptability in the policy community, is that it can be *trained by example*: a model that characterizes a set of sequences can be constructed without reference to the underlying rules used to code those sequences. This contrasts with the interval-level aggregative methods using event data scales such as those proposed by Azar & Sloan (1975) or Goldstein (1992). These scales, while of considerable utility, assign weights to individual events in isolation and make no distinction, for example, between an accusation that follows a violent event and an accusation during a meeting.[5] 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.

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: It is simple to prefix an HMM with a "background" state that simply gives the distribution of events generated by a particular source (e.g. Reuters/WEIS) when no crisis is occurring and this occurs in the models estimated below. A model can simply cycle in this state until something important happens and the chain moves into later states characteristic of crisis behavior.

There is a clear interpretation to each of the parameters of the A and B matrices, which allows them to be interpreted substantively. More generally, there is 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

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

those applications also indicates that the method is relatively robust. While there is always a danger in applying the *technique du jour* to whatever data on political behavior happen to be laying around, the HMM appears unusually well suited to the problems of generalizing and classifying international event data sequences, a task for which there are at present no particularly satisfactory solutions.

6.3. Discriminating BCOW War and Nonwar Crises

6.3.1. Data

The hidden Markov models were first estimated using the BCOW sequences studied in Chapter 4. The BCOW events were re-coded into WEIS categories according to the translation table listed in the Appendix. The four subsets of crises listed in Table 1 were analyzed.[6] The short names (e.g. *pastry*) correspond to the BCOW file identifiers. "Training" sequences were used to estimate the HMM matrices for the war and nonwar sequences; the system was tested with the remaining "test" sequences.

In contrast to the design in Chapter 4—which distinguished with separate codes whether events were occurring between the principal actors in the conflict, the principals and outside actors, and so forth—this study looks at simple directed-dyadic sequences involving the principal actors ("Side A" and "Side B") identified in the BCOW data set. This was done to provide comparability with a general event stream such as one generated by Reuters, where the "sides" of a conflict are not necessarily evident. The HMMs are therefore trying to model the general characteristics of "dyads involved in a crisis" rather than making distinctions based on the role of various actors.

[6] The BCOW crises not included in the studies in Chapter 4 were generally those whose length in events is very long (e.g. Suez or the Cuban Missile Crisis); or those that could not be easily classified into war or nonwar categories (e.g. Trieste). The HMM method is less sensitive to the length of a crisis so it should be possible to analyze the longer crises using the technique.

In order to record the passage of time in the various crises, days where nothing occurred were assigned a **00** non-event code; this is by far the most common "event" in the sequences. Sequences were coded from the beginning date to the ending date of the crisis as reported in the BCOW data set. When the BCOW data set reported multiple events on a single day, all of these were included. This is consistent with the structure of the hidden Markov model because the events observed on a particular day could occur as multiple observations from a single state of the model. In contrast, the parallel event sequences and Levenshtein metric in Chapter 4 assumed a strict temporal ordering. In those models, the fact that some days have multiple events while other days contain zero or one events complicates the estimation of the model. Dyads containing fewer than 20 BCOW events were not included in the analysis. Dyadic sequences typically contained about 30 to 70 actual events, although in a few cases there were over 200 events. When the nonevent days were added, most of the sequences contained between 200 and 300 events.[7]

The Levant data were WEIS-coded with KEDS from Reuters lead sentences obtained from the NEXIS data service for the period April 1979 through February 1997. The sequences that were tested were filtered of any of the WEIS codes that did not occur in the translated BCOW data (see Appendix) and a **00** nonevent was added for each day in which no events were recorded. As in the BCOW sequences, multiple events occurring in the same day are kept in the sequence.

6.3.2. Estimation Algorithm

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

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

handle the BCOW and WEIS sequences. The estimation algorithms were also generalized to handle estimation of a left-right-left model, and we implemented the Viterbi algorithm described in Rabiner (1989) in order to estimate the most likely state sequence.[8]

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

Consistent with the CASCON and SHERFACS approaches, the models we estimated used 6 states. Some additional experiments were done using a 12-state model and this produced much the same results.[9] Adding additional states to the models would strain neither memory nor

---

[8] The Meyers & Whitson code is clean, well-documented, and survived translation to run correctly the first time. Either our C code or their C++ code should 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, estimation equations were rewritten 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 is not implemented in the revised program 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 chapter to be done within a single program. With that caveat, the code is available on request.

[9] The 12-state models resulted in about a 4% improvement in the total likelihood in both the war and nonwar training cases. The classification accuracy is generally similar to that of the 6-state model—including the cases that were misclassified—with 3 errors in the war test cases and 6 in the nonwar. Curiously, only 6 of the states in the nonwar model and 7 of the states in the war model have high (>0.85) recurrence probabilities (including

computing time but, as noted below, a small number of states seems to be sufficient for the BCOW crises. Because the Baum-Welch algorithm is a numerical estimation method that is dependent on the initial values assigned to the probabilities, we ran at least 512 experiments with the matrices initialized to different random sets of probabilities, and then selected the model that had the highest total probability for the cases in the training set. A spot-check of the best-fitting results generated by separate runs of 128 experiments showed an extremely high correlation ($r>0.99$) between the alpha probabilities computed for each of the training cases, so the algorithm appears to be finding a global maximum in terms of these.[10] There is less convergence between the probabilities in the A and B matrices, although these are generally similar. This is presumably due to the fact that various combinations of recurrence probabilities and observed symbol probabilities can produce almost identical likelihoods for the training sequences.

6.3.3. Results

The HMMs estimated from the nonwar and war BCOW crises (translated into WEIS codes) are reported in Table 2 and Figure 3; Table 2 also reports the events in the transition vectors that have relatively high probabilities. The matrices are quite plausible, as are the differences between them; both models generated large recurrence probabilities on all six states. Both of the models successfully match all of their training cases —in other words, all of the nonwar training cases show a higher likelihood of fitting the nonwar model than the war model, and vice versa for the war training cases. The HMM thus meets the minimal requirements of any machine-learning

---

the absorbing state), indicating that most of the remaining states do not contribute substantially to the likelihood of the model. While the original 6-state configuration was chosen to mirror the Butterworth (1976) and CASCON (1989, 1997) schema, it seems to be close to optimal on the basis of the empirical tests as well.

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

approach: it can successfully classify its training cases. Because the set of 83 parameters used in the model (5 recurrence probabilities and 6 vectors of 13 symbol probabilities) are substantially smaller than the several thousand events in the training sets, it is unlikely that this fit is tautological.

The nonwar matrix begins with a series of cooperative events in state A. As conjectured, the distribution of the probabilities in this vector is close to that of the vector of marginal probabilities of events in the training set: the two vectors correlate with r=0.95 for all true events, and r=0.98 when the nonevent is included. The model then passes the time with nonevents in state B before escalating into conflictual events in state C. The transition between states B and C is likely to be either a consult, promise or request. State D is generates another sequence of nonevents, and then state E is dominated by just three event types: promise (probability 0.81), approve (probability 0.10) and agree (probability 0.08). State E rather conspicuously appears to represent the "dispute resolution" phase of the crisis. The absorbing state settles back into a mix of cooperative and conflictual (but nonviolent) events.

The war matrix shows a very different pattern. State A primarily generates nonevents, again closely reflecting the marginal probabilities of events in the training set: the correlation is r=0.82 for the true events and r=0.9995 when the nonevent is included.[11] State B involves a mix of mediating (consult, promise and request; total probability 0.37) and confrontational (accuse, demonstrate, seize and force; total probability 0.30) events. In state C, force has the highest probability. In contrast to the nonwar model, nonevents have high probabilities in the transition vector, indicating that the shift between states is signaled by a change in the distribution of events rather than a single triggering events. States D and E are dominated by nonevents and a mixture of conciliatory and confrontational events, and the absorbing state once is more dominated by force events. My guess is that states D and E are most likely the result of situations where the

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

BCOW data include a period of peace negotiations following the cessation of hostilities, whereas the absorbing state is used to model the cases where hostilities continue until virtually the end of the data (specifically the Schleswig-Holstein War and Italo-Ethiopian War). The presence of force events in the transition vectors of states D and E is consistent with this interpretation and the recurrence probability on state E is so high (0.9946; for state D it is 0.9858) that it could virtually serve as an absorbing state itself.

The results of the split-sample testing are reported in Table 3, which gives the α log-likelihood values for the fit of various dyadic sequences using the HMMs estimated on the training cases. The war model classifies somewhat more accurately than the nonwar model, but both models do quite well and the cases that are incorrectly classified are concentrated in a set of plausible exceptions rather than distributed randomly.

