Forecasting Conflict in the Balkans using Hidden Markov Models

Philip A. Schrodt

Department of Political Science

University of Kansas

Lawrence, KS 66045 USA

phone: +1.785.864.9024 fax: +1.785.864.5700

p-schrodt@ukans.edu

August 2000

Paper presented at the American Political Science Association meetings
Washington, DC

Electronic copies of this paper are available at http://www.ukans.edu/~keds/papers.html and at APSA PROceedings webs site: http://PRO.harvard.edu. This work was partially supported by an EPSCoR start-up grant from the National Computational Science Alliance and utilized the NCSA SGI/CRAY Origin2000 parallel computing system.

Abstract

This study uses hidden Markov models (HMM) to forecast conflict in the former Yugoslavia for the period January 1991 through January 1999. The political and military events reported in the lead sentences of Reuters news service stories were coded into the World Events Interaction Survey (WEIS) event data scheme. The forecasting scheme involved randomly selecting eight 100-event "templates" taken at a 1-, 3- or 6-month forecasting lag for high-conflict and low-conflict weeks. A separate HMM is developed for the high-conflict-week sequences and the low-conflict-week sequences. Forecasting is done by determining whether a sequence of observed events fit the high-conflict or low-conflict model with higher probability.

Models were selected to maximize the difference between correct and incorrect predictions, evaluated by week. Three weighting schemes were used: unweighted (U), penalize false positives (P) and penalize false negatives (N). There is a relatively high level of convergence in the estimates—the best and worst models of a given type vary in accuracy by only about 15% to 20%. In full-sample tests, the U and P models produce at overall accuracy of around 80%. However, these models correctly forecast only about 25% of the high-conflict weeks, although about 60% of the cases where a high-conflict week has been forecast turn out to have high conflict. In contrast, the N model has an overall accuracy of only about 50% in full-sample tests, but it correctly forecasts high-conflict weeks with 85% accuracy in the 3- and 6-month horizon and 92% accuracy in the 1-month horizon. However, this is achieved by excessive predictions of high-conflict weeks: only about 30% of the cases where a high-conflict week has been forecast are high-conflict. Models that use templates from only the previous year usually do about as well as models based on the entire sample.

The models are remarkably insensitive to the length of the forecasting horizon—the drop-off in accuracy at longer forecasting horizons is very small, typically around 2%-4%. There is also no clear difference in the estimated coefficients for the 1-month and 6-month models. An extensive analysis was done of the coefficient estimates in the full-sample model to determine what the model was "looking at" in order to make predictions. While a number of statistically significant differences exist between the high and low conflict models, these do not fall into any neat patterns. This is probably due to a combination of the large number of parameters being estimated, the multiple local maxima in the estimation surface, and the complications introduced by the presence of a number of very low probability event categories. Some experiments with simplified models indicate that it is possible to use models with substantially fewer parameters without markedly decreasing the accuracy of the predictions; in fact predictions of the high conflict periods actually increase in accuracy quite substantially.

The Sequence Recognition Approach to Political Forecasting

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

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

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

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

Analogical reasoning is an easy task for the human brain, one that is substantially easier than sequential or deductive reasoning. Most experimental evidence suggests that human memory is organized so that when one item is recalled, this naturally activates links to other items that have features in common, and these are more likely to be recalled as well (Anderson 1983; Kohonen 1984).

Because analogies are so prevalent in human political reasoning, it would be helpful to have some computational method for systematically assessing the similarity of two sequences of political events. In Schrodt (1991), I posed this problem in the following manner:

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

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

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

Hidden Markov models

Over the past decade the hidden Markov model (HMM) has emerged as one of the most widely used techniques for 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 general model).sequence comparison method. 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 modeling the control of cellular phone networks, computer recognition of American Sign Language and—inevitably—the timing of trading in financial markets. The purpose of this project is to apply this technique to the problem of forecasting conflict in the former Yugoslavia.

A sequence is "noisy" when it contains missing, erroneous and extraneous elements, and consequently the sequence cannot be classified by simply matching it to a set of known "correct" sequences. A spelling program, for example, would always mark "wan" as an incorrect spelling of "one" because written English usually allows one and only one correct spelling of a word. Spoken English, in contrast, allows a wide variation of pronunciations, and in some regional dialects, "wan" is the most common pronunciation of "one". A computer program attempting to decipher spoken English needs to provide for a variety of different ways that a word might be pronounced, whereas a spelling checker needs only to know one.

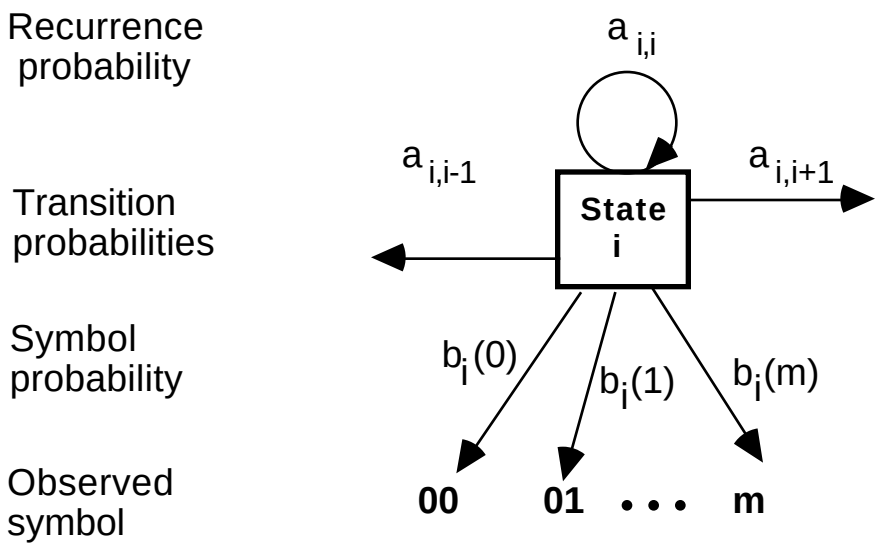
An HMM is a variation on the well-known Markov chain model, one of the most widely studied stochastic models of discrete events (Bartholomew 1975). The standard reference on HMMs is Rabiner (1989), which contains a thorough discussion of the estimation techniques used with the models as well as setting forth a standard notation that is used in virtually all contemporary articles on the subject. Like a conventional Markov chain, a HMM consists of a set of n discrete states and an n x n matrix $[A] = \{a_{ij}\}$ of *transition probabilities* for going between those states. In addition, however, every state has a vector of *observed symbol probabilities* that combine into a second matrix $[B] = \{\mathbf{b}_j(k)\}$ corresponding 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". 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)

While HMMs can have any type of transition matrix, the model that I will focus on in this chapter is called a "left-right model" because it imposes the constraint that the system can only remain in its current state or move to the next 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 parts of a word may be spoken slowly or quickly but in normal speech the ordering of those parts is never modified.

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

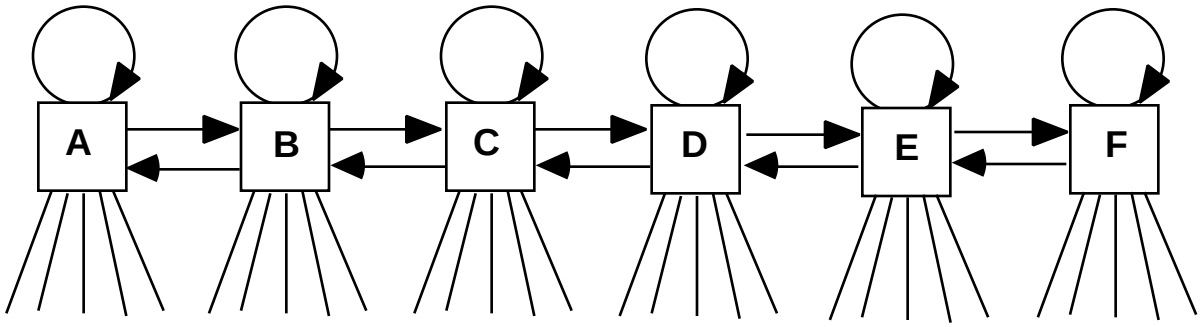


A series of these individual elements form an HMM such as the 5-state model illustrated in Figure 2. This illustrates a “left-right-left” model, where a process can make a transition to the previous state, the next state, or remain in the same state.[1]

In empirical applications, the transition matrix and symbol probabilities of an HMM are estimated using an iterative 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 coefficients 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."

[1] This is a generalization of the “left-right” model commonly used in speech recognition, where transitions are only allowed to the next state. In a left-right model, the final state of the chain is an "absorbing state" that has no exit probability and recurs with a probability of 1.

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



Because the Baum-Welch algorithm is an expectation-maximization method, it should, in theory, be possible to use the standard tools of maximum likelihood methods to compute asymptotic estimates of the standard errors of the estimates of the parameters in the [A] and [B] matrices, as well as comparing different models using likelihood ratios. In practice, however, this does not seem to be done, at least in the literature I've surveyed. The reason for this seems to be related to the local solutions provided by the Baum-Welch algorithm. As illustrated in the experiments below, the variance of the parameter estimates found in these local solutions is very large, although a variety of differing parameters appear to yield roughly similar estimates for the joint probability of the sequences.

After a set of models has been estimated, that set 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. 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 $10^{-(T+1)}$: 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. These sequence probabilities are consequently *extremely* small, even if the sequence was in fact generated by one of the models, but the only important comparison is the *relative* fit of the various models. The measure of fit usually reported is the log of the probability; this statistic is labeled α (alpha).

(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.)

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. The HMM that has the highest probability—for example the HMM corresponding to the numeral "three"—gives the best estimate of the number that was spoken.

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. The states of the model are unobserved, but have a close theoretical analog in the concept of crisis "phase" that has been explicitly coded in data sets such as the Butterworth international dispute resolution data set (Butterworth 1976), CASCON (Bloomfield & Moulton 1989, 1997) and SHERFACS (Sherman & Neack 1993), and in work on preventive diplomacy such as Lund (1996). For example, Lund (1996:38-39) outlines a series of crisis phases ranging from "durable peace" to "war" and emphasizes the importance of an "unstable peace" phase. In the HMM, these different phases would be distinguished by different distributions of observed WEIS events found in the estimated $\mathbf{b_j}$ vectors. 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 by the Baum-Welch algorithm from a different random starting point might be quite different—local maximization is also true of most other techniques for analyzing sequences. Furthermore, the computational efficiency of the Baum-Welch algorithm allows estimates to be made from a number of different starting points. The HMM model, being stochastic rather than deterministic, is specifically designed to deal with noisy input and with indeterminate time; both of these are present in international event sequences.

HMMs are *trained by example*—model that characterizes a set of sequences can be constructed without reference to the underlying rules used to code those sequences. This provides a close parallel to the method by which human analysts generalize sequences: They typically learn general characteristics from a set of archetypal cases.

HMMs do not require the use of interval-level scales such as those proposed by Azar and 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. The HMM, in contrast, uses only the original, disaggregated events and models the context of events by using different symbol observation probabilities in different states. An event that has a low probability within a particular context (that is, a specific hidden state) lowers the overall probability of the model generating the sequence. In aggregative scaling methods, events have the same weight in all contexts.

While most existing work with event data aggregates by months or even years, the HMM requires no temporal aggregation. This is particularly important for early warning problems, where critical periods in the development of a crisis may occur over a week or even a day. The HMM is relatively insensitive to the delineation of the start of a sequence. It is simple to prefix an HMM with an initial "background" state that reflects the distribution of events generated by a particular source (e.g. Reuters/WEIS) when no crisis is occurring. A model can simply cycle in this state until something important happens and the chain moves into the later states characteristic of crisis behavior.

There is a clear interpretation to each of the parameters of the [A] and [B] matrices, which allows them to be interpreted substantively; this contrasts with techniques such as neural networks that have a very diffuse parameter structure. More generally, the fit of the model has a familiar probabilistic interpretation. Finally—and not insignificantly—the HMM technique has already been developed and is an active research topic in a number of different fields. The breadth of those applications 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.

Data and Forecasting Model

Data

The event data used in this study were machine-coded using the 2-digit (22 category) WEIS system from the lead sentences in Reuters stories obtained from the NEXIS data service for the period January 1991 through May 1997 and the Reuters Business Briefing service for June 1997

through January 1999. These reports were coded using the Kansas Event Data System (KEDS) automated event data coding program (Gerner et al. 1994; Schrodt, Davis & Weddle 1994).

The KEDS coder does some simple linguistic parsing of the news reports—for instance, it identifies the political actors, recognizes compound nouns and compound verb phrases, and determines the references of pronouns—and then employs a large set of verb patterns to determine the appropriate event code. Only the lead sentences were coded and a sentence was not coded if it contained six or more verbs or no actor was found prior to the verb (sentences meeting these criteria have a greater-than-average likelihood of being incorrectly coded by KEDS). Schrodt & Gerner (1994), Huxtable & Pevehouse (1996) and Bond et al. (1996) discuss extensively the reliability and validity of event data generated using Reuters and KEDS. A **00** nonevent was added for each day in which no events were recorded in either direction in the dyad. Multiple events occurring in the same day are kept in the sequence. While the JWAC coding dictionaries sub-divide the various geographical-ethnic groups in the Balkans—for example separately coding Bosnian Serbs—these actors were combined into four primary ethnic groups—Serbs, Croats, Bosnians, and Kosovars—for the purpose of the analysis.

Because Reuters was intensely covering this region during most of the period analyzed, only lead sentences were coded. The KEDS program is capable of coding full stories and has been tested in that capacity (see Schrodt and Gerner 1998), but we have found that full-story coding gives additional information only when marginal actors are involved, or when an area receives only sporatic coverage (for example West Africa; see Huxtable 1997). Neither condition applies to the Balkans: significant events in the conflict almost inevitably received coverage in separate stories (i.e. in lead sentences), and in many cases, a single event would generate multiple stories.

Full-story coding would primarily serve to insert a large number of high-frequency verbal events ("comment" and "consult") into the sequences, and because the HMM models use templates that contain a fixed number of events (this is necessitated by the HMM approach), the effect of full-story coding would be to *reduce* the amount of information in any given sequence.[2] By using lead-sentence coding, we insure that the sequences contain the most important events (subject, as always, to the judgements of Reuters' reporters and editors.)

---

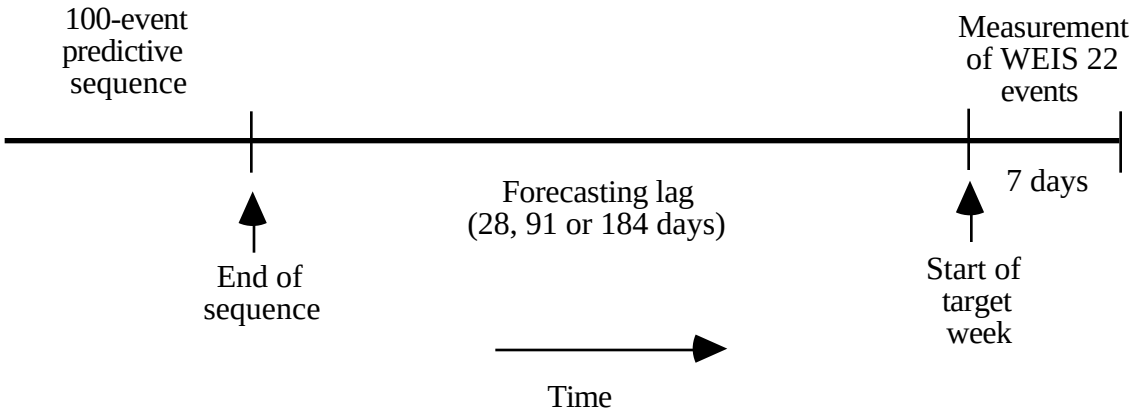
[2] That is, reduce the amount of information from the perspective of information theory: because the "comment" category in WEIS is very frequent, it carries less information than an event that occurs with low probability.

Forecasting Model

The accuracy of the model is accessed on whether it can predict when violence will occur, violence being defined as events coded into the WEIS "22" category. Given the diverse character of conflict in this region over the time period—violence having occurred between Serbs and Croats, Serbs and Bosnians, Serbs and Kosovars, but also occasionally between other groups (e.g. Croats and Bosnians)— I did not attempt to differentiate who was involved in the violence.

The forecast target was whether a week contained more or less than twenty (20) WEIS category **22** ("use of force") events. The threshold of 20 events is somewhat arbitrary, but that threshold seemed to be reasonable for differentiating periods when the Balkans were relatively quiet from those where there was substantial violence. Approximately 20% of the weeks in the data set satisfy the "high conflict" criterion.

Figure 3. Prediction Scheme



Three forecast periods were used

28 days	approximately 1 month
91 days	approximately 3 months
184 days	approximately 6 months

The early warning sequence for each week consisted of the 100 events prior to the first day of the week minus the forecast period;[3] Figure 3 provides a schematic of this model.

[3] Why 100?—because I have ten fingers... The length of the warning sequence is a free parameter—SEQ_SIZE in the program—and other values might work better, depending on the application. In previous work on the Middle East, I have done some experiments with sequences of 50 and 200 events; the results were roughly comparable to the results from 100 event sequences. Given the vagaries of timing in this region—for example the unpredictable and seasonally variable effects of weather on military operations, as well as the large number of

The HMMs were estimated using 16 "forecasting templates," eight for the high-conflict weeks and eight for the low-conflict weeks. A high-conflict template was created from the data by choosing a high-conflict week at random, then getting a 100-event sequence with the appropriate forecasting lag; low-conflict templates are created in a similar manner. Finally, prediction for a particular week is done by computing the fit of a 100-event sequence, obtained with a suitable lag prior to the beginning of the week, and then assigning the "high conflict" or "low conflict" prediction depending on which one of the two models had a higher probability of generating the sequence.