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

		A	B	C	D	E	Abs
recurrence probability		0.96	0.98	0.96	0.99	0.64	1.00
Event							
00	**none**	0.58	0.97	0.33	0.97	0.00	0.85
01	**comment**	0.02	0.00	0.02	0.00	0.00	0.00
02	**consult**	0.07	0.003	0.04	0.00	0.00	0.04
04	**approve**	0.04	0.003	0.07	0.006	0.10	0.20
05	**promise**	0.14	0.006	0.17	0.003	0.81	0.04
06	**grant**	0.00	0.00	0.005	0.00	0.00	0.00
07	**reward**	0.002	0.00	0.00	0.00	0.00	0.00
08	**agree**	0.005	0.00	0.005	0.00	0.08	0.005
09	**request**	0.07	0.002	0.14	0.004	0.017	0.03
12	**accuse**	0.04	0.007	0.08	0.006	0.00	0.01
17	**threaten**	0.002	0.00	0.005	0.00	0.00	0.00
18	**demons**	0.01	0.004	0.11	0.01	0.00	0.004

19	**reduce rel.**	0.00	0.00	0.005	0.00	0.00	0.002
21	**seize**	0.005	0.003	0.005	0.00	0.00	0.00
22	**force**	0.00	0.002	0.005	0.001	0.00	0.002
	transition events	03 (.23)	03 (.30) 05 (.33) 09 (.21)	03 (.20) 09 (.37)	00 (.57) 18 (.26)	00 (.36) 04 (.20) 05 (.22)	NA

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

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

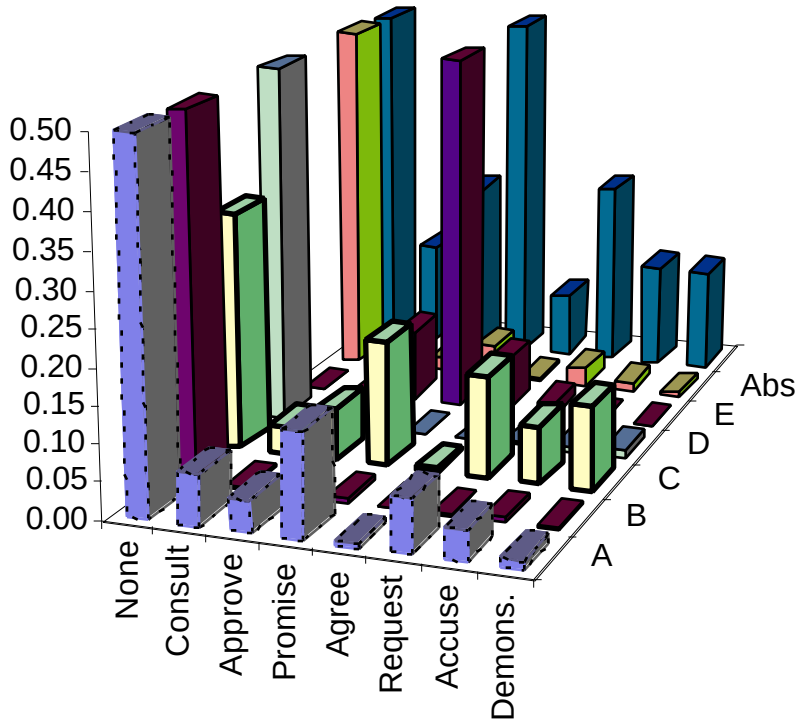


Figure 3a. HMM Event Probabilities: Nonwar crises

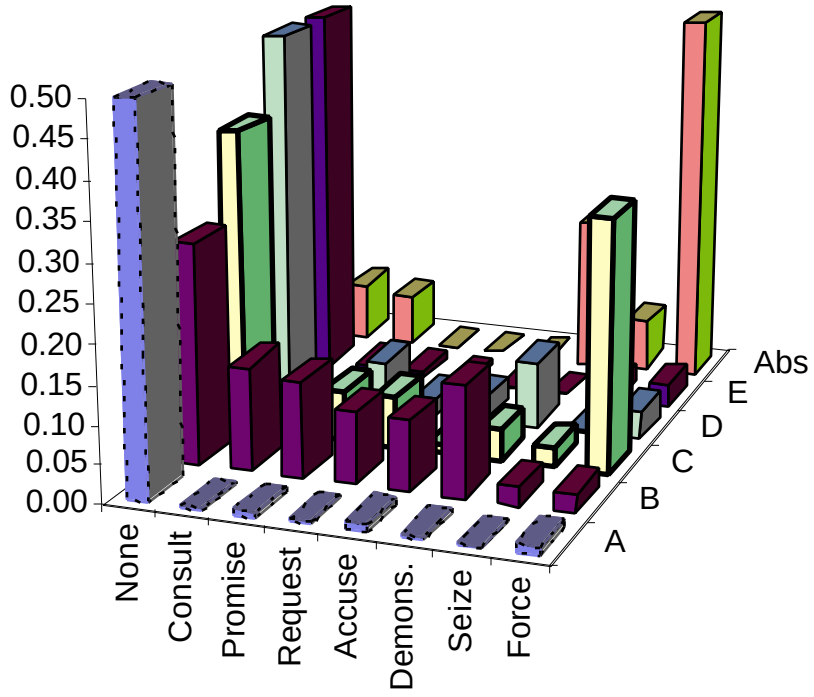


Figure 3b. HMM Event Probabilities: War crises

Table 3: Alpha values for the test cases

Nonwar Test Cases

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

War Test Cases

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

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

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

6.3.4. Increasing the allowed transitions: LRL models

The LR model assumes that a crisis has a linear development—once it moves to a new state, it cannot go back. An alternative conceptualization of the crisis process—and one more consistent with the informal literature—is that a crisis can go iterate through a number of periods of escalation and de-escalation before it is completed. This can be modeled in an HMM by simply allowing transitions to the left as well as the right, in other words, a left-right-left (LRL) model such as the 6-state model illustrated in Figure 2. In contrast to the LR model, every state is accessible from every other state. As before, we assume that crises that have a clear "beginning" and "end" and thus the sequences are assumed to start in State A.

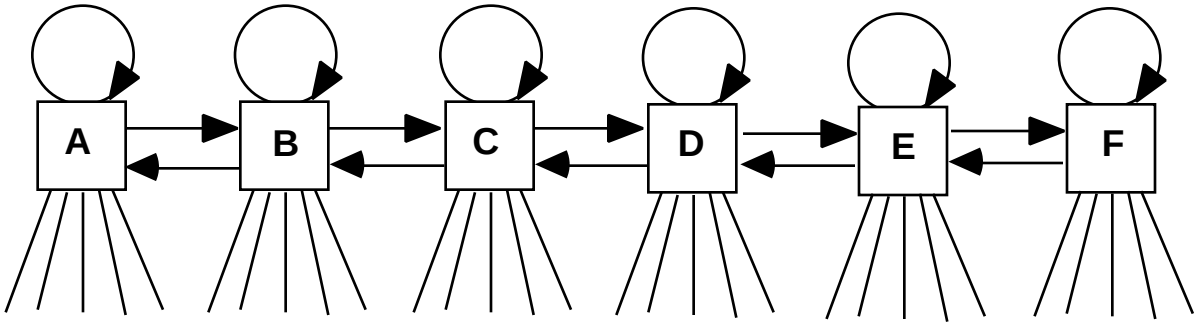


Figure 2. A left-right-left (LRL) hidden Markov Model

We repeated the experiments on the BCOW model 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 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 our 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.[12]

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 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

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

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.[13]

---

[12] 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 we examined.

[13] 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 $b_{k}j$ 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

In order to further explore the distribution of the estimates of models, we computed the mean and standard deviation of the parameter estimates on 2048 Monte Carlo experiments with the LR 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—although 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.[14]

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.

---

these patterns are consistent with the LRL model capturing closely linked reciprocal or tit-for-tat behavior—quite possibly reported in a single news report—in the event data stream.

[14] 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.

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 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.

6.3.5. Using the BCOW models to measure conflict in the Middle East

The next set of calculations determines whether the HMMs could be used to reveal anything about a contemporary political situation. The nonwar and war HMM models were first re-estimated using both the training and test cases.[15] Figures 4, 6 and 7 show the log-likelihood fit of the two models to three of the densest dyads from the Reuters-based Levant data set: ISR>PAL, SYR>LEB and ISR>LEB. The two lines below the X-axis are the alpha log-likelihoods; the line near the X-axis is the $\alpha_{war} - \alpha_{nonwar}$ difference. The WEIS sequences used

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

to generate the fit were generated by taking the 100 events prior to the end of each month. This sequence typically covers about two months, although it is shorter in times of intense activity. Because all of the sequences are the same length, their values can be compared over time.

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

The first thing that is conspicuous in the figures is that the nonwar and war alpha curves track each other very closely.[16] This probably reflects the effects of the presence or absence of nonevents; these are much prevalent in the BCOW dyads than in these politically-active Levantine dyads. Periods with a high intensity of activity—for example the Palestinian *intifada* and various Syrian and Israeli interventions in Lebanon—consistently show much lower alpha values than periods of low activity. This reduction in alpha is probably due in large part to the fact that actual events (as distinct from nonevents) have a low probability (see Table 3) in most of the states of both HMMs.[17]

[16] Similar results can also be generated using an LRL model.