In this preliminary analysis, the templates have been chosen at random from *any* week in the data set, so while the computed *forecast* is strictly predictive (i.e. the prediction technique uses no information beyond that available at the beginning of the week minus FORECAST days), the *estimation* of the predictive model is not, because the model can use templates that occurred after the time of the forecast. (This approach was used because I anticipated that some periods of time would provide higher-quality templates than others, though this turned out not to be the case so far as I can determine). Nonetheless, only a very small amount of the data[4]—approximately 4%—is being used to characterize the complete sequences, so the model is certainly not tautological. It is straightforward to switch the estimation is a purely predictive mode; this will be done for the final analysis in this project once the various elements of the model have been finalized. Two additional estimation schemes that are purely predictive—they estimate the model on a set of data that occurs before the period where the accuracy is assessed—are also estimated; they will be discussed below.

The forecasting model used the following eight relationships

Serbia -> any target	Croatia -> any target
any source -> Serbia	any source -> Croatia
Bosnia -> any target	Kosovo -> any target
any source -> Bosnia	any source -> Kosovo

Following the approach in Schrodt (2000), the multiple interactions were modeled by incrementing the WEIS code for the Nth dyad by (N-1)*22, so for example the {any source -> Serbia} events have codes **23** through **44** (corresponding to the original WEIS codes **01** to **22**) the {Croatia -> any target} events have codes **89** through **110**, and so forth. If no event occurred with either dyad,

---

diplomatic interventions during the course of the conflict—it is unlikely that the model will be very sensitive to the length of the sequence.

[4] 1600 events in the templates out of about 42,000 events in the total sequence.

the **00** nonevent was assigned to the day. The resulting model contains 177 event codes (8*22 + 1).

This approach involves a relatively high level of aggregation, since it does not differentiate the source or target of interactions with the four principals in the conflict. This was done for three reasons. First, if I had tried to further differentiate sources and targets, the number of codes required in the model would have increased geometrically: For example looking only at the four principals plus the international community would have involved 20 dyads, which would mean 441 codes, which leads to a parameter space of dimension 5,324, which makes consistent estimation that much more difficult. Second, because of th Serbia first Croatia, then Bosnia, then a Croat-Bosnian alliance,a template chosen at any particular time would primarily consist of zero values for almost all of the dyads, and therefore would not generalize easily. This approach, in contrast, essentially asks "What do Serbian (or Croatian, Bosnia, Kosovar) patterns look like when Serbia is interacting with someone, anyone." Finally, the specific sources and targets are generally consistent during a given pef time, so while the dyadic behaviors are aggregated, it is simple to identify the actors responsible for most of the activity at any given point in time.

Estimation Algorithm

The HMM parameters were estimated 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. I translated this code from the Solaris C++ environment to a Macintosh CodeWarrior ANSI C environment, in the process combining Meyers and Whitson's separate driver programs for training and testing into a single program, and modifying the input format to handle the WEIS sequences. I then extended the code to handle the left-right-left (LRL) model, and implemented the Viterbi algorithm described in Rabiner (1989) in order to estimate the most likely state sequence. In the process of extending the model to the LRL form, I rewrote the estimation equations to correspond exactly to those in Rabiner—the Meyers & Whitson implementation differed slightly from Rabiner's equations, presumably because their models estimate a separate vector for "transition symbols." These new procedures produce estimates similar to those of Meyers & Whitson when all probabilities to previous states are forced to zero. This source code is available on the KEDS project web site: http://www.ukans.edu/~keds/software.html.

The program requires about 2 Mb of memory for a system using 177 codes, 6 states and 100-event sequences. The largest arrays required by the program are proportional to (M+T)*N, where M is the number of possible event codes, T is the maximum sequence length and N is the number

of states. Consistent with the CASCON and SHERFACS approaches, and with earlier work that I did on the Middle East (Schrodt 1999, 2000), the models I estimated used 6 states.

On average, estimating a single 177-code, 6-state HMM model using 8 template sequences takes about 0.80 seconds on a 350 Mhz Apple Macintosh G3, and calculating the probability of a given sequence—i.e. doing a prediction—is practically instantaneous. Unfortunately, this speed is canceled out by the fact that the Baum-Welch algorithm is not very efficient at finding global maxima, and therefore a combination of Monte Carlo and genetic algorithm methods were required to locate good models [see discussion below]. In the protocols I've been using, a standard run requires estimating 49,664 models (= 2* {[(24*64)+16]*16}) so it takes around 11 hours. This limited full evaluation to a couple of runs per day on a personal, so it was difficult to experiment with a large number of parameters. However, I would reiterate that this is only a constraint on *estimating* the optimal model—once models have been established, new data can be evaluated against these very, very quickly.

[In the course of doing this research, I adapted the programs to run on a parallel supercomputer at the National Center for Supercomputer Applications; this experience is described at `http://www.ukans.edu/~keds/NCSA.html`. The supercomputer definitely allowed greater experimentation; these results are briefly summarized below.]

As noted above, the optimization surface of an HMM calculated using the Baum-Welch algorithm is characterized by a very large number of local maxima. This means that the resulting the parameters—and the predictive accuracy—of the model depend heavily on the initial approximation for those parameters. In other words, the point where the estimation routine stops depends heavily on where it starts, and those starting points occur in a 2156-dimensional space —2 models * [(6 states * 177 codes) + 16 transition probabilities].

In order to efficiently explore this surface, I used a combination of Monte Carlo methods and genetic algorithms. This was done as follows (the capitalized variable names correspond to constants in the estimation program):

1. A set of GA_VECT initial starting points for a pair of models—one for the high conflict templates and one for the low conflict templates—was chosen (typically GA_VECT = 32). The initial parameter values were chosen from a uniform [0,1] distribution, then normalized so that they summed to 1.

2. An HMM was estimated for each of these starting points, then the "fitness" of the estimated model was computed. "Fitness" was some combination of correct minus incorrect predictions (see below).

3. After all of the models had been computed, the vectors were sorted in descending order of fitness. Only the top GA_SURVIVE (=0.25) * GA_VECT vectors were saved.

4. The remaining (1 - GA_SURVIVE) * GA_VECT vectors were replaced by reproducing the surviving vectors using the standard genetic algorithm method: two "parents" were chosen at random (the likelihood of choosing being proportional to their fitness), then a new vector was created by taking the first NR elements of the first vector (where NR is a random number) and the remaining elements from the second vector. In addition, elements were randomly mutated with probability GA_MUTATE (=0.10).

This process was repeated for GA_MAX_EON (=64) iterations, then the best-fitting model was saved, and a set of predictions was made using that best-fitting model. Finally, the GA was repeated with a number of different starting points and templates (GA_EXPER, which was set at 16 but repeated multiple times in some cases).

The GA seems to work fairly efficiently, though it does not converge to a single model in the Balkans case. However, the range of optimal values is relatively limited, typically varying by around 15% - 20% between the best and worst models. This variation, while not ideal, is considerably more limited than what I got using a simpler Monte-Carlo estimation.[5] The GA is not, however, a “magic bullet” and is itself quite time-consuming—I did a small number of experiments where I estimated models using a pure Monte-Carlo search that involved the same number of evaluations as the GA, and the GA gives only about a 10%-20% improvement in accuracy for the same amount of work. Other complex optimization approaches such as simulated annealing might produce better results.

While I used GA_MAX_EON =64 in virtually all of the experiments, most of the optimization tends to occur early in the process.[6] I did some experiments (on a simpler model, Israel-Lebanon) with altering the various parameters of the GA—GA_VECT, GA_SURVIVE and GA_MUTATE—and the optimization doesn't seem very sensitive to these values.[7] (I did a run of 16

---

[5] This variation is partly due to the different choice of templates for high and low conflict periods: some templates are going to be more "typical" than others, and consequently will provide a better fit. In a separate experiment on Israel-Lebanon data where a fixed set of templates was used, the variation among experiments was in the range of 10%.

[6] Virtually all of the optimization occurs in the first 32 eons of the genetic algorithm, and one could probably set GA_MAX_EON =32 with almost no loss of accuracy. This would increase the speed of the experiments by a factor of 2.

[7] I did a run of 16 experiments on the 3N model for the Balkans data with GA_VECT=48 and GA_SURVIVE=0.33—in other words, keeping 16 rather than 8 of the "parent" vectors—and this made no discernible difference in either the overall accuracy or the rate of convergence.

experiments on the 3N model for the Balkans data with GA_VECT=48 and GA_SURVIVE=0.33 —in other words, keeping 16 rather than 8 of the "parent" vectors—and this made no discernible difference in either the overall accuracy or the rate of convergence.)

Alternative Measures of Fitness

The fitness of the model-pairs is computed using

$$\sum (w_{TP} + w_{TN} — w_{FP} — w_{FN})$$

where the w_{aa} correspond to the weights for a

TP: true positive	high conflict predicted when high conflict occurs
TN: true negative	low conflict predicted when low conflict occurs
FP: false positive	high conflict predicted when low conflict occurs
FN: false negative	low conflict predicted when high conflict occurs

This is equivalent to a weighted "right minus wrong" criterion:

(#correct predictions) minus (#incorrect predictions)

The sum is over all of the weeks in the data set (from 1 January 1991 to 26 January 1999); each week is classified into one of these categories.

Because the data set is strongly skewed towards low-conflict weeks (80% of the cases), this model will generally under-predict high-conflict: In particular a null model that predicts *only* low-conflict for all weeks would have an impressive 80% accuracy, but it will also be quite useless. This is the perennial early warning problem of balancing Type I and Type II errors: how should a model balance the possibility of false alarms with the possibility of missing actual cases of high conflict. One can create a [useless] model with zero false alarms by *never* predicting high conflict, and one can also create an equally useless model that misses none of the cases of high conflict by *always* predicting high conflict. There is no simple way around this tradeoff.

The simplest way of doing this is to differentiate reward and penalize some types of predictions over others. I experimented with three different weighting systems:

		w_{TP}	w_{TN}	w_{FP}	w_{FN}
Unweighted	1.0	1.0	1.0	1.0	
Entropy, FP	1.68	0.21	1.68	0.21	
Entropy, FN	1.68	0.21	0.21	1.68	

The entropy-based weights are suggested by the information theory computation of the entropy of a sequence

$$E = -\sum_{i=1}^{C} p_i \ln(p_i)$$

where

p_i = the proportion of category i in the data and

C = the number of distinct categories.

and ln() is the natural logarithm. By taking only the log of the proportion, one gets the weights

-ln(0.186) = 1.68 high conflict

-ln(0.814) = 0.21 low conflict

These weights were used as the "reward" for a correct prediction for each type of week. I also looked at two different ways of weighting the incorrect predictions: one using the high-conflict weight to penalize false positives; the other uses the high-conflict weight on false negatives.[8]

Results

Models were estimated for the three forecasting horizons and the three weighting systems, a total of nine models. At least sixteen Monte-Carlo experiments were run for each model; in some cases there are a larger number of models because I was able to salvage results from assorted experiments.

In all of the tables and figures, the numerical prefix (1, 3, and 6) refers to the forecasting horizon; the letter prefix (U, P, N) refers to the fitness weighting. In some cases, I analyze

[8] There is no particular reason that the penalty for incorrect predictions needs to have the same value as the reward for correct predictions, but this will do for a first approximation.

separately the best-fitting models (accuracy ≥ 0.795 for U and P models; accuracy ≥ 0.495 for N models) and for all models. Unless otherwise noted, the discussion will refer to the "best" models, but the "all" results are presented for purposes of comparison and to give some sense of the variance found in the estimates. In all of the analyses, the U and P models produce very similar results, with the N model being distinctive.

Overall Accuracy

The overall accuracy of the models is shown in Tables 1 and 2; additional tables show the four-way classification accuracy. Table 1 summarizes the number of models evaluated, the total number of observations (i.e. total number of weeks predicted across all of the models), and the average accuracy (true positive + true negatives). The key result here is that the overall accuracy for the best U and P models consistently show almost exactly 80% accuracy; the best N models somewhat less consistently show about 52% accuracy. The difference between the "best" and "all" model sets[9] is not dramatic—only about 5%—and quite surprisingly, there is very little drop-off in accuracy as the time horizon increases.

[This last characteristic worries me a bit, but I have thoroughly checked through the program code to make sure that this is not an artifact, and I can't find any errors. In addition, I have also used these programs to estimate models for a number of other conflict regions and the pattern of small decreases in accuracy—rather than accuracy increasing and decreasing randomly—is found consistently. The reason for this lack of sensitivity to time lags may be due to the episodic character of violence in the Balkans and elsewhere. The data set is characterized by two extended high-conflict periods—May-93 to June-94 and April-95 to October-95—with the remainder of the period having only sporadic conflict, often lasting only a couple of weeks. These gross characteristics may be predictable quite far in advance.]

[9] "All" includes the "Best" models

Table 1
Summary of Estimated Models

	Best Models			All Models		
Model	# Models	# Obsrv	% Correct	# Models	# Obsrv	%Correct
1-U	10	4090	80.7%	22	8998	78.5%
3-U	7	2800	80.6%	21	8400	76.9%
6-U	2	772	81.2%	16	6176	75.9%
1-P	7	2863	80.8%	29	11861	77.6%
3-P	11	4400	80.8%	54	21600	76.0%
6-P	3	1158	80.0%	16	6176	76.9%
1-N	15	6135	55.2%	16	6544	54.2%
3-N	8	3200	52.8%	16	6400	49.0%
6-N	7	2702	53.7%	16	6176	47.7%

Tables 2a and 2b show the accuracy of the forecasts broken down by high-conflict and low-conflict weeks respectively. The "observed" column gives the percentage of the weeks that were correctly forecast : this proportion is $\frac{TP}{TP+FN}$ for high conflict and $\frac{TN}{TN+FP}$ for low conflict. It is the percentage of time that a high or low conflict week would have been predicted correctly.

The "forecast" column, in contrast, gives the percentage of the weeks that were forecast as having high or low conflict actually turned out to have the predicted characteristic. This is $\frac{TP}{TP+FP}$ for high conflict and $\frac{TN}{TN+FN}$ for low conflict. It is the percentage of time that a type of prediction is accurate.

As indicated in the discussion of the weighting systems, the N-type models operate very differently than the U- and P-type models. As shown in Table 1, U and P models have a high overall accuracy. However, this accuracy comes almost entirely from correctly forecasting low-conflict weeks—U and P models predict about 95% of these weeks correctly, but correctly predict only about 25% of the high-conflict weeks. N-type models, in contrast, predict about 85% of the high-conflict weeks correctly (and 45% of the low-conflict weeks).

Table 2a
High Conflict Weeks

	Best Models		All Models	
Model	**Observed**	**Forecast**	**Observed**	**Forecast**
1-U	24.2%	55.5%	31.3%	45.7%
3-U	26.9%	57.1%	27.9%	41.6%
6-U	30.1%	63.3%	32.3%	42.2%
1-P	21.3%	55.0%	29.3%	40.8%
3-P	22.7%	57.1%	29.0%	37.9%
6-P	24.3%	57.1%	25.9%	42.6%
1-N	92.8%	28.6%	92.67%	28.1%
3-N	86.8%	27.3%	88.1%	25.9%
6-N	86.2%	28.5%	88.5%	26.3%

Table 2b
Low Conflict Weeks

	Best Models		All Models	
Model	**Observed**	**Forecast**	**Observed**	**Forecast**
1-U	95.1%	83.1%	90.5%	83.8%
3-U	94.7%	83.2%	89.7%	82.6%
6-U	95.2%	83.3%	87.9%	82.6%
1-P	95.6%	83.0%	89.5%	83.7%
3-P	95.7%	82.7%	87.9%	82.9%
6-P	95.06%	82.2%	90.6%	82.0%
1-N	46.5%	96.5%	45.3%	96.4%
3-N	44.6%	93.4%	39.6%	93.3%
6-N	45.5%	92.9%	37.4%	92.8%

From the perspective of forecasting, the N-type models are best at warning against the possible "bolt out of the blue,". However, this comes at a price of a lot of false alarms: when an N-type model forecasts a high-conflict week, there is only about a 30% chance that this will occur.[10] A high-conflict prediction by a U- or P-type model, meanwhile, will result in an actual high-conflict week in about 60% of the cases.

Time Series Analysis

Figures 4 through 6 show display the accuracy of the various models as a 5-week centered moving average.[11] (In the event you are reading a photocopy of this, the original graphs are in glorious color; these can be found in the electronic version of the paper. The correlation between the three forecast periods is quite high, so you are not missing much in black-and-white.) "Accuracy" is computed as the percentage of the "best" models that correctly classify each week. For reference, Figure 7 shows the target sequence of WEIS 22-type events, at the same scale as the remaining figures (the heavy line is a 5-week centered moving average).

There are some general patterns to the accuracy. All three models are very accurate at the beginning and end of the series, prior to about April 1992 and following April 1997. During the period October-94 to October-97, the U/P and N models are almost perfect mirror images of each other—when the accuracy of one type of model is high, the other is low, and vice-versa. (This also means that the models are making opposite predictions: for example following October-95, the P models almost always [correctly] predict low conflict, whereas the N-models predict high conflict.) This is not merely a visual effect: the correlation (r) for the 3-N versus 3-P models is -0.59 for Oct-91 to Mar-97, and -0.28 for the entire period.

Curiously, the N-model takes a very long time—almost eighteen months—to recover accuracy after the implementation of the Dayton Accords, which might suggest that despite the fact that these agreements resulted in a sudden cessation of hostilities, the preconditions for violence remained in place and Dayton may in fact have had a major impact in actually enforcing peace. The accuracy in the 1993 period is medium for both models; this is probably partly due to the fact that the high-low classifications are fluctuating rapidly during this period.

---

[10] However, as noted below, some of these false positives are "near misses" in the sense of predicting high conflict in weeks that fall just short of the 20-event threshold. Furthermore, many of these errors occur during the period of the implementation of the Dayton Accords, and if that period is eliminated from the analysis, the forecast accuracy of the N models is probably closer to 50%.

[11] The moving average is used for purposes of clarity—the weekly accuracy fluctuates very rapidly in places, resulting in a graph that is very difficult to read.

Figures 8 and 9 [attempt to] show the accuracy for high-conflict weeks only for the P and N models. These figures reinforce the points made earlier about the relative accuracy of the two approaches, and are primarily important for what they *don't* show: (1) there is no clear differentiation in the predictive accuracy of the 1, 3 and 6-month forecasts and (2) the accuracy is fairly consistent over time, except for the tails of the series (including 1992 for the N model).

Figure 4.

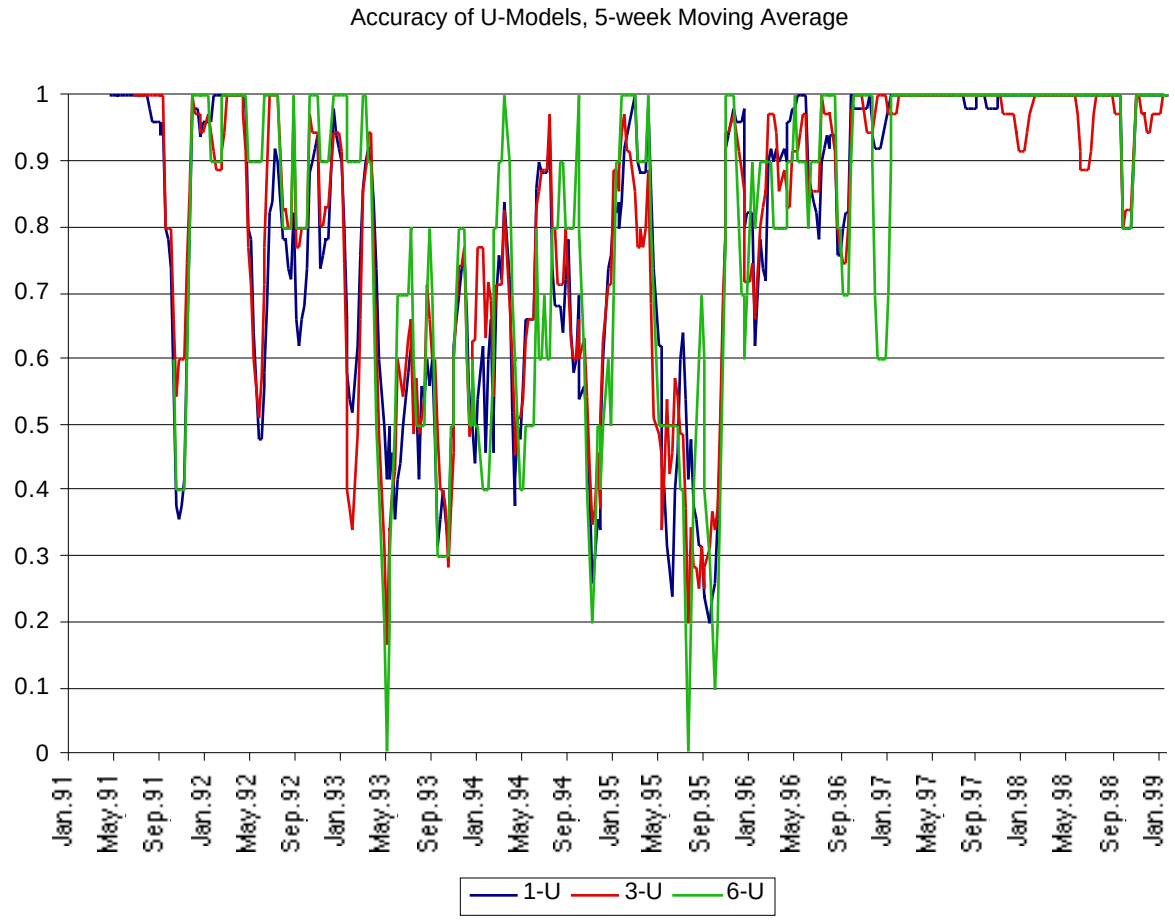


Figure 5.

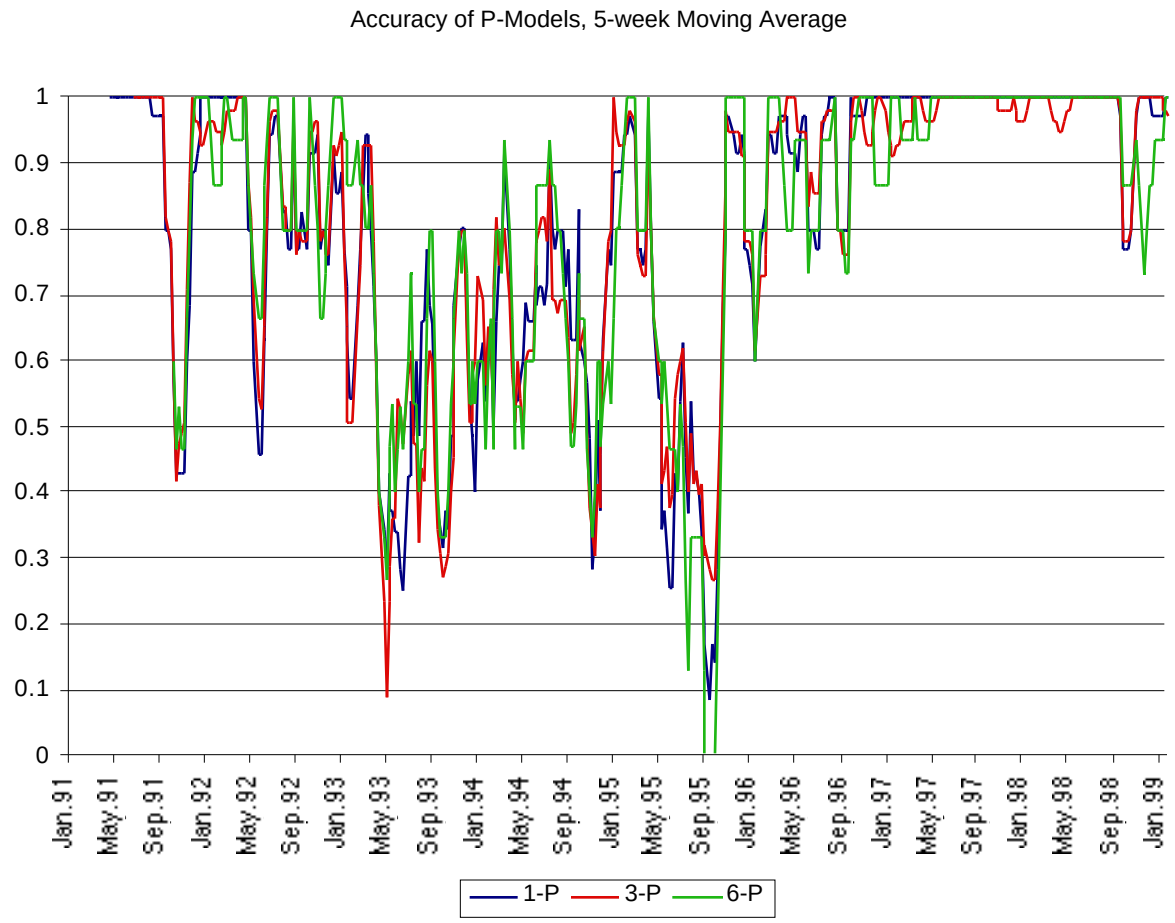

Accuracy of P-Models, 5-week Moving Average
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Jan.91
May.91
Sep.91
Jan.92
May.92
Sep.92
Jan.93
May.93
Sep.93
Jan.94
May.94
Sep.94
Jan.95
May.95
Sep.95
Jan.96
May.96
Sep.96
Jan.97
May.97
Sep.97
Jan.98
May.98
Sep.98
Jan.99
1-P
3-P
6-P

Figure 6.

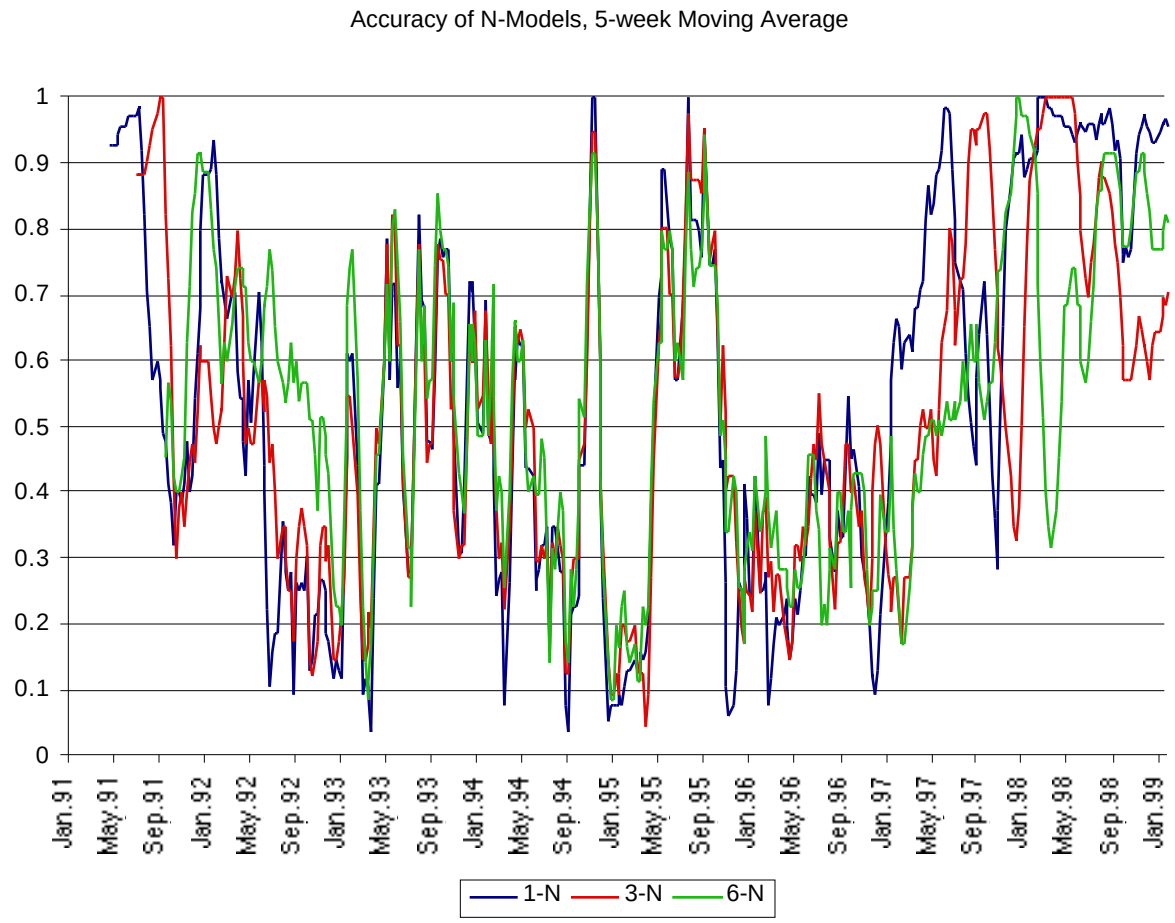

Accuracy of N-Models, 5-week Moving Average
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Jan.91
May.91
Sep.91
Jan.92
May.92
Sep.92
Jan.93
May.93
Sep.93
Jan.94
May.94
Sep.94
Jan.95
May.95
Sep.95
Jan.96
May.96
Sep.96
Jan.97
May.97
Sep.97
Jan.98
May.98
Sep.98
Jan.99
1-N
3-N
6-N

Figure 7.

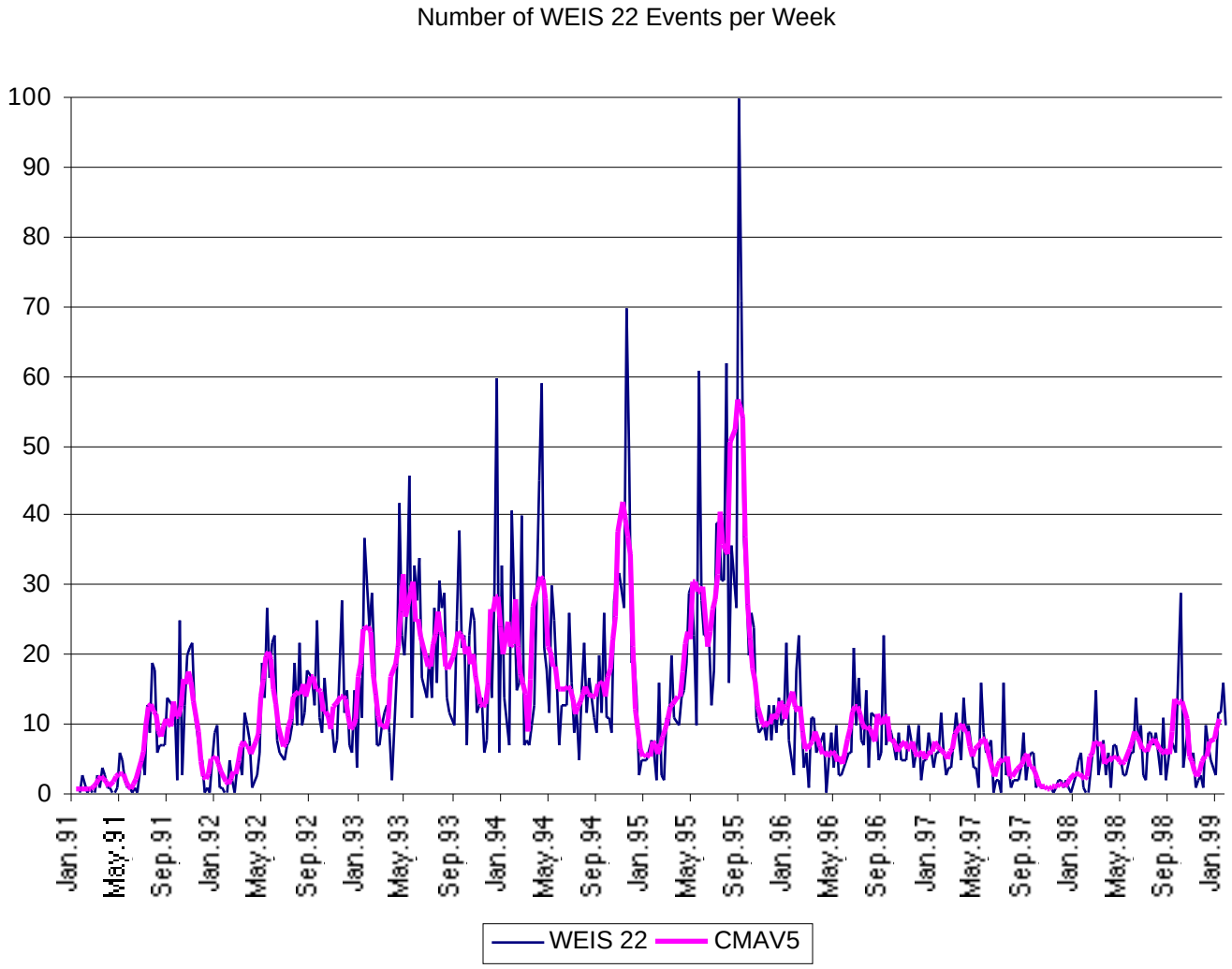

Number of WEIS 22 Events per Week
100
90
80
70
60
50
40
30
20
10
0
Jan.91
May.91
Sep.91
Jan.92
May.92
Sep.92
Jan.93
May.93
Sep.93
Jan.94
May.94
Sep.94
Jan.95
May.95
Sep.95
Jan.96
May.96
Sep.96
Jan.97
May.97
Sep.97
Jan.98
May.98
Sep.98
Jan.99
WEIS 22
CMAV5

Figure 8.