[17] This may also be due in part to the crudeness of the BCOW to WEIS translation. For example BCOW contains a "continuous military conflict" code that we translated into a single WEIS force event. In fact, such codes presumably indicate multiple consecutive days of WEIS force events. Such sequences are common during the

For contrast, Figure 5 shows the alpha curves for a set of random simulated data that has the same marginal event probabilities as the ISR>PAL data set but no autocorrelation.[18] Three features are evident in this figure. First, as one would expect, the two curves are basically just noise—due to the 100-event sequence length, they are significantly autocorrelated at a lag of one month but beyond one month the autocorrelation pattern is consistent with white noise. Second, the war and nonwar alpha curves themselves are highly correlated ($r = 0.80$; $p<.001$). Finally, the alpha value for the war model is consistently higher than the value for the nonwar model, which is to be expected because around 20% of the events in this sequence are force events.

Figure 8 and Table 4 compare the difference in the HMM alpha log-likelihoods with the Goldstein-scaled time series for the Levant for the period August 1979 to October 1996.[19] Figure 8 shows a relatively close correspondence between the alpha-difference and the Goldstein score for Israel > Palestinian behaviors during most of the period. As noted in Table 5, the correlations between the Goldstein score and the difference between the HMM probabilities is highly significant for all of the dyads, and going as high as 0.67 for Israel > Palestinian behaviors during the period before the Oslo agreements (September 1993).

---

interventions in Lebanon and during the *intifada* but would have no BCOW counterparts given our translation rules.

[18] The marginal probabilities are:

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

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

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

[19] The Goldstein score has been divided by 4 to bring the two measures into scale with each other.