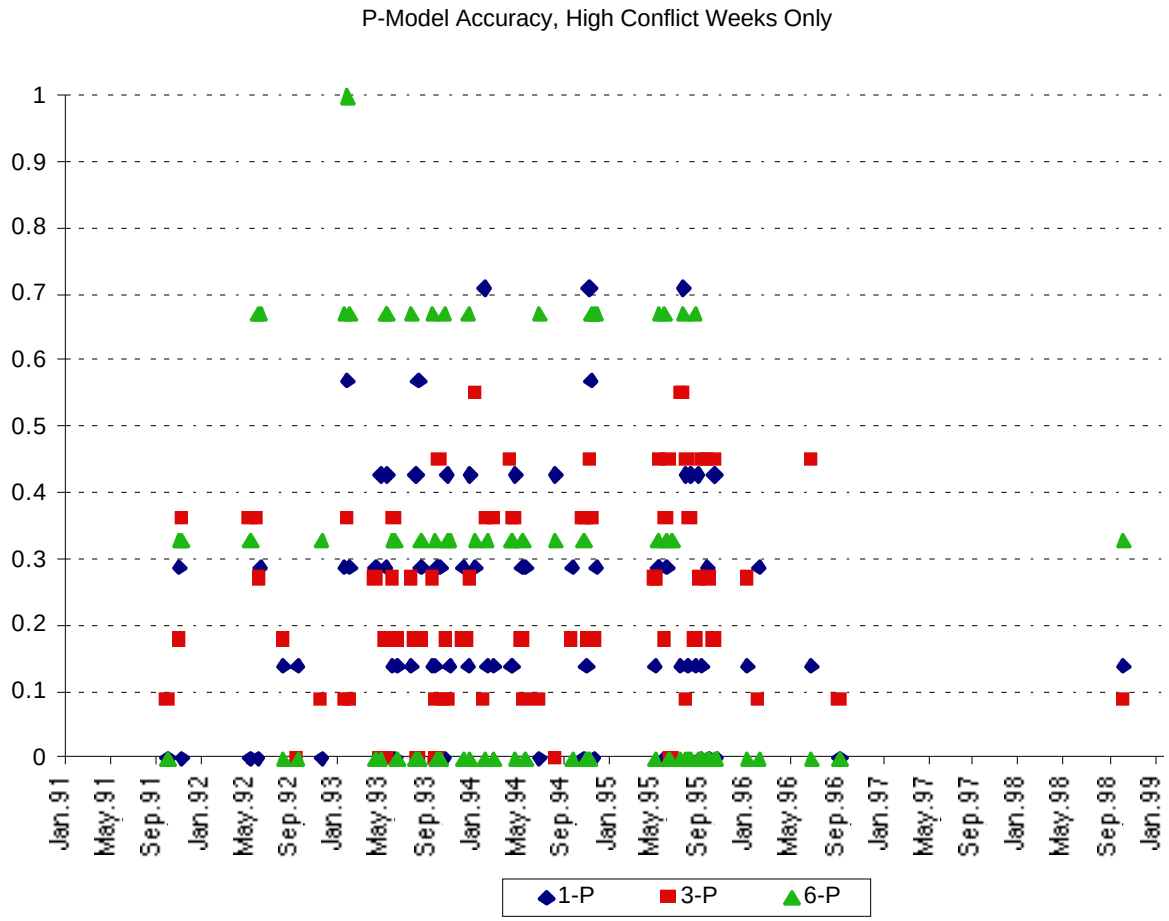
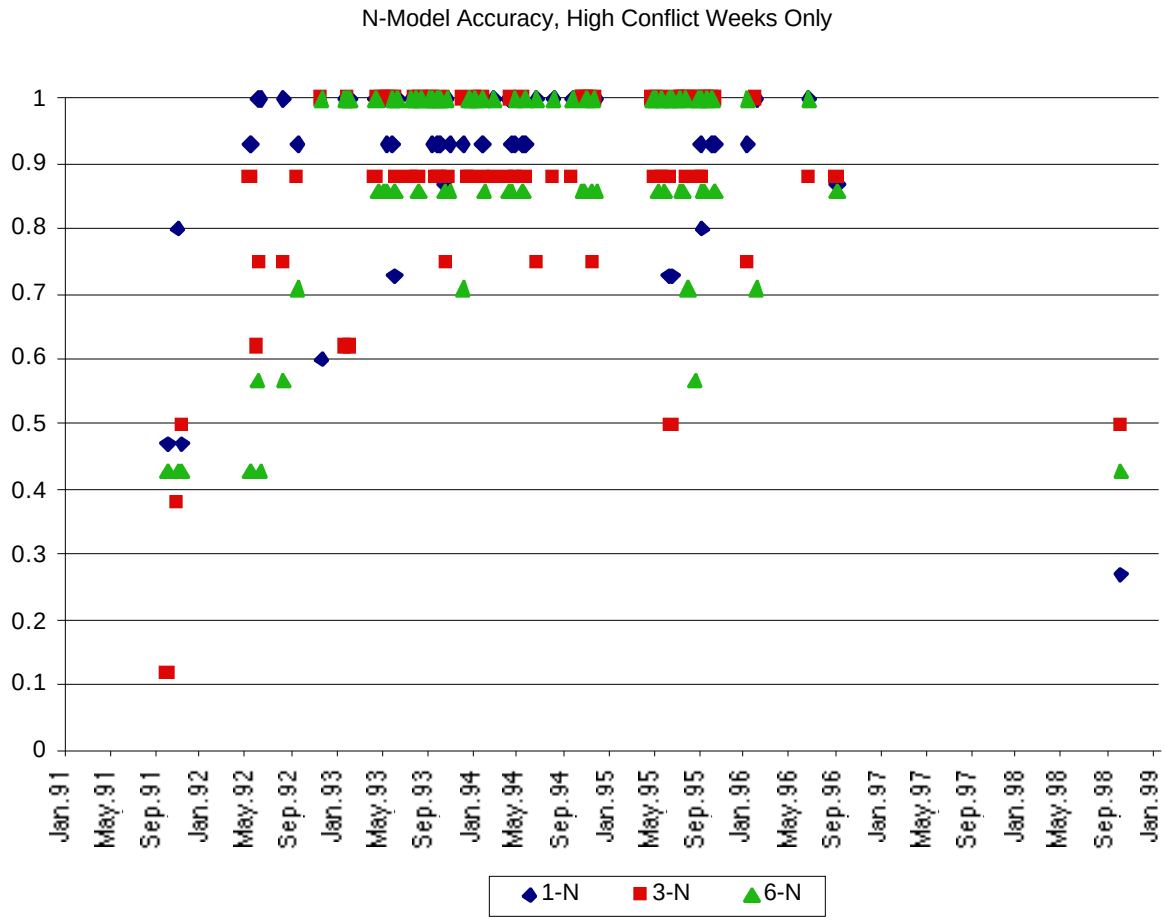

P-Model Accuracy, High Conflict Weeks Only
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Jan.91
May.91
Sep.91
Jan.92
May.92
Sep.92
Jan.93
May.93
Sep.93
Jan.94
May.94
Sep.94
Jan.95
May.95
Sep.95
Jan.96
May.96
Sep.96
Jan.97
May.97
Sep.97
Jan.98
May.98
Sep.98
Jan.99
1-P
3-P
6-P

Figure 9.


N-Model Accuracy, High Conflict Weeks Only
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Jan.91
May.91
Sep.91
Jan.92
May.92
Sep.92
Jan.93
May.93
Sep.93
Jan.94
May.94
Sep.94
Jan.95
May.95
Sep.95
Jan.96
May.96
Sep.96
Jan.97
May.97
Sep.97
Jan.98
May.98
Sep.98
Jan.99
1-N
3-N
6-N

Purely Predictive Models

All of the analyses reported above use templates from the entire period to develop the models. This means that some of the "predictions" of the models—on average, 50% of the period—are in fact retrospective in that they are based on templates that occur after the week that is being classified. In contrast, this section will evaluate two sets of models that are purely predictive: the templates will be chosen from the period prior to the weeks being evaluated.

Two different schemes were used to do this. The entire time period was subdivided by calendar years: 1993, 1994, 1995, 1996, 1997 and 1998. Let C_k refer to the beginning of the first full week of year k.

Prior Templates: Templates were taken from any time prior to C_k; predictions were evaluated on all of the weeks greater than C_k.

Recent Templates: Templates were taken from the time period $C_{k-1} \geq t > C_k$; predictions were evaluated on the weeks $C_k \geq t > C_{k+1}$.[12]

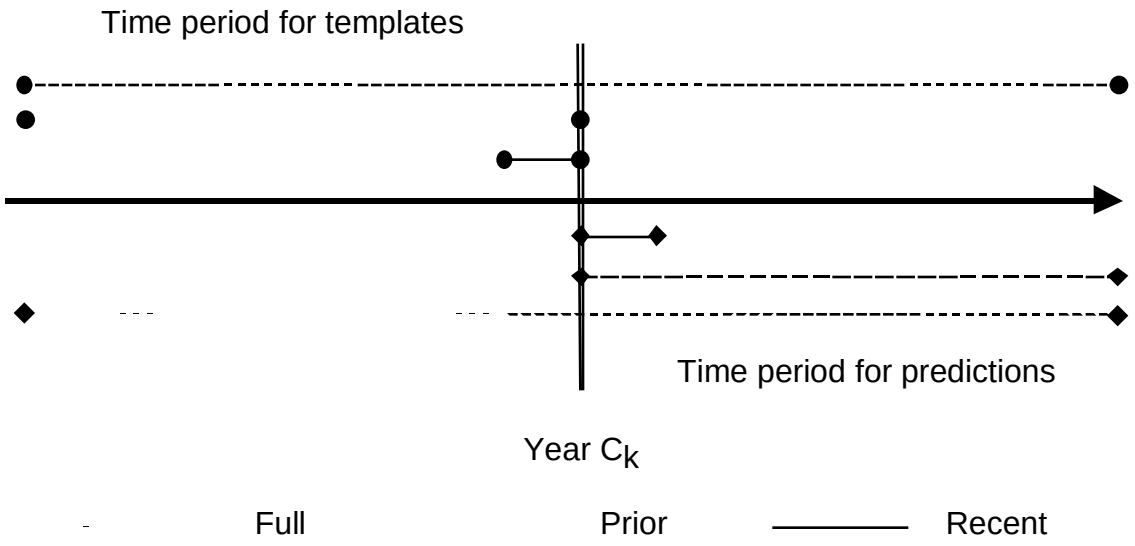
In other words, the "prior" scheme takes templates from any time before the beginning of a year, and then evaluates the accuracy of the prediction on all of the remaining weeks in the data set, whereas the "recent" scheme takes templates only from the previous year (where possible) and evaluates accuracy on a single year. The relationship between these schemes is illustrated in Figure 9.[13]

---

[12] Because of the small number of high-conflict weeks at the beginning and end of the sequence, there were not enough high-conflict weeks to provide an adequate number of templates (8) for some years, so this was implements by chosing high-conflict templates from 1991 and 1992 for the year 1993, and for 1996 to the beginning of the year for 1997 and 1998.

[13] Note: Due to a last minute reorganization of the paper, there is a gap in the table and figure numbers here.

Figure 9.
Prediction Schemes



These changes required relatively few changes in the program, but the estimation times increases by roughly a factor of six: Using 16 Monte-Carlo experiments per year, the prior templates estimation takes approximately 2000 minutes (33 hours) to run on a 350 Mhz Macintosh G3; the recent templates estimation takes approximately 1400 minutes (23 hours). [The difference is due to the fact that the recent model needs to classify only 52 weeks, whereas the prior model evaluates all 313 weeks]. Given these constraints, only the P and N models were evaluated.

The results of this analysis are reported in Tables 5 and 6. In general, both of the predictive analyses mirror the full-sample analysis in the sense that most of the figures are within ±20% of the earlier results. The differences between the P- and N-models that were found in the full-sample analysis continue to be reflected in the predictive analysis, though it seems to be somewhat more attenuated. There also seems to be more of a tendency for the 6-month forecast to be less accurate than the 1-month forecast, though these differences are frequently small (less than 10%) and the pattern is not universal.

As expected, these results differ substantially over time, with the high-conflict years 1993, 1994 and 1995 generally having one pattern (consistently better or worse predictions, depending on the model), 1996 being an intermediate years, and 1997 and 1998 having a single pattern. There are a few exceptions to this—for example the "%High Forecast" indicator for the P-model is uniformly awful—but it holds more often than not. [Also note that the sample sizes in the later

years of the prior analyses are relatively small—particularly for the high-conflict weeks—so the accuracy fluctuates more wildly than it does in the full-sample analysis.]

It should be noted that in many cases, the purely-predictive models perform *better* than the full-sample model, which is frequently not the case in statistical analyses. This is probably due to the model being able to adapt to the changing characteristics of the system, for example the shift in the focus of the conflict from Croatia to Bosnia to Kosovo, as well as adapting to the periods of low conflict. In some situations, this adaptation can be counter-productive, notably as the prior model adapts to the high-conflict period prior to 1996 and then finds almost no conflict to predict after 1996. This can lead to situations where there is a great deal of difference between the relative accuracy in predicting high-conflict and low-conflict weeks. But in a majority of the cases, the short-term adaptation produces substantially better predictions. This will be discussed further in the next section.

Table 5a. Accuracy of P-Models for Prior Forecast Templates and Predictions; Accuracy Computed on Full Period after Year

% Correct

	1993	1994	1995	1996	1997	1998
1	53.57	68.80	74.15	78.96	90.16	94.31
3	60.73	61.86	61.38	81.21	82.75	90.85
6	67.33	59.39	67.34	73.29	89.41	85.71

% High Correct

	1993	1994	1995	1996	1997	1998
1	39.55	27.93	20.83	35.00	12.50	6.25
3	29.11	31.52	35.19	33.75	31.25	0.00
6	23.54	48.54	29.63	40.00	25.00	25.00

% Low Correct

	1993	1994	1995	1996	1997	1998
1	57.76	77.61	81.89	80.37	90.89	95.91
3	70.18	68.41	65.19	82.73	83.24	92.50
6	80.43	61.73	72.82	74.36	90.01	86.82

% High Forecast

	1993	1994	1995	1996	1997	1998
1	21.89	21.19	14.31	5.41	1.27	2.70
3	22.61	17.70	12.79	5.90	1.71	0.00
6	26.47	21.47	13.66	4.76	2.29	3.33

% Low Forecast

	1993	1994	1995	1996	1997	1998
1	76.16	83.32	87.69	97.47	99.11	98.25
3	76.79	82.25	87.39	97.50	99.23	98.07
6	77.86	84.76	87.70	97.48	99.23	98.45

Table 5b. Accuracy of N-Models for Prior Forecast Templates and Predictions; Accuracy Computed on Full Period after Year

% Correct

	1993	1994	1995	1996	1997	1998
1	53.86	40.24	37.35	28.14	42.42	69.53
3	55.93	26.30	26.73	23.10	30.32	56.70
6	67.82	32.08	25.29	21.16	27.84	48.55

% High Correct

	1993	1994	1995	1996	1997	1998
1	44.26	84.18	82.87	96.25	50.00	37.50
3	37.59	91.22	90.05	83.75	56.25	37.50
6	23.20	89.63	93.75	93.75	87.50	56.25

% Low Correct

	1993	1994	1995	1996	1997	1998
1	56.74	30.76	30.75	25.96	42.35	70.11
3	61.42	12.30	17.54	21.15	30.08	57.05
6	81.17	19.67	15.36	18.83	27.28	48.41

% High Forecast

	1993	1994	1995	1996	1997	1998
1	23.44	20.77	14.80	4.00	0.80	2.23
3	22.57	18.32	13.68	3.29	0.75	1.56
6	26.94	19.39	13.85	3.57	1.11	1.94

% Low Forecast

	1993	1994	1995	1996	1997	1998
1	77.29	90.02	92.52	99.54	98.91	98.41
3	76.69	86.67	92.39	97.60	98.66	98.05
6	77.94	89.79	94.42	98.95	99.57	98.38

Table 6a. Accuracy of P-Models for Recent Templates; Prediction Accuracy by Year

% Correct

	1993	1994	1995	1996	1997	1998
1	49.28	52.52	54.81	79.25	96.27	96.75
3	48.80	52.16	47.60	75.12	93.87	97.96
6	47.96	47.12	54.93	51.65	95.43	94.35

% High Correct

	1993	1994	1995	1996	1997	1998
1	33.41	41.25	42.33	28.12	0.00	0.00
3	16.83	55.94	27.27	28.12	0.00	0.00
6	9.13	57.81	40.06	29.69	0.00	0.00

% Low Correct

	1993	1994	1995	1996	1997	1998
1	65.14	59.57	63.96	83.42	96.27	98.64
3	80.77	49.80	62.50	78.95	93.87	99.88
6	86.78	40.43	65.83	53.44	95.43	96.20

% High Forecast

	1993	1994	1995	1996	1997	1998
1	48.94	38.94	46.27	12.16	0.00	0.00
3	46.67	41.06	34.78	9.84	0.00	0.00
6	40.86	37.76	46.23	4.95	0.00	0.00

% Low Forecast

	1993	1994	1995	1996	1997	1998
1	49.45	61.87	60.20	93.43	100.00	98.05
3	49.27	64.39	53.96	93.08	100.00	98.07
6	48.85	60.53	59.96	90.30	100.00	98.00

Table 6b. Accuracy of N-Models for Recent Templates; Prediction Accuracy by Year

% Correct

	1993	1994	1995	1996	1997	1998
1	48.08	42.07	47.00	52.95	76.20	94.59
3	47.96	43.03	44.23	39.98	74.16	97.36
6	47.12	41.59	49.64	28.89	77.04	94.11

% High Correct

	1993	1994	1995	1996	1997	1998
1	42.55	73.75	83.24	53.12	0.00	0.00
3	31.73	81.56	73.3	62.50	0.00	0.00
6	17.07	86.88	91.19	76.56	0.00	6.25

% Low Correct

	1993	1994	1995	1996	1997	1998
1	53.61	22.27	20.42	52.93	76.20	96.45
3	64.18	18.95	22.92	38.14	74.16	99.26
6	77.16	13.28	19.17	25.00	77.04	95.83

% High Forecast

	1993	1994	1995	1996	1997	1998
1	47.84	37.22	43.41	8.44	0.00	0.00
3	46.98	38.61	41.08	7.62	0.00	0.00
6	42.77	38.5	45.28	7.69	0.00	2.86

% Low Forecast

	1993	1994	1995	1996	1997	1998
1	48.27	57.58	62.42	93.26	100.00	98.01
3	48.46	62.18	53.92	92.57	100.00	98.06
6	48.2	61.82	74.80	92.89	100.00	98.12

Comparison of Forecast Accuracy by Year

The "recent" prediction scheme is the most sensitive to the changing character of the conflict. The annual statistics in Table 6 are also the most straightforward to evaluate, because each prediction is done on 52-weeks, in contrast to the variable number of weeks in the prediction period in the "prior" scheme.

In order to get a systematic evaluation of the effect of changing the level of adaptation in the model, the results from the full-sample and prior analysis were re-analyzed on an annual basis: In other words, rather than assessing the accuracy on a period that extends to the end of the data set, the accuracy for each year (52 weeks) of the data was tabulated. (The *optimization* of the model was done as before—in fact this is simply a retabulation of the existing results, not a new set of estimations.) These results are reported in Tables 7 and 8, which are directly comparable to Table 6.

Figures 10, 11 and 12 show the comparative level of accuracy of the three methods for the 3-month lead. When the relative level of accuracy is compared across the estimation methods, two patterns are evident in these figures.

First, there is generally a single rank ordering of the accuracy of the three methods across time: For example, if the recent and prior schemes (or full-sample and recent schemes) have roughly the same level of accuracy on one year, they will have this on all years. The exceptions that occur in this are usually at the ends of the data set, 1993 and 1998. There is quite a bit of variation in the patterns across the accuracy measures and estimation techniques, however.

Second, the full-sample scheme is almost always better than the prior scheme; when this is no true, the prior scheme and the recent scheme usually have about the same level of accuracy. In all cases, the prior scheme is either the least accurate of the three—in some cases dramatically less accurate, as in Figure 12b , or else it is comparable in value to one of the other two measures.

These results suggest rather strongly that if one is in a predictive mode (as distinct from retrospectively analyzing a period of time, or using one set of interactions to try to predict behavior in a different region), then it is best of use short-term adaptation. In many cases, the short-term model does also as well as the full-sample model, and in Figure 10b, it does substantially better in the later years. HMMs seem to work best when they can "forget"—in the sense of ignoring older information—as well as "learn."

The optimal set of training templates might be one that effective has an "exponential decay" in the sense of having more templates from recent history than from distant history. To a certain extent, this was done already in the latter years in this analysis, since it was necessary to go further

back in time to pick up sufficient examples of high-conflict weeks.[14] Simply accumulating additional information on an on-going conflict as it becomes available—which is effectively what the "prior" scheme does—does not seem to be a good idea: from a statistical perspective, these are non-stationary systems. On the other hand, one would still like to have some way of maintaining a "memory" of the precursors to high-conflict situations in a situation which has been peaceful for some time; this might be provided either by instances of earlier conflict in the region (even if it occurred a number of years earlier), or with some generic templates for conflict in comparable regions. In Schrodt (1999) I was able to use hidden Markov models generated from a set of 19th and 20th century crises (the BCOW data set) to provide a reasonably good measure of conflict in the Israel-Palestine conflict, and based on this, it might be possible to find some good archetypical models for conflict in areas that have generally been conflict-free.

---

[14] However, this result has certainly been affected to some extent by the fact that the Balkans conflict went through four or five very distinct subphases—the initial conflict with Croatia, followed by the conflict with Bosnia, followed by the combined Croatia-Bosnian counter-offensive and brief NATO attacks, followed by the peaceful Dayton period, and ending with the lead-up to the Kosovo conflict. Because the coding system is sensitive to the presence or absence of activity by individual actors, it may to overly sensitive to these shifts and unable to pick up a general pattern for "conflict." It would be interesting to compare these results with a model of a highly-institutionalized conflict such as Israel-Lebanon or Turkey-Kurds.

Figure 10a
%Correct by Year, 3P Models

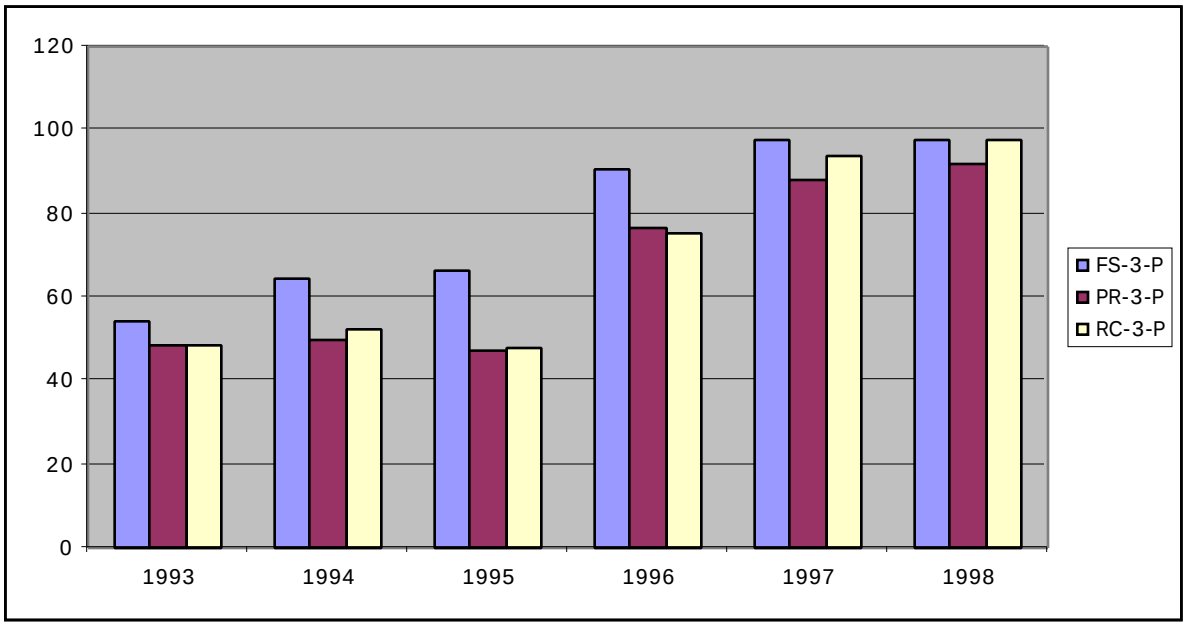


Figure 10b
% Correct by Year, 3N Models

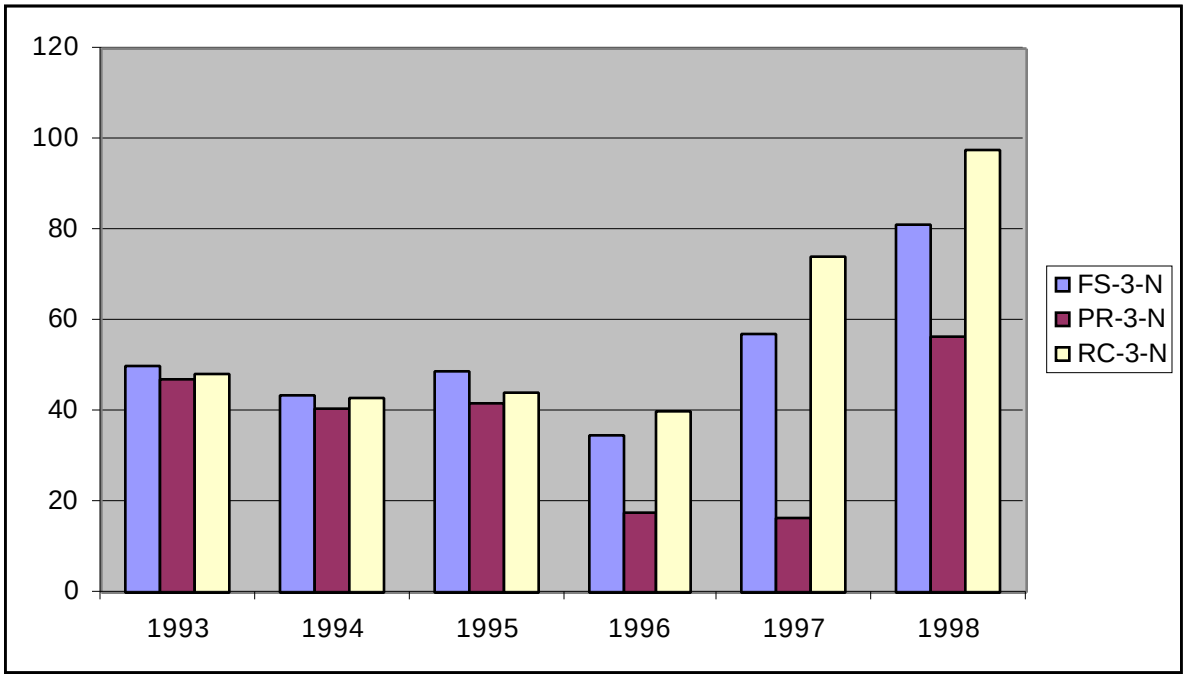


Figure 11a
% High Correct by Year, 3P Models

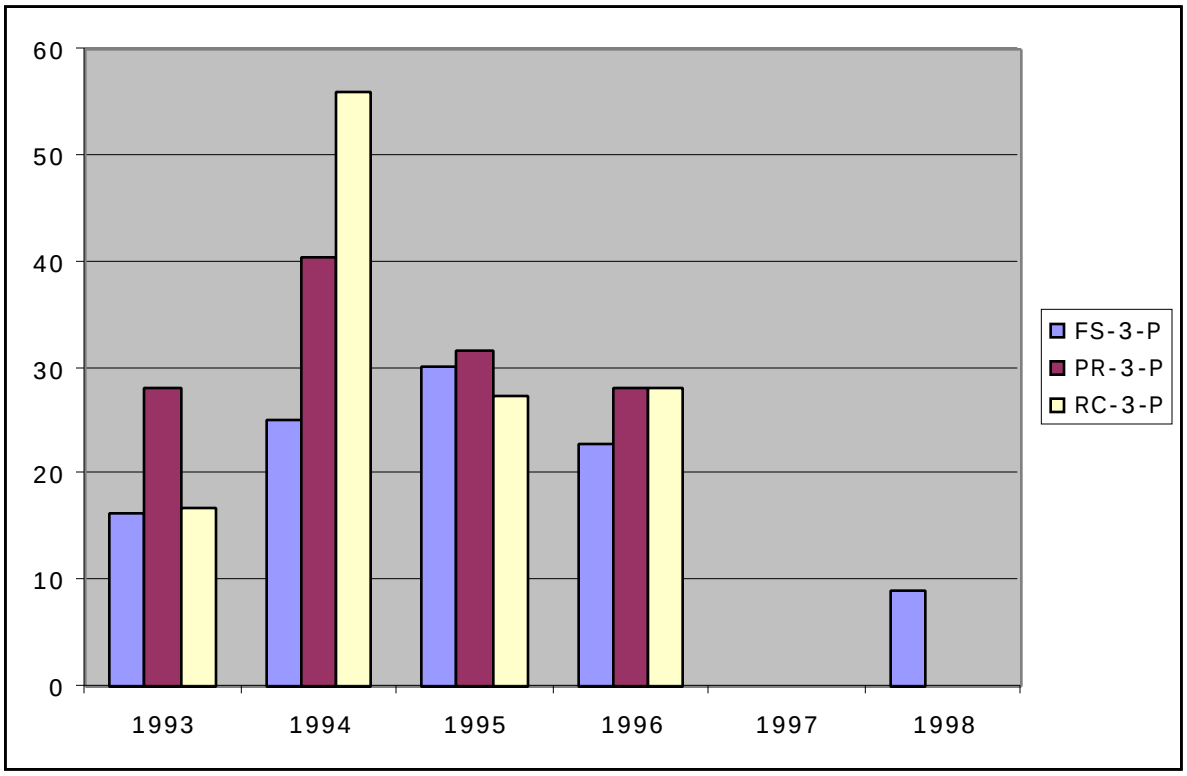


Figure 11b
% High Correct by Year, 3N Models

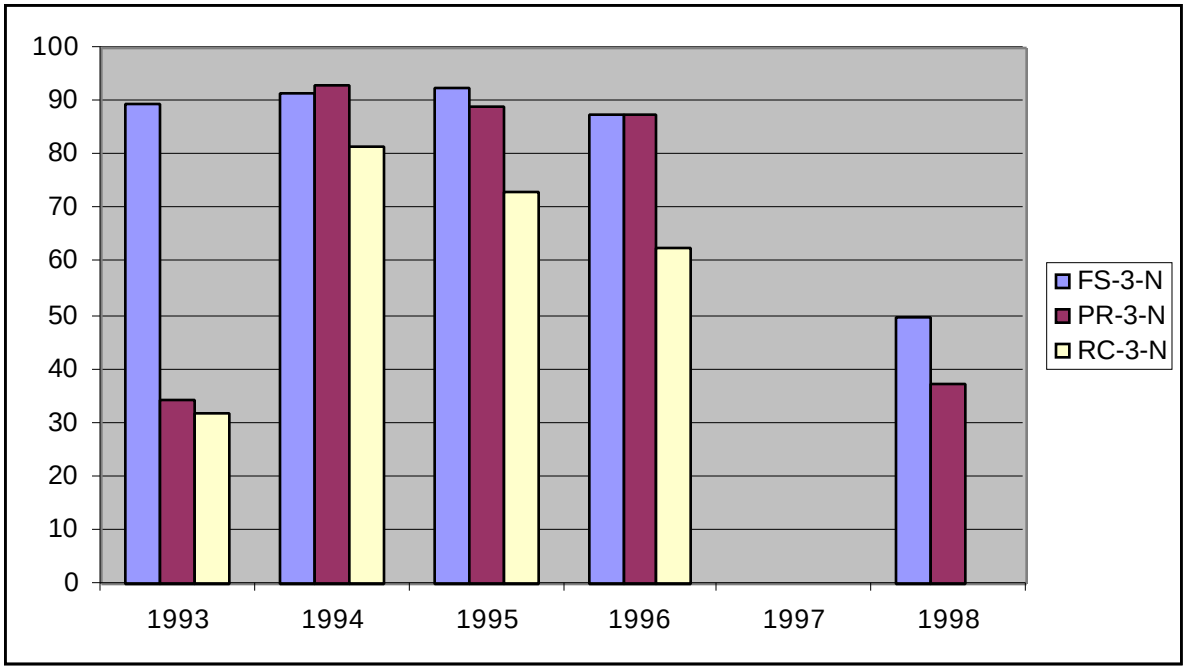


Figure 12a
% Low Correct by Year, 3P Models

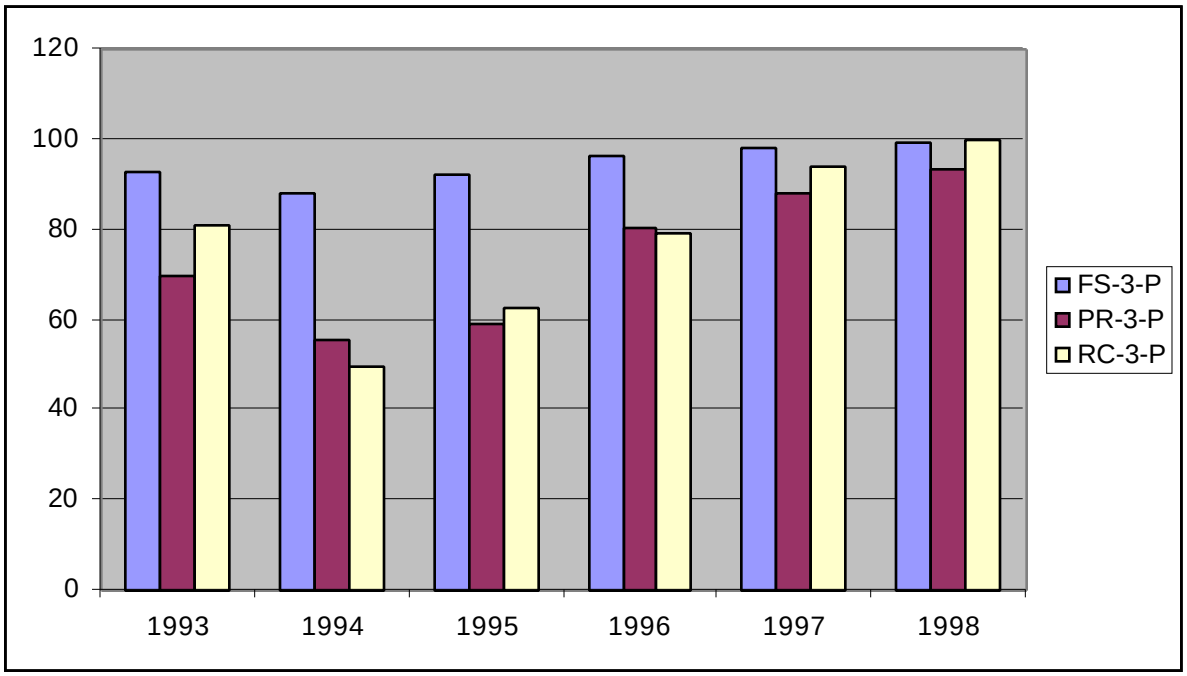


Figure 12b
% Low Correct by Year, 3N Models

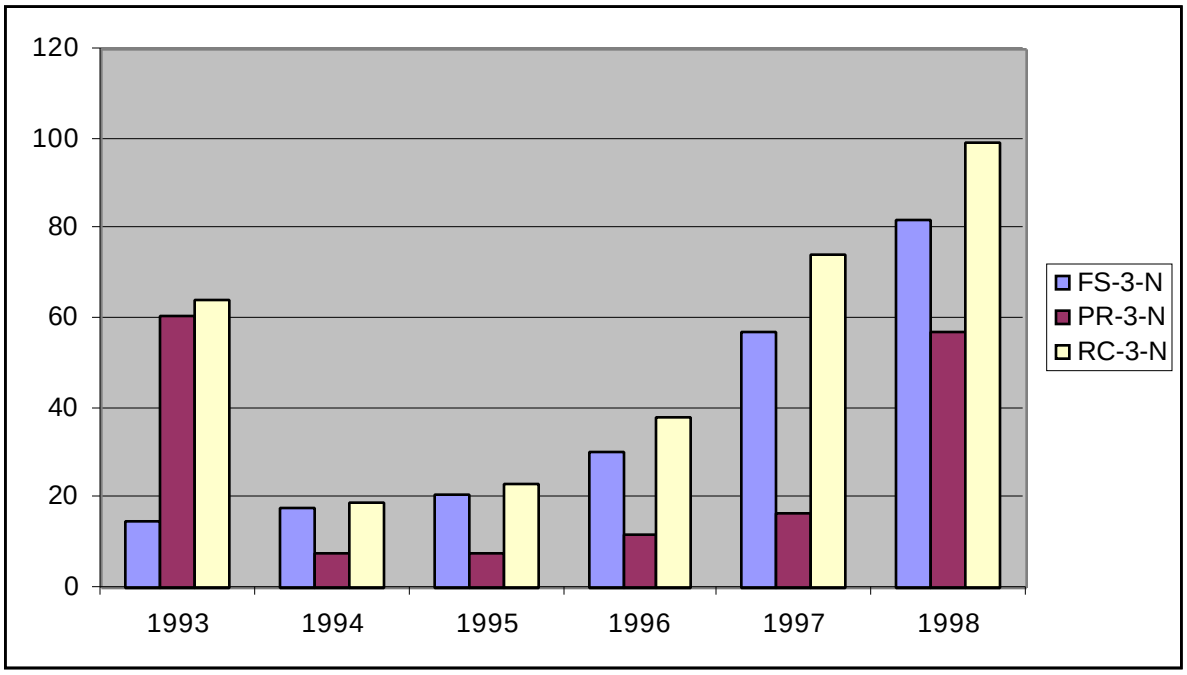


Table 7a. Accuracy of P-Models for Prior Sample Templates; Prediction Accuracy by Year