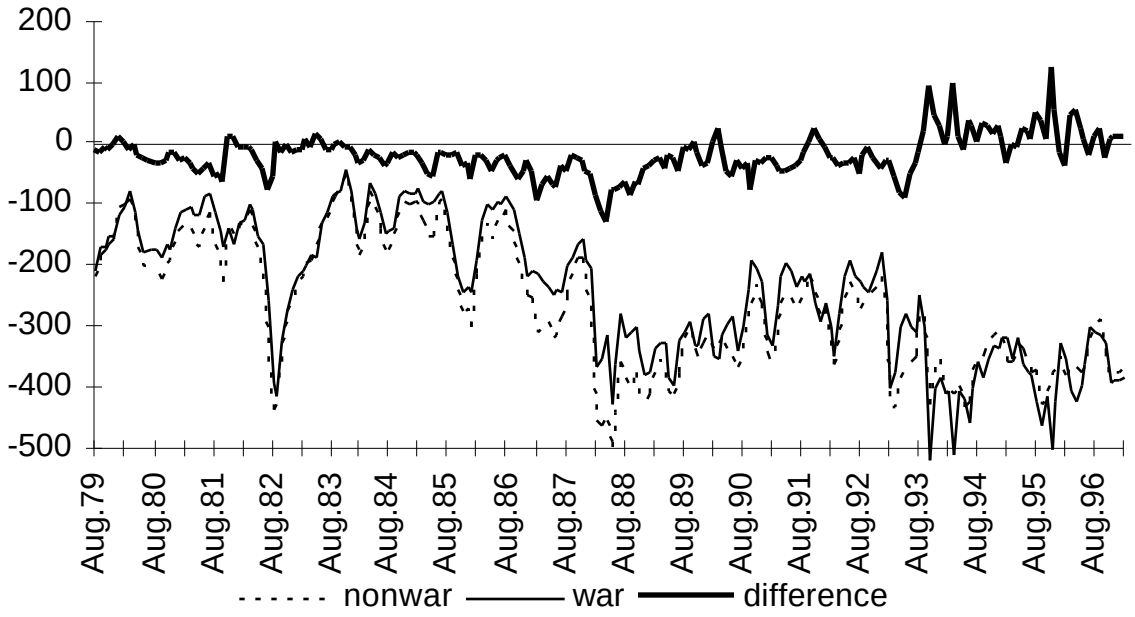


Figure 4. Alphas for Israel > Palestinians

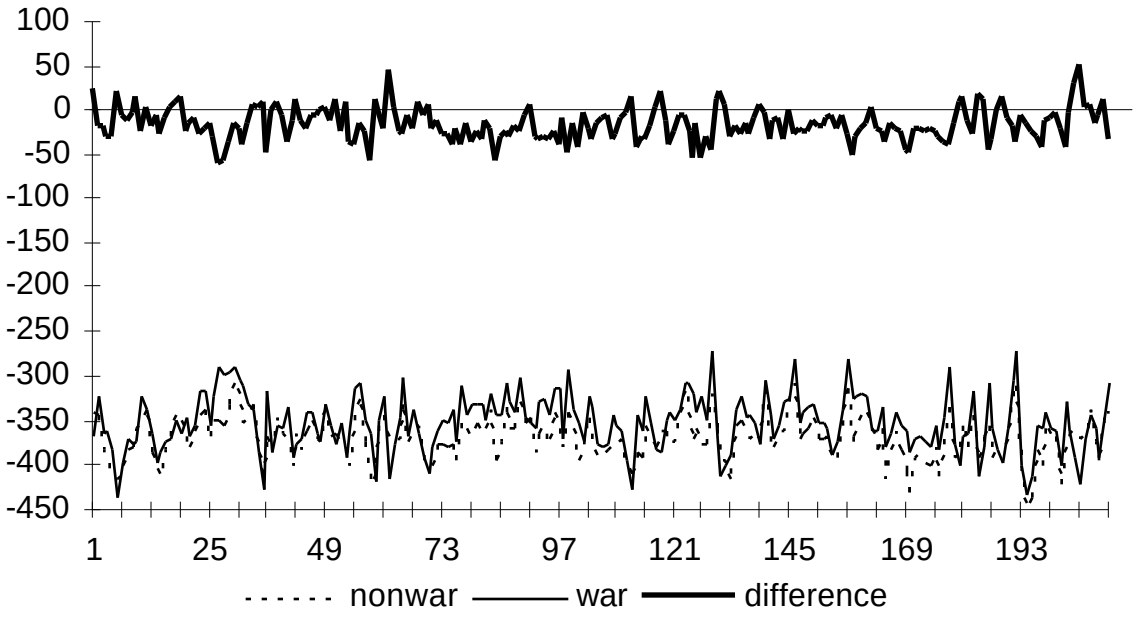


Figure 5. Alphas for random sequences

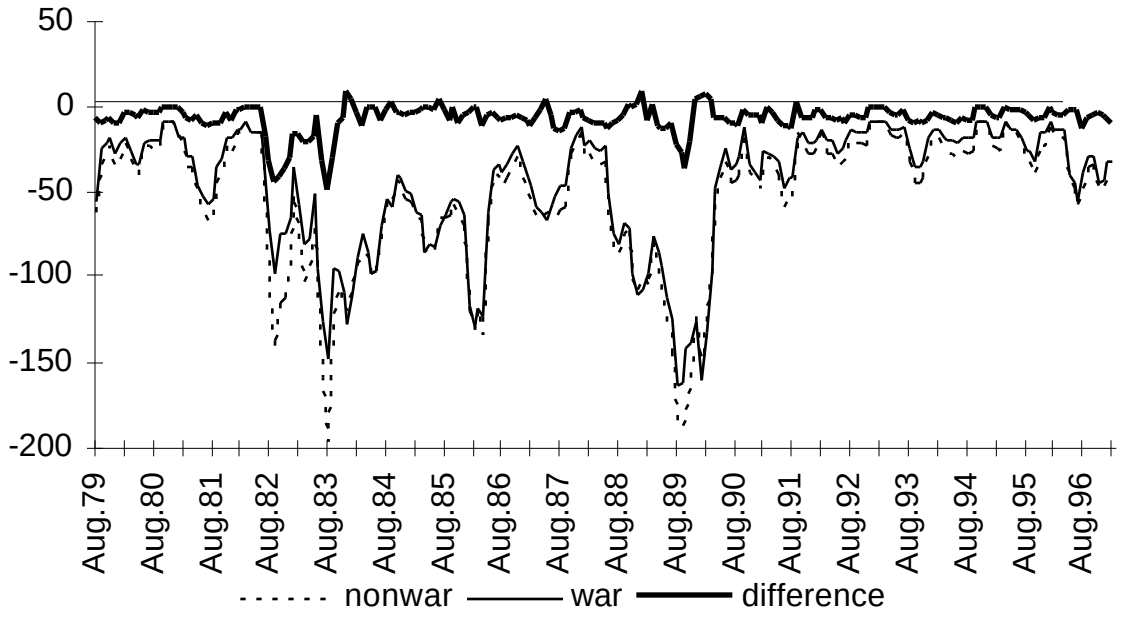


Figure 6. Alphas for Syria > Lebanon

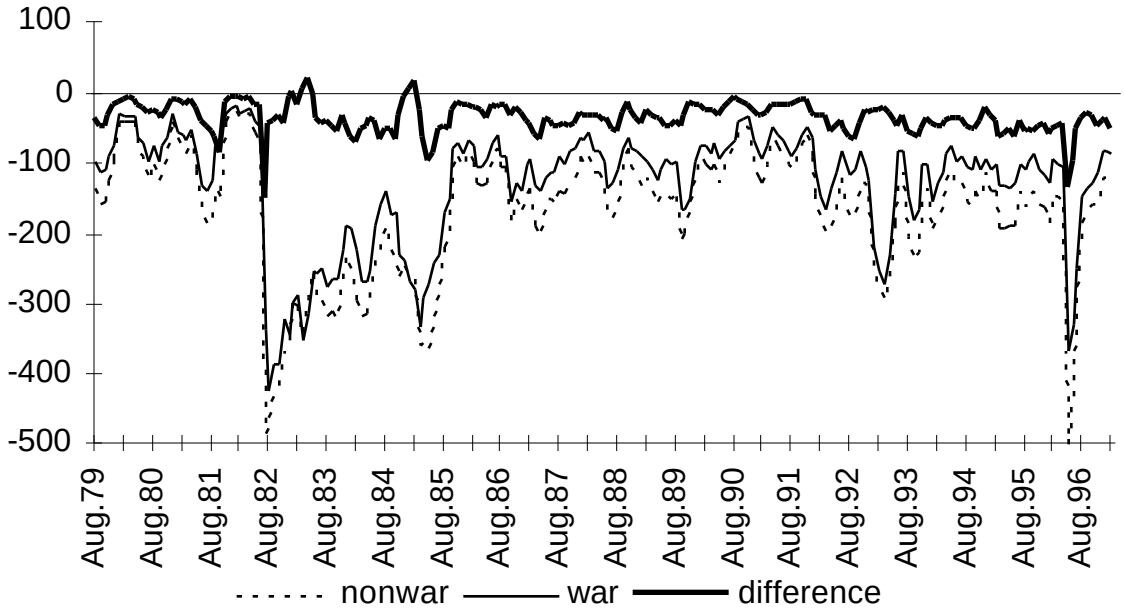


Figure 7. Alphas for Israel >Lebanon

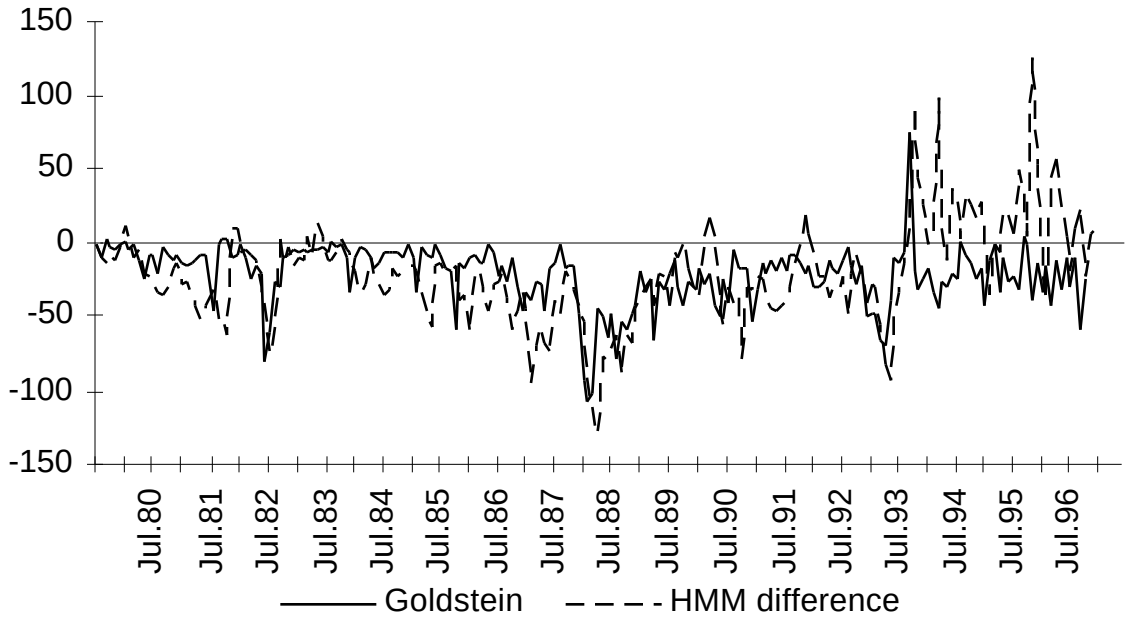


Figure 8**. Comparison of Goldstein a**nd HMM difference scores for I**srael > Palestinians**

Table 4: Correlation between Goldstein scores and HMM difference

Dyad	**N**	**r ***
Israel>Palestinians	206	0.49
Israel>Palestinians, pre-Oslo	170	0.67
Israel>Lebanon	206	0.55
Syria>Lebanon	206	0.40

*all correlations are significant at the 0.001 level

While the alpha-difference and Goldstein scores in Figure 8 generally track each other, particularly on major events such as the invasion of Lebanon and the *intifada*, there are a couple of interesting distinctions. First, the alpha-difference is somewhat more sensitive in measuring the level of conflict (in the sense of moving away from the nonwar model) than is the Goldstein score: for example this is conspicuous in the period prior to the summer of 1981 where there was considerable conflict between Israel and PLO militias then residing in southern Lebanon. Second, the alpha-difference is much more sensitive to periods of negotiations than is the Goldstein score. This is most evident in the post-Oslo period but can also be seen in a positive peaks in October-December 1991 corresponding to the beginning of the Madrid negotiations; the positive point in that November-December 1981 corresponds to the cease-fire between the PLO (in southern Lebanon) and Israel that was brokered by the United States; and the peak in March-June 1983 appears to correspond to a series of prisoner-exchange negotiations brokered by Austria.[20]

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

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

6.4. Early Warning using HMMs

Our final analysis focuses on forecasting the occurrence of 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. 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.

As discussed earlier in Chapter 4, 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"[21]: A conspicuously successful strategic innovation is unlikely to succeed twice precisely because it was successful the first time. More

[21] "...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).

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 adaptive agents, there is little point in trying to make predictions based on past behaviors.[22] 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.[23] 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

---

[22] 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.

[23] A detailed discussion of the concept of coadaptation is beyond the scope of this chapter, 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.

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 we've outlined above—antagonists fight, on repeated occasions, over the same issues, often over the same territory, and without resolution.

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.

6.4.1. Measures of Tit-for-Tat Conflict

The focus of our 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. 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.[24]

Two different predictive targets are being used: the number of TFT incidents, and the Goldstein-scaled score of the ISR>LEB conflict.[25] 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.

[24] 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 we 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).

[25] 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.

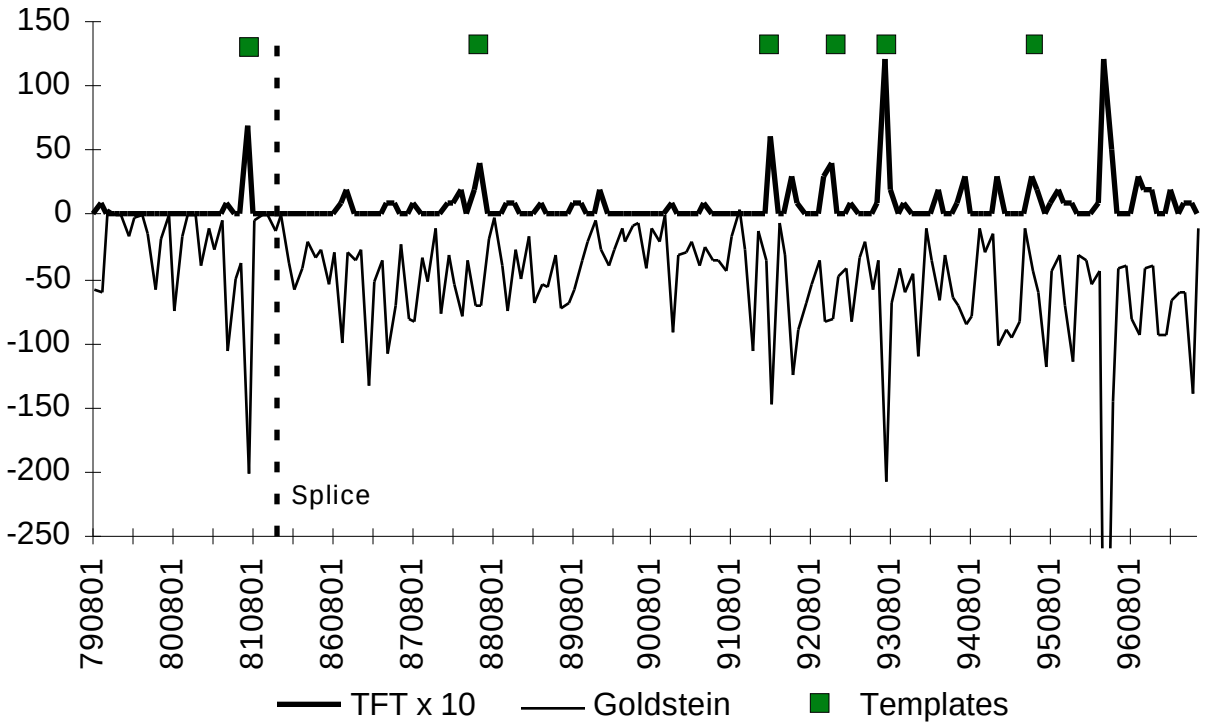


Figure 3. Time series of the TFT and Goldstein scores

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. 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.[26] A custom program is used to compute the cross-correlations appropriately despite the splice in the data set.[27] The resulting sample size is around 160 and the critical values of r for a two-tailed significance test are

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

[27] 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...

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.

6.4.2 Early Warning using the Fit of an HMM

To develop the HMM early warning model by using analogies, we 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 we are 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, we also computed a "background" HMM that consisted of the 100 events prior to each of ten randomly chosen dates.[28] The background model is likely to be necessary because the fit of the HMM is very sensitive to the number of non-events and we 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).

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

The early warning sequence for each template was the 100 events *prior* to the first day of the template month.[29] 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 we are 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.

[29] Why 100?—because we have ten fingers... The length of the warning sequence is a free parameter and other values might work better, depending on the application. We 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.

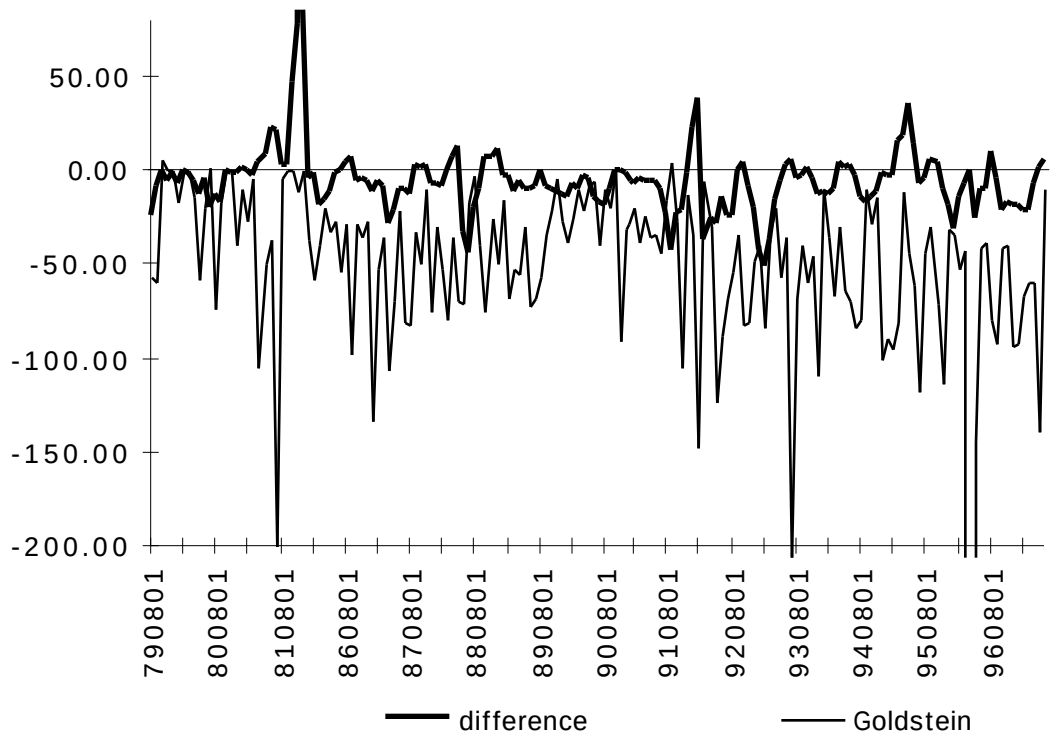


Figure 4. Time series of difference between template and background alphas with Goldstein scores

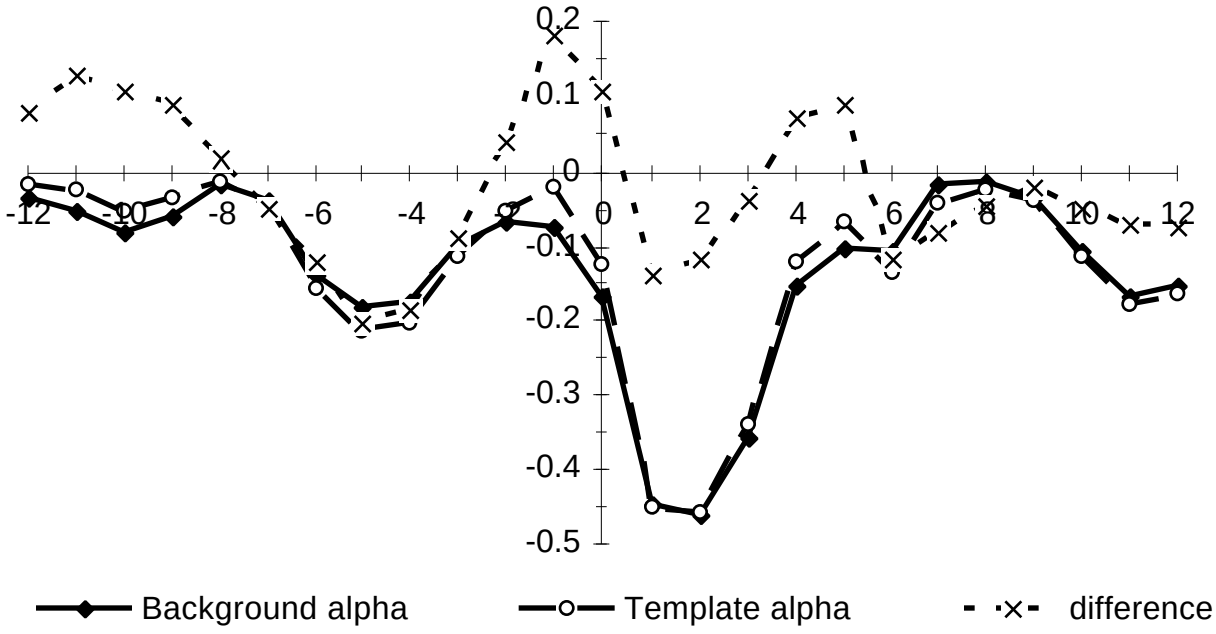


Figure 5. Cross-correlation of TFT with background and template model alphas

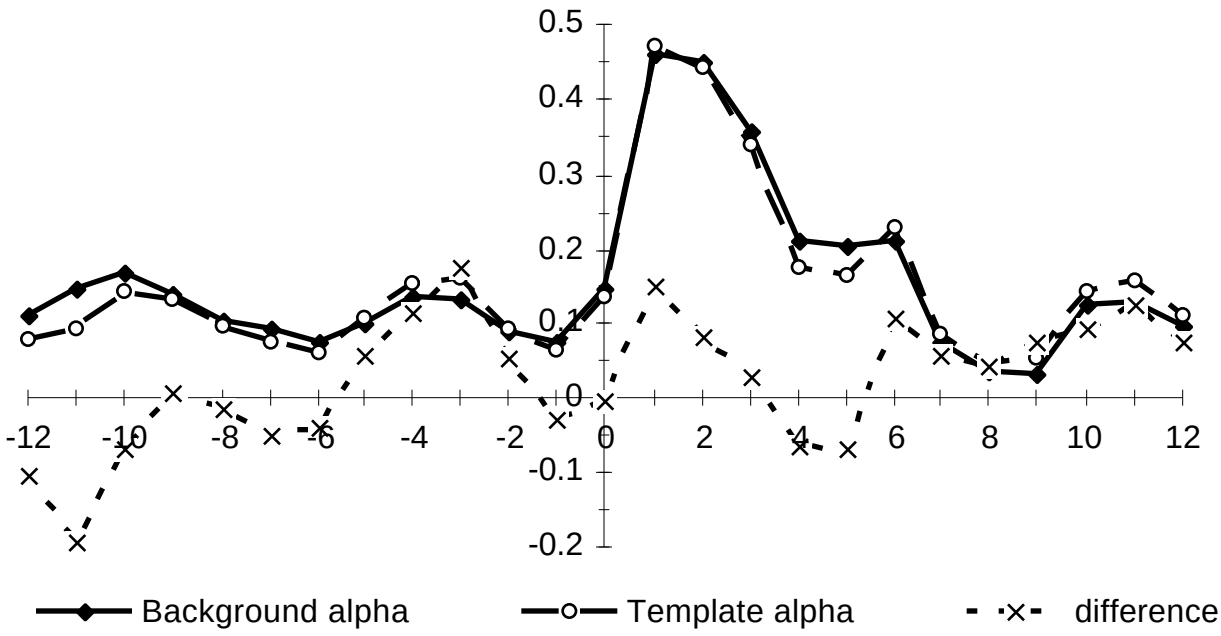


Figure 6. Cross-correlation of Goldstein scores with background and template model alphas

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 our enthusiasm for this approach are quickly dampened upon noticing that the cross-correlations of background and template models are almost indistinguishable. It was 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 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).[30] 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.