% Correct

	1993	1994	1995	1996	1997	1998
1	48.44	51.08	52.88	67.45	95.31	94.11
3	48.80	49.88	47.36	76.30	88.34	91.71
6	47.00	49.52	53.37	65.57	93.39	87.74

% High Correct

	1993	1994	1995	1996	1997	1998
1	39.18	35.31	19.03	34.38	---	6.25
3	28.12	40.31	31.53	28.12	---	0
6	23.32	56.56	29.26	39.06	---	25.00

% Low Correct

	1993	1994	1995	1996	1997	1998
1	57.69	60.94	77.71	70.15	95.31	95.83
3	69.47	55.86	58.96	80.23	88.34	93.50
6	70.67	45.12	71.04	67.73	93.39	88.97

% High Forecast

	1993	1994	1995	1996	1997	1998
1	48.08	36.10	38.51	8.59	0	2.86
3	47.95	36.34	36.04	10.40	0	0
6	44.29	39.18	42.56	8.99	0	4.26

% Low Forecast

	1993	1994	1995	1996	1997	1998
1	48.68	60.12	56.69	92.91	100	98.12
3	49.15	59.96	54.01	93.19	100	97.95
6	47.96	62.43	57.80	93.16	100	98.37

Table 7b. Accuracy of N-Models for Prior Sample Templates; Prediction Accuracy by Year

% Correct

	1993	1994	1995	1996	1997	1998
1	46.15	39.30	42.91	14.98	31.01	69.95
3	47.24	40.75	41.95	17.45	16.47	56.49
6	46.27	42.43	45.67	17.57	23.56	48.56

% High Correct

	1993	1994	1995	1996	1997	1998
1	41.35	87.19	83.81	98.44	---	37.50
3	34.13	93.12	88.92	87.50	---	37.50
6	24.76	90.62	94.60	93.75	---	56.25

% Low Correct

	1993	1994	1995	1996	1997	1998
1	50.96	9.38	12.92	8.16	31.01	70.59
3	60.34	8.01	7.50	11.73	16.47	56.86
6	67.79	12.3	9.79	11.35	23.56	48.41

% High Forecast

	1993	1994	1995	1996	1997	1998
1	45.74	37.55	41.37	8.05	0	2.44
3	46.25	38.75	41.35	7.49	0	1.68
6	43.46	39.24	43.47	7.95	0	2.09

% Low Forecast

	1993	1994	1995	1996	1997	1998
1	46.49	53.93	52.10	98.46	100	98.29
3	47.81	65.08	48.00	92.00	100	97.89
6	47.39	67.74	71.21	95.70	100	98.26

Table 8a. Accuracy of P-Models for Full Sample Templates; Prediction Accuracy by Year

% Correct

	1993	1994	1995	1996	1997	1998
1	56.32	65.11	62.64	90.03	100	97.53
3	54.55	64.16	66.61	90.57	97.90	97.38
6	62.18	64.74	62.18	87.42	98.72	95.51

% High Correct

	1993	1994	1995	1996	1997	1998
1	25.56	25.18	20.27	14.29	---	14.29
3	16.43	25.23	30.08	22.73	---	9.09
6	32.89	25.00	16.67	0.00	---	33.33

% Low Correct

	1993	1994	1995	1996	1997	1998
1	86.41	89.78	91.67	96.21	100.00	99.16
3	92.66	88.14	92.26	96.10	97.90	99.11
6	90.00	89.58	95.56	94.56	98.72	96.73

% High Forecast

	1993	1994	1995	1996	1997	1998
1	64.79	60.34	62.5	23.53	---	25.00
3	69.12	56.70	73.20	32.26	0	16.67
6	75.76	60	73.33	0	0	16.67

% Low Forecast

	1993	1994	1995	1996	1997	1998
1	54.27	66.01	62.66	93.22	100	98.33
3	52.58	65.68	65.26	93.84	100	98.23
6	58.54	65.65	60.99	92.05	100	98.67

Table 8b. Accuracy of N-Models for Full Sample Templates; Prediction Accuracy by Year

% Correct

	1993	1994	1995	1996	1997	1998
1	50.26	40.38	46.28	30.19	70.00	93.46
3	50.00	43.75	48.80	34.43	57.21	81.25
6	54.40	45.60	48.08	33.69	55.77	73.90

% High Correct

	1993	1994	1995	1996	1997	1998
1	95.86	97.79	94.70	95.00	---	26.67
3	89.74	91.72	92.59	87.50	---	50.00
6	92.44	92.91	90.14	89.29	---	42.86

% Low Correct

	1993	1994	1995	1996	1997	1998
1	10.77	9.82	15.69	24.90	70.00	94.77
3	14.93	18.08	20.87	30.10	57.21	81.86
6	20.31	20.25	21.17	29.15	55.77	74.51

% High Forecast

	1993	1994	1995	1996	1997	1998
1	48.19	36.60	41.51	9.36	0	9.09
3	48.21	37.46	42.74	9.27	0	5.13
6	50.96	38.44	42.24	9.33	0	3.19

% Low Forecast

	1993	1994	1995	1996	1997	1998
1	75.00	89.29	82.42	98.39	100	98.51
3	62.26	80.33	81.54	96.72	100	98.82
6	75.00	84.21	77.05	97.09	100	98.52

Assessment of differences between the high-conflict and low-conflict models

The HMM has a very diffuse coefficient structure: the high and low models each have 1,078 parameters (177 observation probabilities plus 16 transition probabilities), so determining how to predict a week fits the high or low conflict category involves 2,156 parameters. Add the fact that this analysis has looked at three different weighting schemes and three forecasting horizons, and one is looking at 19,404 parameters; if the prospective and recent models are included, there are 58,212 parameters. Add the fact that we are using Monte Carlo methods, so that each of those conditions was duplicated at least 16 times, and the analysis has actually produced in excess of 931,392 parameter estimates. That's a lot of parameters...

On the positive side, these parameters all have straightforward interpretations (in contrast, for example, to the weights of a neural network). The larger observation probabilities correspond to the behaviors that are being "watched" by the model in order to make a prediction. In this analysis, the transition probabilities are of interest primarily to determine the number of Markov states that are actually found in the process. The *differences* between the high and low models will show the characteristics of the system that are most likely to distinguish between pre-conflict and pre-peace periods.

In order to make sense of this mass of information, I'm focusing the analysis somewhat:

- Only the full-sample runs have been analyzed—these are likely to be most representative of the results of the entire sample.
- Only the P and N models are compared—as in the prediction analysis, the parameter estimates of the U and P models are very similar and with very few exceptions, anything true of the P model will also be true of the N model.
- Most of the analysis will focus on the 3-month forecast—there are some differences between the forecasting horizons, and these will be explored below—but in general the major differences are found between the high and low conflict models, and between the P and N weighting systems.
- Most of the analyses will be done using the "reduced" models (see discussion below)

Because the parameter estimates result from Monte Carlo estimation, they will be analyzed statistically. As before, the "All" results refer to the analysis of all of the Monte Carlo runs; the "Best" results are only those with accuracy ≥ 0.795 for P models and accuracy ≥ 0.495 for N models.

State Reduction Criteria

This section gives details on the algorithm used to reduce the number of states in a model through elimination and combinion of states that will occur very rarely. In this discussion, the notations p_i, r_i, and n_i refer to the transitions to the previous state, same state, and next state respectively.

"Small" probabilities are—somewhat arbitrarily—defined as those less than 0.1; "large" probabilities are defined as those greater than 0.9.[15];

Small values of r_i occur in two circumstances:

• transient states: r_i is small; p_i and p_{i+1} are not large. In this case, the system will simply pass through the state quickly and that state will be responsible for very few events in the sequence.

• cyclical pairs: r_i and r_{i+1} are small; n_i and p_{i+1} are large. In this case, the system will remain inside this pair of states for a significant period of time, but switch between state i and i+1 for every other event

While other combinations are logically possible—for example one could have a cyclical triplet—virtually all of the cases of small r_i values fall into these categories.

The transient case can be simply eliminated from the model. The only situation where this removes any important information is when p_i is very small, in which case the state is acting as a one-way valve—once the system has passed through the state, it will not return to the earlier states. These situations are fairly common—in fact in a number of cases the estimated p_i is zero—but I have not tried to analyze them separately.

The cyclical pair, in contrast, can be mathematically reduced to a single state. The pair forms a Markov chain of the form

$$p_i \longleftrightarrow \begin{bmatrix} r_i & n_i \\ p_{i+1} & r_{i+1} \end{bmatrix} \longleftrightarrow n_{i+1}$$

Because $r_i + n_i < 1$ and $p_{i+1} + r_{i+1} < 1$, the middle matrix is not strictly a Markov chain, so these are standardized by row, e.g. r_i is replaced with $\frac{r_i}{r_i + n_i}$.

A Markov chain relatively quickly reaches an equilibrium distribution where it spends a known

---

[15] r_i has a highly bimodel distribution...

proportion of time in each of the states. Feller (1968: 432) shows that the Markov chain

$$\begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

has the equilibrium distribution $[\frac{q}{p+q}, \frac{p}{p+q}]$. This formula will be used to give the weights to the observation probabilities when reducing cyclical states. (In practice, n_i and p_{i+1} are sufficiently close in value that this gives much the same results as simply weighting the two observation vectors equally)

Interpreting the t-test tables

Much of this analysis is based t-test tables such as the following:

HIGH CONFLICT
PARAMETER T-TESTS AGAINST MARGINALS FOR MODEL P3
State 0

	BOS>	>BOS	CRO>	>CRO	SER>	>SER	KSV>	>KSV	
YIELD	-5.127	***	---	---	+++	-3.948	---	---	
COMMNT		---	8.004	---	2.495	2.321	12.233	-4.772	-3.849
CONSLT		---	7.845	---	---	2.329	+++	-6.469	-8.034
APPROV		---	-3.651	---	+++	---	-4.874	---	-11.946
PROMIS		---	---	---	---	---	-3.150	---	-3.085
GRANT		-2.735	-2.976	---	---	+++	---	---	-2.719
REWARD		---	---	---	---	---	---	---	-4.266
AGREE		-3.535	-3.592	-5.826	-7.120	-3.418	-2.324	-13.352	-10.067
REQEST		---	+++	+++	-2.234	-2.466	-5.893	---	-2.247
PROPOS		-6.552	---	---	+++	---	---	---	-4.674
REJECT		---	---	+++	---	3.707	+++	---	-2.901
ACCUSE		+++	---	---	3.033	2.546	---	---	---
PROTST		-2.252	-3.447	+++	---	-2.396	---	---	-11.946
DENY	+++	-3.255	---	---	2.502	-6.648	---	+++	
DEMAND		---	---	+++	---	---	---	---	---
WARN		-2.141	---	-2.679	***	---	-3.734	---	---
THREAT		-2.636	---	---	---	---	---	---	---
REDUCE		---	---	---	---	---	---	---	-2.531
EXPEL	***	---	---	---	---	---	---	-8.707	
SEIZE	---	---	---	---	***	---	---	---	
FORCE		---	-2.262	-3.914	---	---	-2.341	---	---
NULL	---	-2.135	---	---	---	-3.501	-3.000	-8.602	
NONEVT		-5.414							

These tables are designed to make it relatively easy to determine where the significant relations in the estimated observation probabilities exist. The first three lines—whose content will vary with the individual test—identify the test and the observation vector: In this case the vector is for State 0 of the P weights, 3-month forecast, high-conflict model; the t-test compares the estimated probabilities with the marginal probabilities.

The matrix of t-test has the event codes in the rows and the dyadic interactions *in the columns. The rows are labeled with the WEIS codes (these are either mnemonics, as in this example, or two-digit numerical codes in some of the earlier runs). The columns are labeled with the dyads: the notation BOS> refers to all actions where Bosnia is the *source*; >BOS refers to all actions where Bosnia is the *target*. (the notation <BOS is used in some earlier versions of the program). The abbreviations are BOS = Bosnia; CRO = Croatia; SER = Serbia; KSV = Kosovo.
Only the significant t-tests are shown. The following thresholds are used:

$\|t\| > 2.0$	Print the actual value; this is approximately the critical value for the 0.05 significance level
$2.0 \geq \|t\| > 1.72$	Print "***" ; this is approximately the critical value for the 0.10 significance level
$1.72 \geq \|t\| > 1.32$	Print "+++"; this is approximately the critical value for 0.20 significance level
$1.32 \geq \|t\|$	Print "---"

The selected critical values are for a t-test with 20 degrees-of-freedom. Note that the sample size varies depending on the individual model, so these critical values are only approximate; in the case of some of the "best" tests, they slightly over-estimate the significance level.[16]

When the reduced forms of the high and low models are being compared, there is an additional problem of comparing the different number of models where a particular state exists (in other words, because of state reduction, there will be fewer instances of models containing four, five or six states than models having three states). There [obviously] must be at least two cases of a particular state in both models (e.g. high and low; 1-month and 6-month) for the t-test to be computed.

Interpreting the t-test maps

In order to provide a means of visualizing all of the parameters simultaneously, this report includes a number of color-coded maps that show the t-tests for all of the codes and dyads. The

---

[16] It would be straightforward to adjust these tables so that they reflect true significance levels, but since the levels of significance tests are themselves rather arbitrary—contrary to popular belief in social science statistical circles, there is little evidence that Moses descended from Mt. Sinai bearing tablets that said "$\alpha < 0.05!$"—this didn't seem to justify the trouble involved in entering a t table into the program used to generate these tables. In all cases, the highest t-values are substantially greater than 2.0, and more generally, the number of significant t-values far exceeds what would be expected by chance under the null hypothesis of no differences, so the critical value of 2.0 is not particularly important in the analysis.

red areas indicate t-test that have high positive values; the blue areas have high negative values; and the green areas are close to zero.

These maps are produced from surface-plots in Microsoft Excel, which introduces a few quirks. First, the "points" being plotted are the *intersections* of the grid line, not the squares outlined by the grid (ideally, these plots should be done as a grid of appropriately colored squares, but this isn't an option in Excel). Excel also interpolates between adjacent points, so there may be several color bands between two values. Technically, the codes are discrete so this is meaningless, though because WEIS codes form a rough cooperation-to-conflict continuum, it makes some sense in the horizontal direction (and probably enhances one's ability to make sense of the map). Second, only eight ranges are plotted, and in fact the "3 - 4" range is actually >3.0 for positive numbers and < -3.0 for negatives. Finally, in order to keep the scaling consistent across the graphs, in a few cases I put an artificial "-3.0x" value in the KSV>:YIELD category.

Parameter comparisons

This section will discuss each of the following comparisons.

- Number of states in the reduced models;
- Pattern of coefficients that are significantly different from the observed distribution of the codes in the data;
- Pattern of coefficients that are significantly different in the high and low models
- Comparison of 1-month and 6-month forecast models

The t-test tables themselves are very lengthy—each requires about 2500 lines, or roughly 40 single-spaced pages for the full model—so with a few exceptions these have not been included in the paper. In some instances, summary-maps of the distributions will be used instead.

Number of states in the models

As noted above, both transient and cyclic states occurred frequently in the estimates, indicating—with one possible exception—that the six-state framework was adequate to account for the observed sequences. The tables that follow show the proportion of non-transient and cyclic states in original matrices, and the average number of states in the reduced matrices, by weight and forecast period.

As usual, the U/P and N models tend to mirror each other. In the P models, the high model has about 20% cyclic states, while the low model has only about 1% to 2%; in the N model this ratio is reversed. In all of the low models, only about 60% to 70% of the states in the low model

are non-transient (that is, neither transient nor cyclical); the high model in the U/P wieghts has even fewer non-transient state, but in the N model around 90% of the states are non-transient.

These characteristics are reflected in the average number of states in the reduced models. All of the models except the N-weight high model average about 4.5 states. In most cases, the low models have fewer states than the high models, though this difference is generally only around a 0.3 to 0.6 states. The standard deviations are relatively low.[17]

The one exception is the N-type, high model, where the reduced models still have nearly six states. This would certainly suggest that moving to a larger model for this situation (or for models with similar weights) might be appropriate. This may be due to the fact that the N-type models specialize in accurately forecasting the more complex, high-conflict category, which may require the greater level of detail that is possible with a model with a greater number of states. P-type models, in contrast, are better with low-conflict weeks, and there is a consistent tendency for low-conflict models to require fewer states than high-conflict models.