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

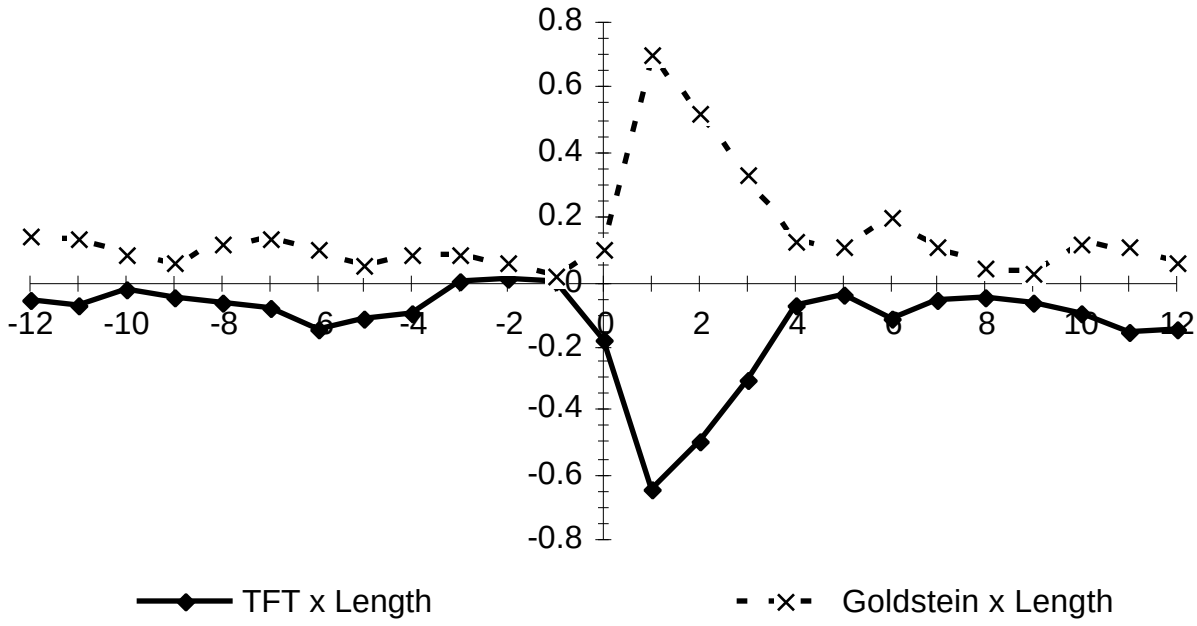


Figure 7. Cross-correlation of TFT and Goldstein with length of sequence in days

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 correct direction...) with the TFT series.[31] 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 $\Delta>2.0$ 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

[31] 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.

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.

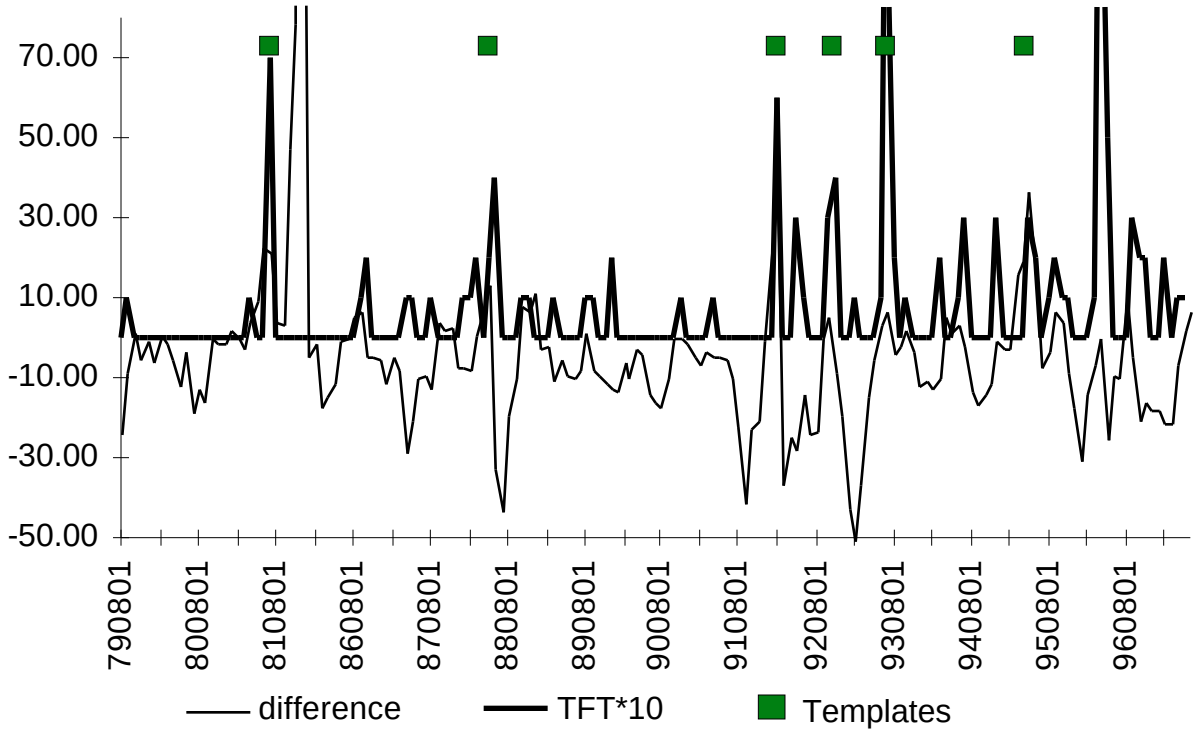


Figure 8. Time series of difference between template and background alphas with the TFT scores

6.4.3. 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. The Viterbi algorithm in the HMM estimation procedure 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, one 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, we 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, our 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 we 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 we'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 the desired early-warning indicator, 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. A plot of the distribution of these cross-correlations over 512 estimated models 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 $-0.1 < r < 0.1$.[32] 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, note that models with

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

high Q_{BC} cross-correlations emerge quite consistently from this technique—in other words, we are presenting typical results of a search across 128 or 256 models, not the best results achieved over months of computation. The early warning models are rare, but not very rare.

To provide a stronger test, we estimated some models using a split sample design. The data set was divided in half at July 1990, then we found the HMM that maximized the Q_{BC} x TFT cross-correlation for the data prior to July 1990. The cross-correlations were then calculated for only the second half of the data (t ≥ 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,[33] 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, 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).

[33] 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...

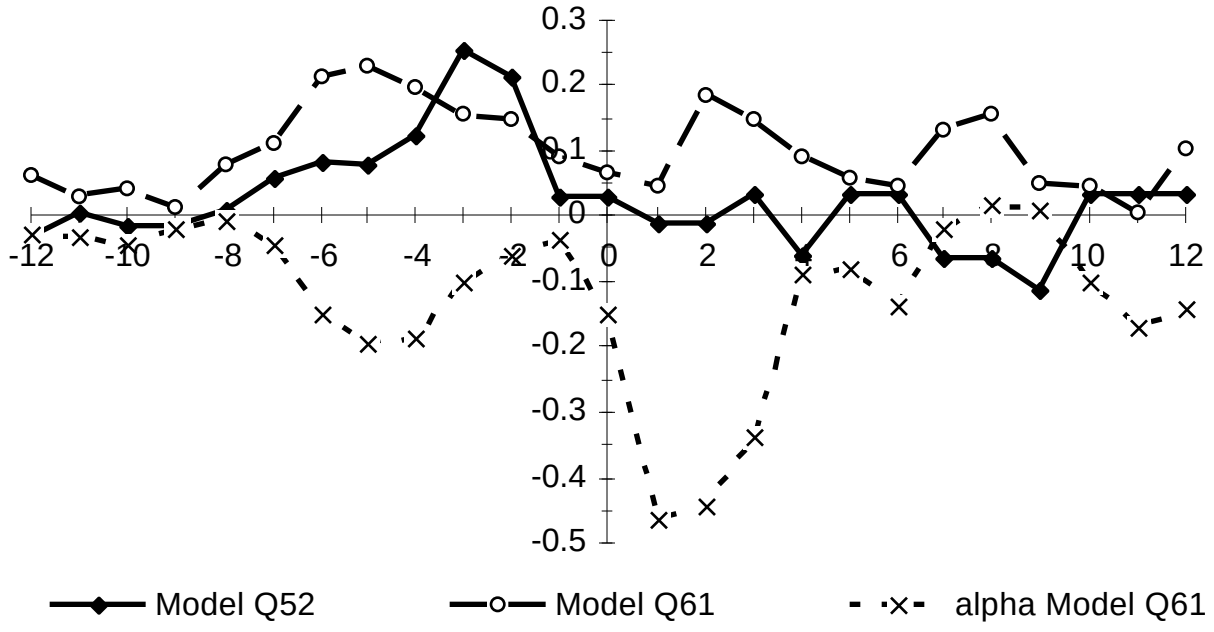


Figure 9. Cross-correlation of TFT with for Q_{BC} and alpha for Models Q52 and Q61