[17] Because the distribution is bounded on the right—no model has more than six states—this is probably not normally distributed. The standard deviations of the "best" models are actually higher than those of the complete set of models, which runs contrary to my expectations; this may be partly a function of the low sample sizes.

Table 9a.
Proportion of Non-transient and cyclic states in original matrices
All Models

Model	Total states	Non-transient	Cyclic	%Non-transient	%Cyclic
U1					
HIGH:	132	69	25	52.27%	18.94%
LOW:	132	91	1	68.94%	0.76%
U3					
HIGH:	126	65	26	51.59%	20.63%
LOW:	126	81	2	64.29%	1.59%
U6					
HIGH:	96	57	14	59.38%	14.58%
LOW:	96	74	0	77.08%	0.00%
P1					
HIGH:	174	93	37	53.45%	21.26%
LOW:	174	115	8	66.09%	4.60%
P3					
HIGH:	324	184	69	56.79%	21.30%
LOW:	324	204	5	62.96%	1.54%
P6					
HIGH:	96	55	18	57.29%	18.75%
LOW:	96	72	1	75.00%	1.04%
N1					
HIGH:	96	87	1	90.62%	1.04%
LOW:	96	55	14	57.29%	14.58%
N3					
HIGH:	96	92	1	95.83%	1.04%
LOW:	96	51	17	53.12%	17.71%
N6					
HIGH:	96	89	2	92.71%	2.08%
LOW:	96	53	21	55.21%	21.88%

Table 9b.
Proportion of Non-transient and cyclic states in original matrices
Best Models Only

Model	Total states	Non-transient	Cyclic	%Non-transient	%Cyclic
U1					
HIGH:	60	30	11	50.00%	18.33%
LOW:	60	38	1	63.33%	1.67%
U3					
HIGH:	42	21	9	50.00%	21.43%
LOW:	42	30	2	71.43%	4.76%
U6					
HIGH:	12	8	1	66.67%	8.33%
LOW:	12	8	0	66.67%	0.00%
P1					
HIGH:	42	26	8	61.90%	19.05%
LOW:	42	27	2	64.29%	4.76%
P3					
HIGH:	66	41	12	62.12%	18.18%
LOW:	66	49	1	74.24%	1.52%
P6					
HIGH:	18	12	3	66.67%	16.67%
LOW:	18	12	0	66.67%	0.00%
N1					
HIGH:	90	81	1	90.00%	1.11%
LOW:	90	53	13	58.89%	14.44%
N3					
HIGH:	48	46	0	95.83%	0.00%
LOW:	48	27	7	56.25%	14.58%
N6					
HIGH:	42	37	1	88.10%	2.38%
LOW:	42	22	10	52.38%	23.81%

Table 10a.
Number of States in the Reduced Models
All Models

Model	Total models	Non-transient	Mean per model	Stdev per model
U1				
HIGH:	22	94	4.27	0.686
LOW :	22	92	4.18	1.230
U3				
HIGH:	21	91	4.33	0.642
LOW :	21	83	3.95	0.898
U6				
HIGH:	16	71	4.44	0.788
LOW :	16	74	4.62	1.111
P1				
HIGH:	29	130	4.48	0.725
LOW :	29	122	4.21	1.349
P3				
HIGH:	54	249	4.61	0.591
LOW :	54	209	3.87	1.504
P6				
HIGH:	16	73	4.56	0.704
LOW :	16	73	4.56	0.998
N1				
HIGH:	16	88	5.50	1.061
LOW :	16	67	4.19	1.014
N3				
HIGH:	16	93	5.81	0.527
LOW :	16	68	4.25	0.750
N6				
HIGH:	16	91	5.69	0.583
LOW :	16	74	4.62	0.484

Table 10b.
Number of States in the Reduced Models
Best Models Only

Model	Total models	Non-transient	Mean per model	Stdev per model
U1				
HIGH:	10	41	4.10	0.700
LOW :	10	39	3.90	1.136
U3				
HIGH:	7	30	4.29	0.700
LOW :	7	32	4.57	0.728
U6				
HIGH:	2	9	4.50	0.500
LOW :	2	8	4.00	1.000
P1				
HIGH:	7	34	4.86	0.350
LOW :	7	29	4.14	1.125
P3				
HIGH:	11	53	4.82	0.386
LOW :	11	50	4.55	0.782
P6				
HIGH:	3	15	5.00	-0.000
LOW :	3	12	4.00	0.816
N1				
HIGH:	15	82	5.47	1.087
LOW :	15	64	4.27	0.998
N3				
HIGH:	8	46	5.75	0.661
LOW :	8	34	4.25	0.829
N6				
HIGH:	7	38	5.43	0.728
LOW :	7	32	4.57	0.495

Comparison of observation probabilities to the marginal distribution of the codes in the data

The significant t-tests for some of the early states of the P3 and N3 models are shown in the following tables and maps. In general, the observation probabilities in both the high and low matrices differ substantially from the marginal (background) probabilities in the dataset as a whole. This is not surprising for the high model, but somewhat surprising for the low model, since about 80% of the weeks in the data are in the low-category. The fact that the low model also shows a large number of differences is probably indicative of the fact that these models were selected to maximize the *difference* between high and low forecast behaviors, so even the low model picks up distinctive patterns.

Keep in mind that because the observation probabilities must add to 1.0, disproportionately high probabilities on some codes must be compensated by disproportionately low probabilities on some others; furthermore, those probabilities are summed across all of the dyads. Consequently while a large positive t-value always means that the behavior is more common than would be expected by chance, a large negative value can either mean that the behavior is being ignored completely by the model, or simply that it does not receive emphasis proportional to that of the high-frequency events (this, for example, is probably what is going on in the series of negative values for Kosovo in the P3-LOW model, and Bosnia in the N3-LOW model). (The very large t-values—those in excess of 10.0, or even in excess of 100.0—are usually associated with codes that occur only rarely and hence have standard deviations and means close to zero; these are usually associated with Kosovo.)

Beyond this, several general points are evident from these tables. First, all of the models, whether high or low, P or N, disproportionately pick up on WEIS "comment" and "consult" codes towards Bosnia (and, in some cases, Croatia), and usually from Serbia. Second, there is a consistent pattern of the high models having a probability of non-events that is much lower than expected by chance, where the low models have a non-event probability that is higher than chance (though usually with t-scores about 2 to 3, whereas the non-event t-scores in the high models are in the negative tens or hundreds). This reinforces what is becoming one of the fundamental of event-data-based early warning: the *existence* of a report is at least as important as the *content* of the report. Third, and unsurprisingly, violent behavior by Serbia—demonstrations, reduced relations, expulsions—is important in the high models, but not the low models.

TABLE 11a.
HIGH CONFLICT PARAMETER T-TESTS AGAINST MARGINALS FOR MODEL P3

State 1

	BOS>	>BOS	CRO>	>CRO	SER>	>SER	KSV>	>KSV
YIELD	---	---	---	-4.238	---	+++	---	---
COMMNT	---	8.235	---	+++	***	12.012	-3.440	---
CONSLT	---	7.563	---	+++	2.009	2.731	---	+++
APPROV	---	-4.286	---	+++	---	-5.017	---	-3.614
PROMIS	---	---	---	---	---	---	---	---
GRANT	+++	***	+++	---	+++	---	---	---
REWARD	---	---	-2.993	---	---	---	---	-3.429
AGREE	-2.598	***	+++	-2.042	***	-3.112	-8.630	-5.273
REQEST	---	-2.792	---	-2.085	-2.733	-3.798	---	---
PROPOS	---	---	---	+++	---	---	---	---
REJECT	+++	---	---	+++	---	---	---	---
ACCUSE	-9.528	---	+++	-3.216	+++	-3.624	-540.692	-7.584
PROTST	---	---	---	+++	---	+++	---	-3.639
DENY	---	-5.058	+++	---	---	-2.934	***	---
DEMAND	---	-3.096	-2.291	---	---	---	---	---
WARN	-2.089	---	---	---	***	-4.514	---	---
THREAT	---	---	---	---	+++	---	---	---
DEMONS	---	---	+++	---	2.782	---	---	-2.246
REDUCE	+++	---	---	---	***	---	---	-2.246
EXPEL	---	2.311	---	---	***	---	---	---
SEIZE	-8.816	-2.981	-11.002	-2.422	---	-8.630	-6.423	---
FORCE	---	---	---	---	---	+++	---	-2.019
NONEVT	-41.565							

State 2

	BOS>	>BOS	CRO>	>CRO	SER>	>SER	KSV>	>KSV
YIELD	---	---	---	+++	---	---	---	---
COMMNT	---	7.904	---	---	---	10.426	-3.245	-2.345
CONSLT	---	8.000	-2.955	---	+++	---	-2.779	-5.296
APPROV	---	---	---	---	---	-3.263	---	-6.024
PROMIS	---	+++	+++	---	***	---	---	***
GRANT	---	---	---	---	+++	---	-5.419	---
REWARD	---	---	***	---	---	+++	---	-3.694
AGREE	---	---	---	---	---	---	-8.745	-4.759
REQEST	---	-4.392	***	***	-2.466	-5.508	---	---
PROPOS	---	---	---	---	***	+++	---	-3.079
REJECT	---	-2.099	---	-2.302	---	-2.125	---	-3.416
ACCUSE	-2.589	---	-3.751	-6.533	-4.144	-2.394	---	-5.694
PROTST	***	---	+++	---	---	---	---	-6.014
DENY	---	-2.279	---	---	---	-2.895	+++	---
DEMAND	+++	-4.858	+++	---	---	-2.903	---	---
WARN	---	---	---	---	2.157	---	---	---
THREAT	---	---	---	---	---	---	---	---
DEMONS	---	---	---	---	2.368	---	---	-4.265
REDUCE	---	+++	---	+++	+++	---	---	-4.265
EXPEL	---	+++	+++	-3.159	---	***	---	---
SEIZE	---	-3.868	-11.873	+++	+++	-5.171	---	-2.486
FORCE	---	---	***	---	+++	-3.516	***	-4.386
NONEVT	-41.449							

TABLE 11b.
LOW CONFLICT PARAMETER T-TESTS AGAINST MARGINALS FOR MODEL P3

State 1

	BOS>	>BOS	CRO>	>CRO	SER>	>SER	KSV>	>KSV
YIELD	-3.021	---	---	---	+++	---	-2.284	-2.110
COMMNT	-3.305	10.613	2.106	---	-3.004	9.739	***	-3.513
CONSLT	***	9.396	---	---	---	---	***	-3.819
APPROV	+++	---	---	---	---	---	---	-2.744
PROMIS	-2.344	-2.992	---	---	***	+++	-3.689	-3.730
GRANT	---	---	---	---	-4.142	+++	-3.875	-4.072
REWARD	---	-3.185	-2.208	2.410	---	-2.269	---	-2.746
AGREE	-2.804	-2.036	---	---	---	-2.036	-5.106	-7.254
REQEST	-2.395	---	+++	---	***	---	---	-3.731
PROPOS	-2.210	-4.489	***	---	-2.334	-4.089	-3.875	-3.569
REJECT	---	---	---	+++	---	***	-2.918	-2.986
ACCUSE	-2.065	+++	2.044	+++	***	---	---	---
PROTST	-2.785	---	2.339	---	+++	2.194	---	---
DENY	2.091	---	---	---	-2.785	---	---	-6.134
DEMAND	+++	-2.317	+++	***	---	+++	-3.011	---
WARN	---	---	-4.645	---	-2.154	-2.096	---	---
THREAT	-3.787	---	2.334	2.251	---	+++	***	-5.018
DEMONS	+++	---	+++	2.280	---	---	-4.318	-19.038
REDUCE	***	-3.826	---	-2.170	---	---	---	---
EXPEL	---	---	---	---	-2.245	-2.006	---	---
SEIZE	-2.293	---	---	2.501	-3.489	---	---	***
FORCE	-8.186	-7.382	---	+++	-3.163	-4.335	---	-2.699
NONEVT	2.902							

State 2

	BOS>	>BOS	CRO>	>CRO	SER>	>SER	KSV>	>KSV
YIELD	-2.686	---	---	---	---	---	---	---
COMMNT	-5.039	4.700	---	---	---	6.708	---	---
CONSLT	-3.229	5.660	---	***	---	-3.625	---	---
APPROV	-6.613	-4.744	---	-3.211	-3.862	-2.880	---	---
PROMIS	---	+++	-2.790	---	---	-4.696	---	---
GRANT	-8.145	+++	+++	+++	-7.440	-4.532	***	-3.407
REWARD	-3.781	---	-2.062	---	---	---	---	---
AGREE	-2.036	-3.552	-2.022	---	-2.725	-2.387	-6.433	-2.687
REQEST	***	---	---	---	---	---	-4.067	---
PROPOS	+++	---	---	---	---	---	+++	---
REJECT	-4.314	---	---	---	---	---	---	---
ACCUSE	-6.935	---	---	---	-2.818	---	---	---
PROTST	-2.355	---	---	+++	---	---	---	---
DENY	---	-4.904	---	***	-6.368	-2.382	---	---
DEMAND	-9.357	---	***	---	---	-2.876	---	---
WARN	-45.513	---	---	---	-3.054	---	+++	---
THREAT	-8.092	-5.045	---	---	-3.503	+++	---	---
DEMONS	-3.597	-3.133	---	+++	+++	-6.083	---	-5.202
REDUCE	---	---	---	---	+++	---	---	---
EXPEL	---	-2.126	---	---	-2.691	-2.326	-43.497	---
SEIZE	-11.142	-2.142	---	---	-3.741	---	---	---
FORCE	-2.038	-2.700	---	---	-2.911	-2.719	---	---
NONEVT	1.993							

TABLE 11c.
HIGH CONFLICT PARAMETER T-TESTS AGAINST MARGINALS FOR MODEL N3

State 1

	BOS>	>BOS	CRO>	>CRO	SER>	>SER	KSV>	>KSV
YIELD	---	---	---	-2.471	---	---	---	---
COMMNT	---	7.624	---	---	***	9.578	***	-33.386
CONSLT	---	4.392	---	---	---	---	-10.062	-20.709
APPROV	---	---	---	---	---	-5.779	---	---
PROMIS	-5.755	---	---	+++	-2.907	---	---	-81.591
GRANT	---	---	---	---	2.504	---	---	+++
REWARD	+++	---	---	-3.524	---	+++	---	-98.116
AGREE	---	---	---	-2.325	---	-2.612	-41.931	-20.400
REQEST	---	---	---	---	-5.145	---	---	---
PROPOS	---	---	---	---	***	---	---	-68.371
REJECT	---	---	---	---	---	---	---	---
ACCUSE	---	---	-6.945	---	-2.089	---	+++	---
PROTST	---	***	---	---	---	---	---	---
DENY	---	-7.021	-2.367	-32.017	-2.243	-5.772	---	---
DEMAND	---	---	+++	---	-2.795	---	---	---
WARN	-3.000	-6.464	---	+++	---	+++	---	---
THREAT	---	---	2.861	***	---	---	---	---
DEMONS	---	-2.641	***	---	+++	---	---	-147.690
REDUCE	---	---	---	---	+++	---	---	---
EXPEL	---	---	+++	---	---	+++	---	---
SEIZE	---	---	-6.470	-2.171	---	-3.422	-3.088	-6.566
FORCE	---	---	***	---	---	-4.190	-14.170	-5.912
NONEVT	-19.223							

State 2

	BOS>	>BOS	CRO>	>CRO	SER>	>SER	KSV>	>KSV
YIELD	---	---	-2.061	-2.775	---	---	---	---
COMMNT	---	6.998	---	---	+++	7.518	-81.591	-18.838
CONSLT	---	6.145	---	---	+++	***	-10.333	-10.953
APPROV	---	-2.859	-2.335	---	---	-2.243	---	---
PROMIS	-2.125	---	---	***	---	---	---	---
GRANT	---	---	---	+++	---	---	---	---
REWARD	+++	***	---	---	---	+++	---	-3.828
AGREE	+++	---	---	-2.048	---	-3.019	-14.131	***
REQEST	---	---	+++	---	-2.579	-3.114	---	-3.474
PROPOS	---	---	---	---	---	---	---	-4.552
REJECT	---	---	+++	---	+++	---	---	-3.349
ACCUSE	+++	---	-2.619	+++	-3.834	-2.413	-81.591	---
PROTST	---	---	---	---	2.169	---	---	---
DENY	+++	-3.073	---	---	---	---	---	---
DEMAND	---	---	---	---	-2.445	---	---	---
WARN	---	---	---	+++	---	-2.001	---	---
THREAT	---	---	+++	***	+++	+++	---	---
DEMONS	-2.412	---	---	***	---	-2.100	---	-36.148
REDUCE	+++	---	---	---	2.635	---	---	---
EXPEL	---	+++	---	+++	---	---	---	---
SEIZE	---	-2.733	-8.688	+++	---	-10.030	-2.066	-4.113
FORCE	---	---	-2.062	---	---	-2.662	-6.771	-10.227
NONEVT	-8.849							