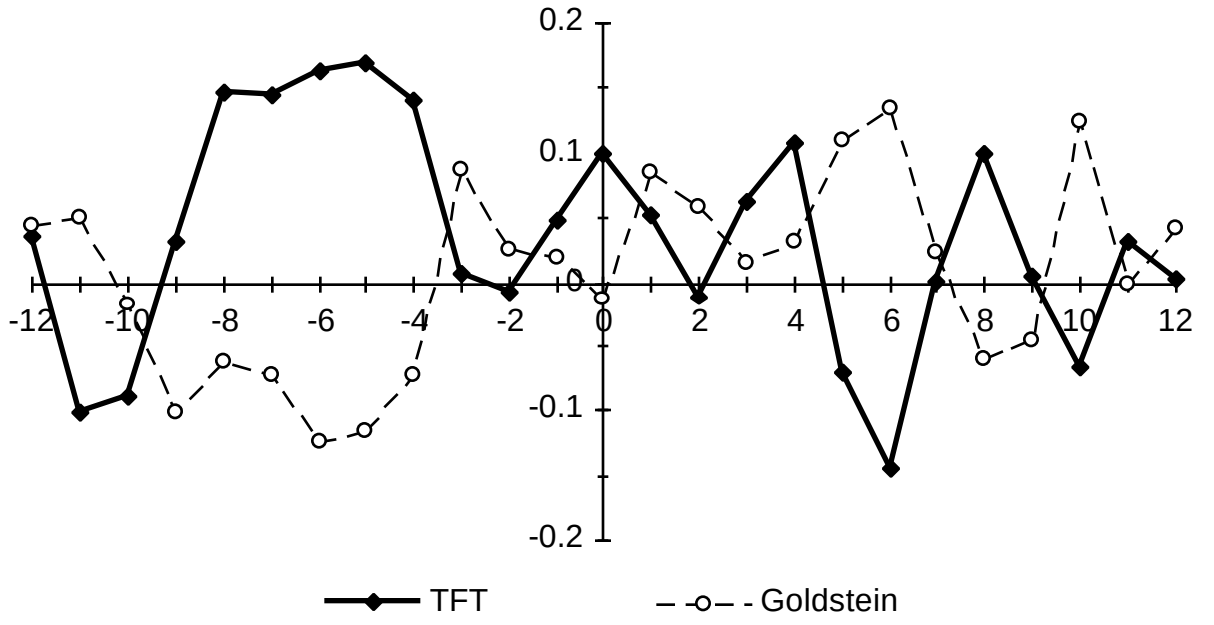


Figure 10. Q_{BC} cross-correlation for the P77 model

We repeated the split-sample tests in a random set of data that have a similar marginal distribution of events but no auto-correlation.[34] 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 (although 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

[34] 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.

14) without compensating consultations and promises, which may be why those states function as early warning indicators.

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 interpretation of "chance" 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.;

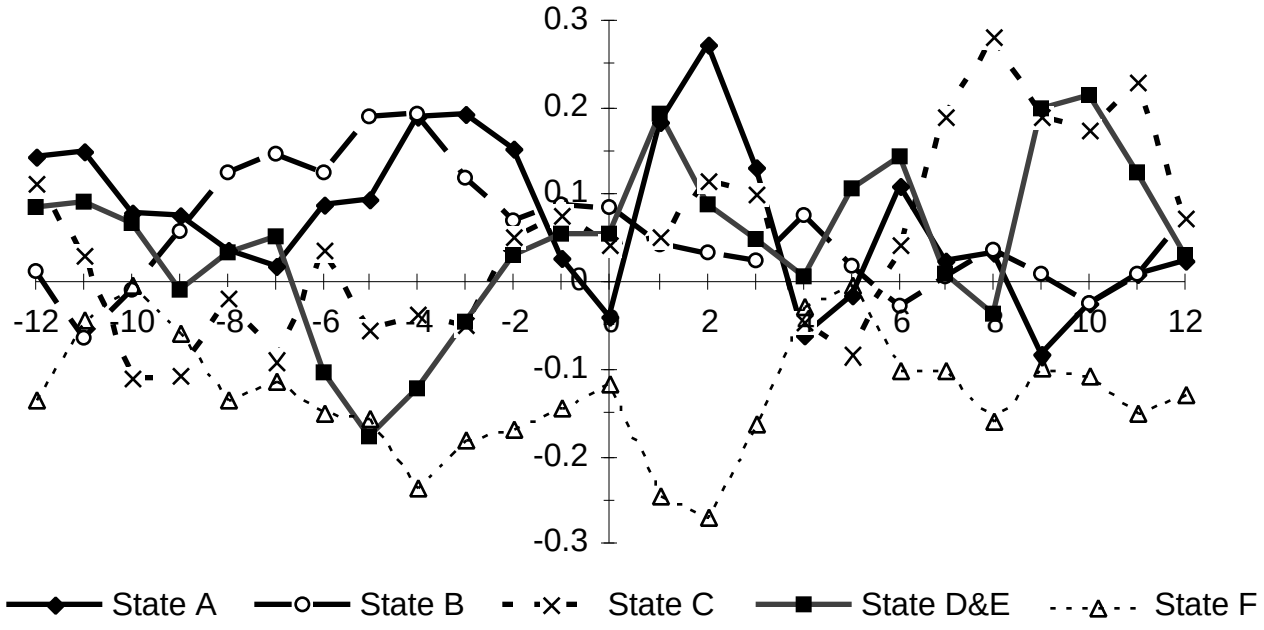


Figure 11. Cross-correlation with TFT by states in the P77 Model

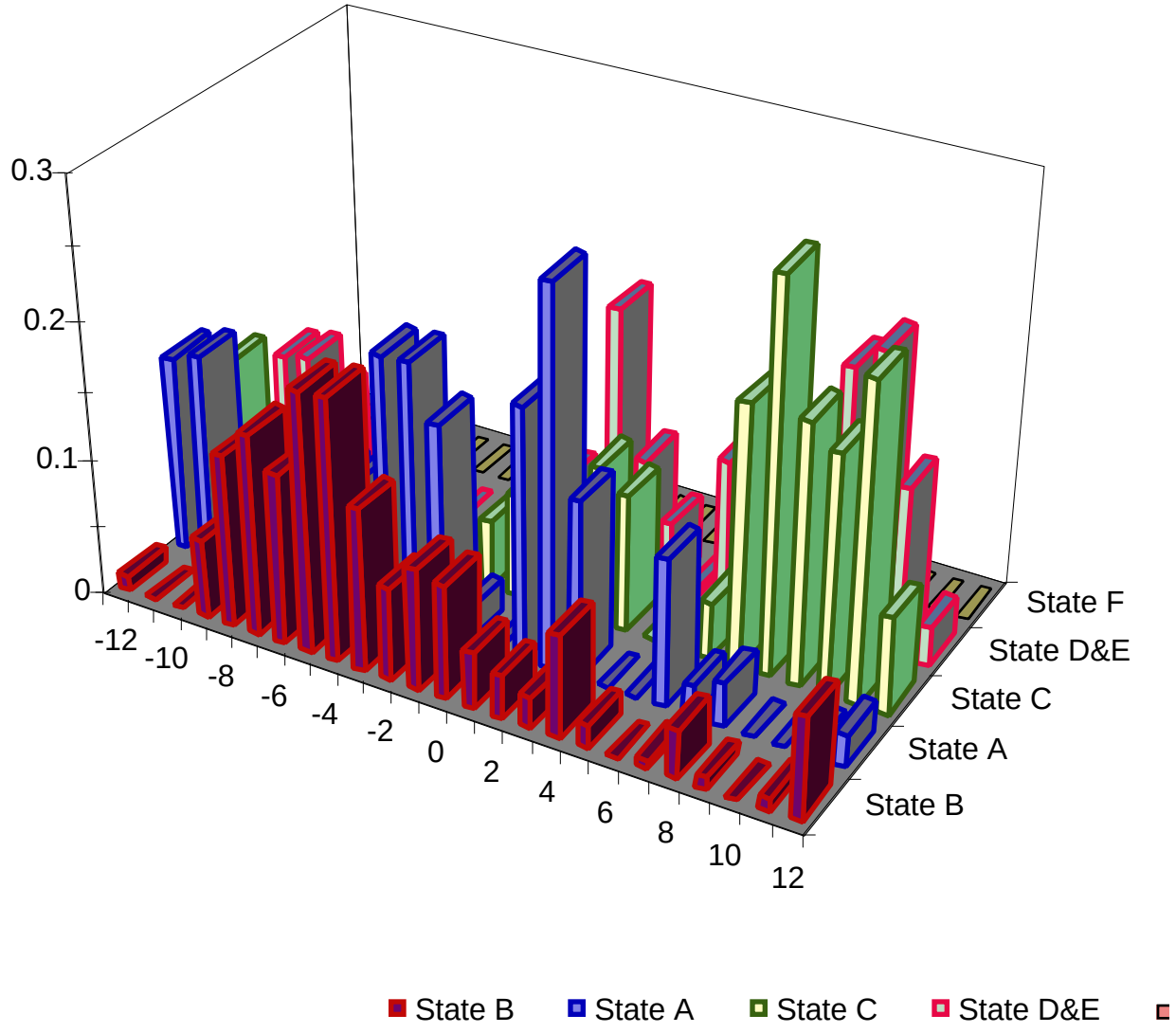


Figure 12. Positive cross-correlation with TFT by states in the P77 Model

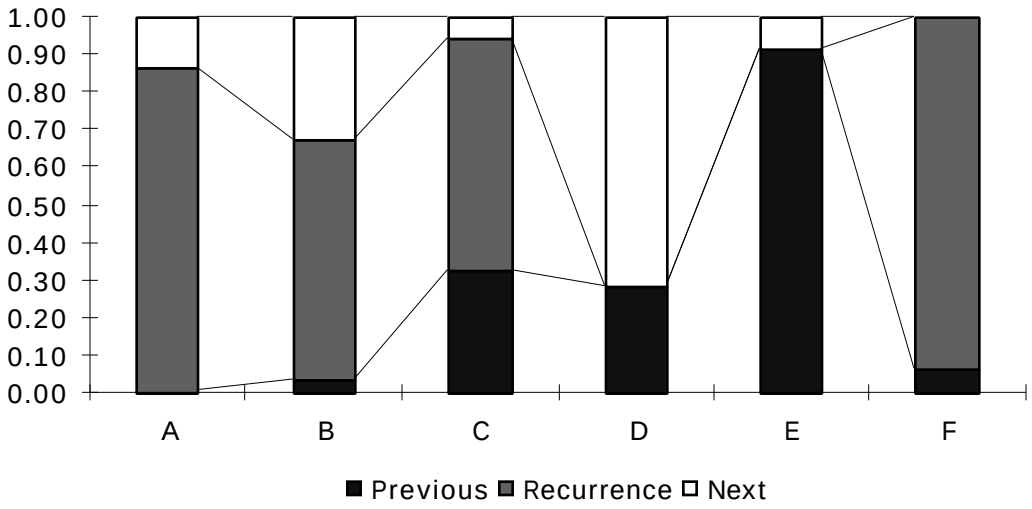


Figure 13. Transition probabilities in the P77 model

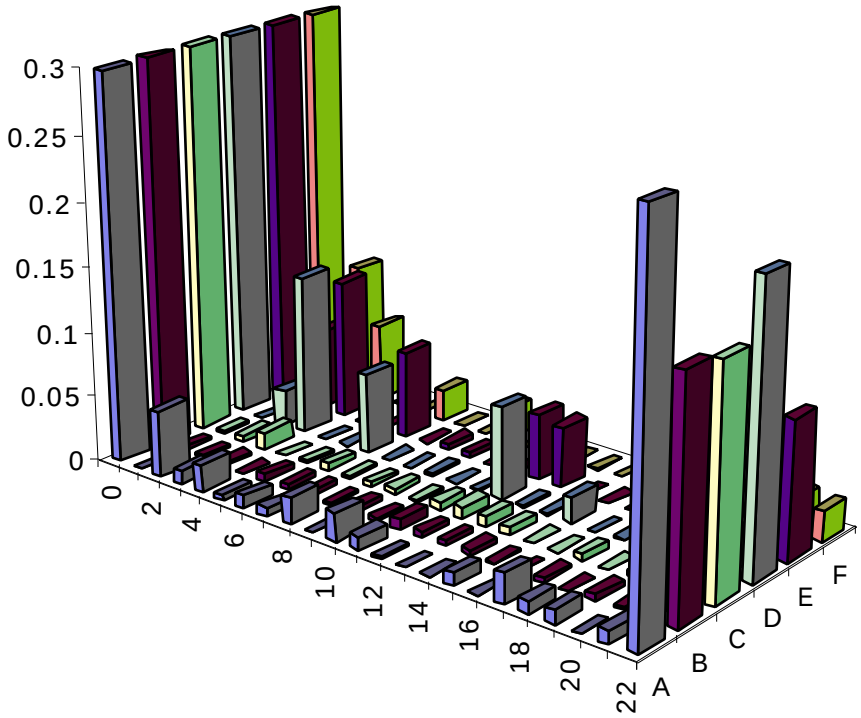


Figure 13. Symbol Probabilities in the P77 Model

6.5. Conclusion

Based on these results, the hidden Markov model appears to be a relatively robust, while still relatively parsimonious, technique for specifying general categories of event sequences. In this concluding section, we will discuss the implications of these results in three domains: theory, estimation and early-warning.

6.5.1. Theory

The HMM, like the Levenshtein metric and the parallel subsequences discussed in Chapter 4, provides a reproducible method of determining how similar a particular sequence of events is to one or more "analogous" sequences. The HMMs were effective in classifying conflicts found in the BCOW cases, in generalizing the BCOW crises to measure conflict in the contemporary Levant, and in providing precedent-based early warning in the protracted conflict in southern Lebanon.

Our 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.[35] Models might also be made more robust by predicating them on the values

[35] 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 a recent early warning conference, we 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.

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.

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

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 sequences 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

[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.

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 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.

None of this is to suggest that the use of precedent and analogies is a panacea. Political forecasting will always be a difficult task, and the literature dealing with the use of precedent in political reasoning focuses at least as much on how analogies can be misused as how they are successfully used.[37] Yet political analysis, unlike weather forecasting or billiards, is a reflexive endeavor: Political behavior is determined in part by how individuals analyze politics. The most common flaws cited in the human use of historical analogy are the undue influence of superficial similarities, the failure to consider the role of background conditions, and a tendency to search only a limited set of candidate examples. These same flaws are likely to be shared by HMMs, so at worst these models may provide a good indicator of possible precedents that human political actors could be considering. At best, a more sophisticated system—perhaps combining HMMs with other techniques—could be developed that specifically avoids some of the problems known to occur in human political pattern recognition.

6.5.2. 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.

---

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

In comparison with earlier techniques for the analysis of event data—which frequently required a great deal of statistical sophistication and "tweaking" of the resulting models—the HMM is sufficiently robust that a basic model could be estimated by an analyst with little or no knowledge of the underlying mathematical methods. In this scenario, the output of a monitoring system would be a list of probable matching sequences and their likelihoods. If the problem of comparability among sequences of different lengths could be worked out, an automated system (using machine-coded event data) could provide a real-time alert whenever the probability of a dyadic behavior matching one of the precursor models exceeded some threshold. This technique is substantially closer to the style of political analysis used in most policy settings, and therefore might be more acceptable than earlier event data efforts that relied on simple quantitative indicators without providing specific historical referents.

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).[38] In addition, Rabiner (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.

---

[38] 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 it is left as an exercise for the reader or his/her research assistants.

[Addendum October 1999: Since this was written, I've done a series of experiments using GAs to estimate HMMs for forecasting conflict in the Balkans. They provide *some* improvement over the Monte Carlo method, but not a dramatic improvement. The local maximum problem has thus far resisted every technique I've thrown at it, and my review of the literature on HMMs indicates that this is a general issue.]

Finally, the war/nonwar crisis distinction used in this study is quite crude. A more sophisticated alternative would be to use Leng's (1993) typology of bargaining strategies—bullying, reciprocating, appeasement, stonewalling, and trial-and-error—to differentiate between dyadic political activities. The probabilities of a dyad fitting each of several different models would then place it in an N-dimensional vector space. This is a straightforward generalization of the Goldstein and Azar-Sloan scales, which place behaviors on a single conflict-cooperation dimension.

The HMM approach 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 that can be modeled using sequence-analysis methods; Poisson conflicts do not. Again, the sequence-analysis approach—the indeterminacy of the HMMs notwithstanding—has the advantages of transparency (a term that we 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.

6.5.3. Early Warning

Our study of southern Lebanon focused on 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.[39]

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

[39] 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...

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.

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, we 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 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).[40] 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,

[40] 1967 and 1973 are both examples of strategic military innovations but the same arguments apply to diplomatic innovations such as Camp David and Oslo.

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, and the WEIS-coded crisis data set being collected by Goldstein and Pevehouse (Goldstein 1997) that covers about a dozen contemporary crises—including the Arab-Israeli conflict, Iran-Iraq, Chechnya, the former Yugoslavia, and the Great Lakes of Africa—is an obvious source for this. Those models then could be used to simply monitor the likelihood of specific crisis precursors, without attempting to aggregate these probabilities into a single quantitative measure or a location in a vector space. A change in the degree of fit to various precursors—for example observing that events seem to indicate a movement from a conciliatory to a bullying bargaining strategy—might also be useful for early warning.

References

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

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

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

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

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

Butterworth, Robert Lyle. 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.

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

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

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

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

Lebow, Richard Ned. 1981. *Between Peace and War: The Nature of International Crises*. Baltimore: Johns Hopkins University Press.

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

Leng, Russell J. 1993a. "Reciprocating Influence Strategies in Interstate Crisis Bargaining." *Journal of Conflict Resolution* 37:3-41.

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

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.

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

Mefford, Dwain. 1985. "Formulating Foreign Policy on the Basis of Historical Programming." In Michael Don Ward and Urs Luterbacher (eds.) *Dynamic Models of International Conflict.* Boulder, CO: Lynn Rienner Publishing.

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

Schrodt, Philip 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, Philip A. 1993. Rules and Co-Adaptation in Foreign Policy Behavior. Paper presented at the meeting of the International Studies Association, Acapulco.

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

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

Appendix 6.A: WEIS equivalents of BCOW codes

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

Physical actions

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

Verbal Actions

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

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

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