TABLE 11d.
LOW CONFLICT PARAMETER T-TESTS AGAINST MARGINALS FOR MODEL N3

State 1

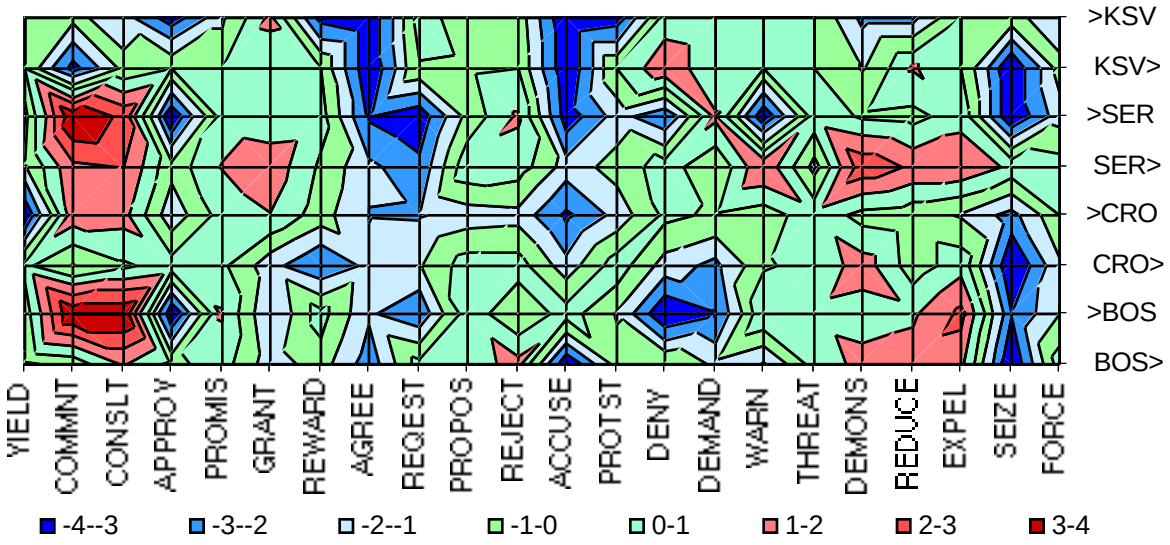
	BOS>	>BOS	CRO>	>CRO	SER>	>SER	KSV>	>KSV	
YIELD	-2.319	***	-2.059	---	---	***	---	---	
COMMNT	---	3.921	---	***	---	6.832	---	---	
CONSLT	-2.326	4.006	---	---	---	---	---	***	
APPROV	-2.238	---	+++	-4.024	+++	---	---	---	
PROMIS	---	+++	---	---	---	-2.332	---	---	
GRANT	---	-6.053	---	---	-9.946	---	---	---	
REWARD	-11.594	-8.933	---	+++	-3.205	---	---	---	
AGREE	-4.863	-5.890	---	---	***	---	-52.673	---	
REQEST	-2.734	+++	---	---	---	---	---	---	
PROPOS	---	-4.190	---	---	-2.441	---	---	-3.804	
REJECT	---	---	---	---	2.043	***	---	---	
ACCUSE	-7.449	---	---	---	---	---	---	---	
PROTST	-2.150	***	---	---	---	+++	---	+++	
DENY	-199.153		-3.634	---	---	***	---	---	---
DEMAND	-7.957	***	---	---	---	---	---	---	
WARN	---	-4.334	***	---	---	-3.851	---	---	
THREAT	-4.137	---	+++	---	---	---	---	---	
DEMONS	-2.660	-3.320	-2.246	---	---	---	---	---	
REDUCE	---	-5.902	---	---	+++	---	---	***	
EXPEL	---	---	---	---	---	---	---	---	
SEIZE	+++	---	-2.795	---	---	***	---	---	
FORCE	-3.540	-3.230	***	---	---	---	---	---	
NONEVT	1.887								

State 2

	BOS>	>BOS	CRO>	>CRO	SER>	>SER	KSV>	>KSV	
YIELD	-3.791	-4.067	+++	---	---	---	---	---	
COMMNT	-4.534	3.671	---	---	-4.211	3.509	---	-17.856	
CONSLT	-3.198	4.490	+++	---	---	---	-9.154	-8.159	
APPROV	---	---	---	---	---	---	---	---	
PROMIS	---	+++	---	---	---	---	---	-4.416	
GRANT	-4.189	-3.838	---	2.529	-4.810	---	---	---	
REWARD	-7.994	-2.711	---	---	+++	-4.244	---	---	
AGREE	---	---	***	***	***	2.180	---	-4.656	
REQEST	-3.705	-3.318	---	---	+++	---	-131.165		-14.831
PROPOS	***	-2.498	---	---	-7.115	-3.055	---	-2.085	
REJECT	---	-2.622	---	---	-3.241	-3.251	---	-131.165	
ACCUSE	---	-2.413	***	***	---	---	-4.718	---	
PROTST	-2.601	---	---	---	-5.090	---	---	---	
DENY	---	---	---	2.079	---	+++	---	---	
DEMAND	-7.780	-3.605	---	---	---	---	-40.279	---	
WARN	---	+++	---	---	-2.512	---	---	---	
THREAT	+++	---	---	2.043	+++	-2.836	---	-73.328	
DEMONS	---	-5.320	---	---	---	---	---	***	
REDUCE	---	-4.059	---	---	---	---	---	---	
EXPEL	---	---	***	---	-2.226	-32.017	---	---	
SEIZE	+++	---	---	---	---	---	---	---	
FORCE	-9.893	-6.976	---	---	+++	***	---	-4.468	
NONEVT	0.908								

Figure 13a
DIFFERENCE-OF-MEANS TESTS BETWEEN ESTIMATED AND MARGINAL PROBABILITIES, 3-MONTH HIGH MODELS STATE 1

P3 - HIGH - STATE 1



N3 - HIGH - STATE 1

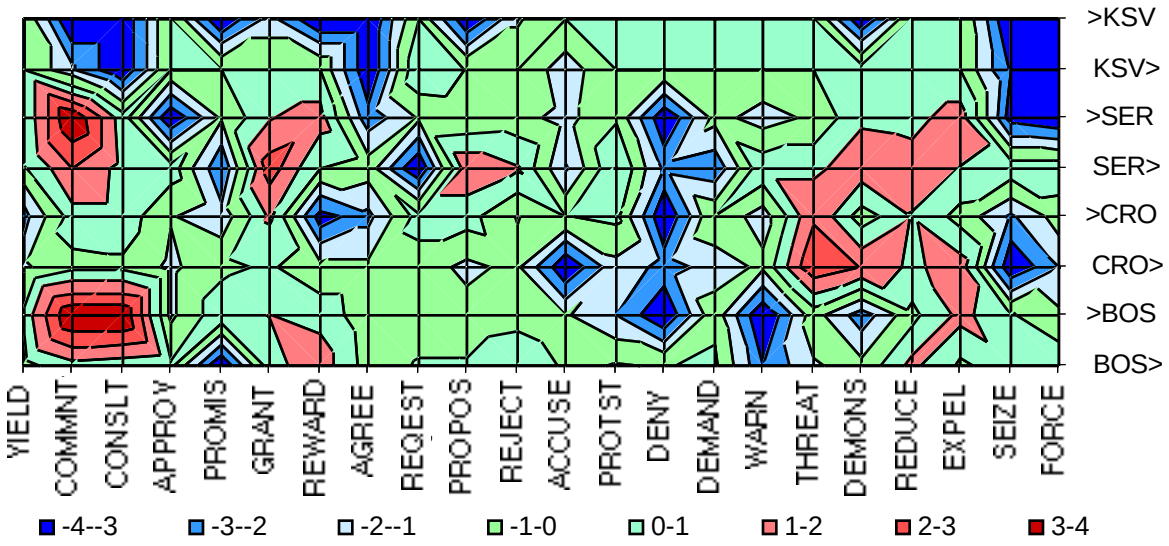
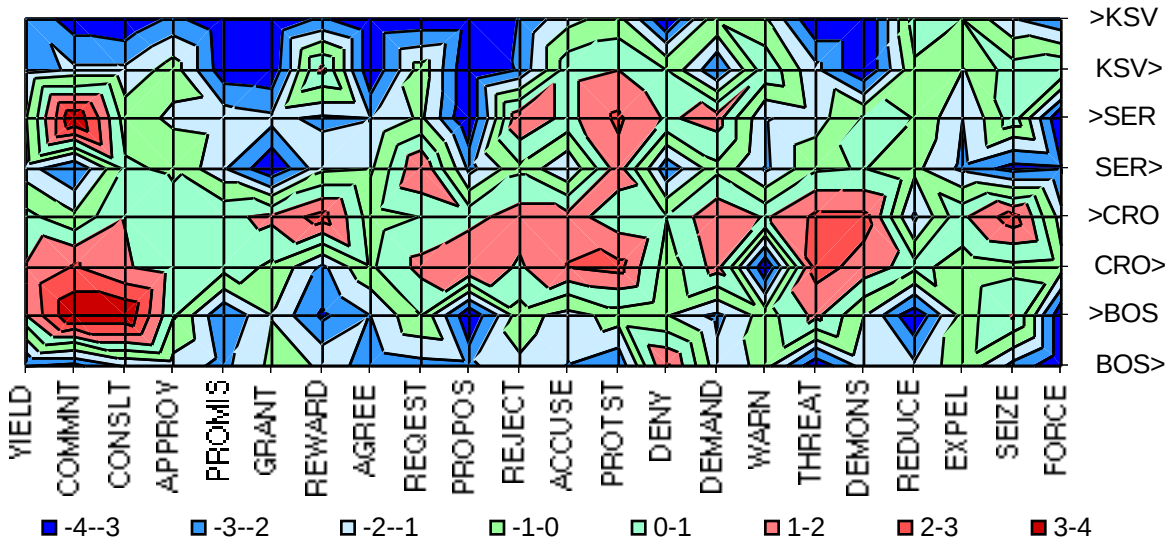
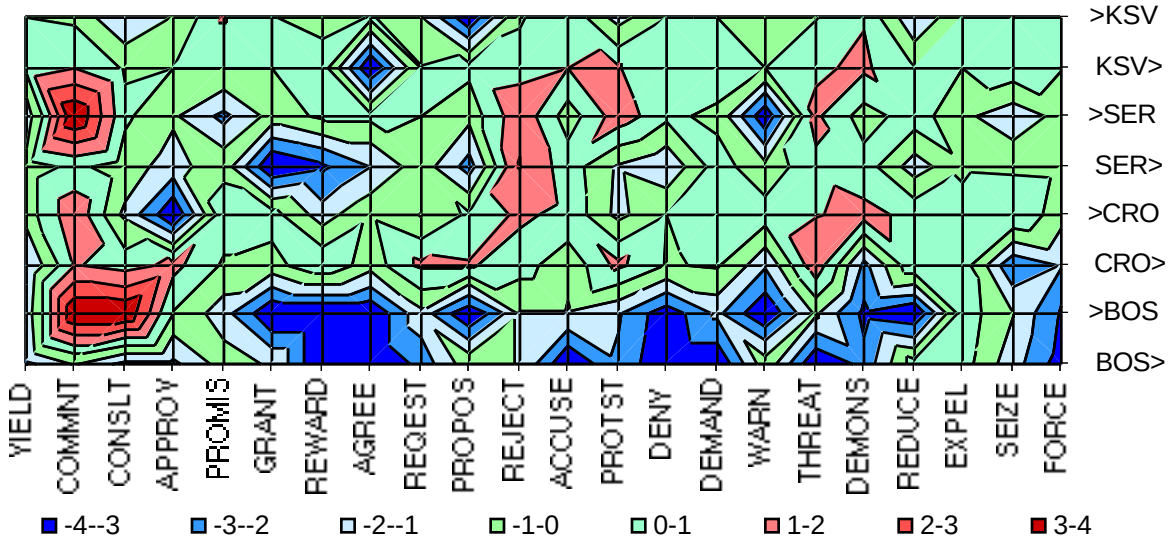


Figure 13b
DIFFERENCE-OF-MEANS TESTS BETWEEN ESTIMATED AND MARGINAL PROBABILITIES, 3-MONTH LOW MODELS STATE 1

P3 - LOW - STATE 1



N3 - LOW - STATE 1



Differences between the high and low models

T–tests were run for the differences between the high-conflict and low-conflict forecasting models. The differences between the observation probabilities in the high and low models generally show more consistent patterns than the parameter values themselves. The tables and maps show the difference of means of the high model minus the low model, so positive values indicate that the high model is higher.

The most obvious feature of both the tables and the maps is that almost all of the values are positive. This is an effect of the much lower number of non-events in the periods preceding a high-conflict period. In the N3 model, most of the differences involve Bosnia and Kosovo in the state 1 model, and in the state 4 model, behavior towards Kosovo—cooperative and conflictual—dominates. These strong results on Kosovo are undoubtedly due in part to the small standard deviations, but emphasis on cooperative events may also reflect attention by the international community towards Kosovo when tensions in the region heat up. Serbian reductions of relations and expulsions are picked up disproportionately in the state 1, but not the state 4, models, and curiously so are grants and rewards. In state 4, the high model looks at agreements across all of the dyads—this probably reflects international attempts to mediate. Interestingly, the maps for all models and the best models are almost identical.

The P3 model is quite different (and, in contrast to the N3 model, the all and best models are different). This model puts far more emphasis on violent behavior that is coming from Serbia, and Kosovo is generally ignored in the state 1 model. The state 4 matrix, in contrast, tends to look similar to the patterns found in the state 4 matrix of the N3 model, albeit with substantially more emphasis on Bosnian behavior.

The tables also report t-tests on the differences between the Markov probabilities in the models. Here one finds a consistent pattern: in the P-type models the recurrence probabilities are consistently (and often dramatically) higher in the high model than in the low model; the opposite is just as consistently true for the N-type models. In the case of the N-type models, this may simply reflect the tendency of the low model to have fewer states than the high model, but that also holds for P-type models, albeit in general the P-type models have fewer states than N-type models. The statistics in the tables as a whole (including those not presented here) provides some evidence, in the form of much lower prior transition probabilities, that the P-type high model is generally moving forward, state by state, whereas the low model simply fluctuates forward and backwards between several different states. This would be consistent with the high model showing escalatory behavior while the low model handles a background of fluctuating behavior. In the N-type model, this pattern is reversed: the high model fluctuates, whereas the low model shows de-escalation.

TABLE 12a.
HIGH .VS. LOW T-TESTS FOR MODEL P3

State 0

Prev_Trans	NaN		Same_Trans		3.224		Next_Trans	-3.224
	BOS>	>BOS	CRO>	>CRO	SER>	>SER	KSV>	>KSV
YIELD	+++	---	---	---	---	---	-2.562	-3.526
COMMNT	3.122	---	+++	***	3.638	---	+++	***
CONSLT	***	---	---	---	2.246	2.350	-3.653	-2.638
APPROV	---	-2.313	---	+++	---	-4.218	-3.161	-2.972
PROMIS	---	---	***	+++	---	-2.335	+++	+++
GRANT	-2.899	-2.898	---	+++	3.576	---	***	---
REWARD	---	---	---	+++	---	---	-3.552	***
AGREE	---	+++	-4.179	-4.162	***	***	-2.406	-3.304
REQEST	---	---	-2.521	-2.360	-2.498	-3.361	---	---
PROPOS	-2.066	---	+++	+++	***	3.143	---	-2.547
REJECT	+++	+++	---	+++	3.196	---	-3.148	-2.087
ACCUSE	---	---	---	2.043	3.196	---	---	---
PROTST	---	-2.677	---	---	***	---	-4.285	-3.109
DENY	***	-2.002	-2.421	---	2.403	-5.277	---	+++
DEMAND	+++	---	***	---	---	---	-2.931	-3.675
WARN	---	---	---	+++	2.294	+++	+++	-2.305
THREAT	---	---	-2.040	---	---	***	-2.981	-3.539
DEMONS	---	---	---	-2.462	---	---	---	---
REDUCE	+++	---	---	---	---	---	---	-2.774
EXPEL	2.063	---	---	+++	2.332	---	---	---
SEIZE	---	---	+++	***	2.308	---	---	---
FORCE	3.293	+++	---	---	3.176	---	-3.339	-3.475
NONEVT	-4.815							

State 1

Prev_Trans	-10.373		Same_Trans		9.481		Next_Trans	-1.410
	BOS>	>BOS	CRO>	>CRO	SER>	>SER	KSV>	>KSV
YIELD	---	---	---	-2.979	---	---	-2.515	-3.784
COMMNT	+++	+++	-2.064	+++	2.952	---	---	---
CONSLT	***	---	---	---	+++	2.532	---	---
APPROV	---	-2.671	---	+++	---	-2.594	-2.539	---
PROMIS	---	***	---	---	+++	---	+++	---
GRANT	---	---	---	---	2.865	---	---	2.149
REWARD	---	***	---	-2.028	---	+++	-2.807	---
AGREE	---	---	---	---	---	---	---	---
REQEST	---	-2.028	+++	+++	-3.284	-2.494	---	---
PROPOS	+++	2.316	---	+++	+++	---	---	---
REJECT	+++	---	---	***	---	---	-2.030	---
ACCUSE	-3.522	---	-2.545	-2.938	---	-3.033	-3.078	-3.003
PROTST	---	---	+++	---	+++	-2.668	-3.676	+++
DENY	+++	-2.138	---	---	+++	-2.404	+++	+++
DEMAND	---	---	-2.322	+++	---	---	-2.220	-3.282
WARN	---	---	---	---	2.630	+++	***	-2.339
THREAT	***	---	***	---	---	---	-3.106	-2.690
DEMONS	***	---	---	-2.154	2.365	---	---	+++
REDUCE	2.066	***	---	---	***	---	---	+++
EXPEL	---	2.188	---	---	2.150	+++	---	---
SEIZE	-2.717	-2.639	-4.498	-3.393	2.065	-4.378	-2.651	---
FORCE	4.043	2.737	---	---	2.122	---	---	---
NONEVT	-4.185							

TABLE 12b.
HIGH .VS. LOW T-TESTS FOR MODEL N3

State 0

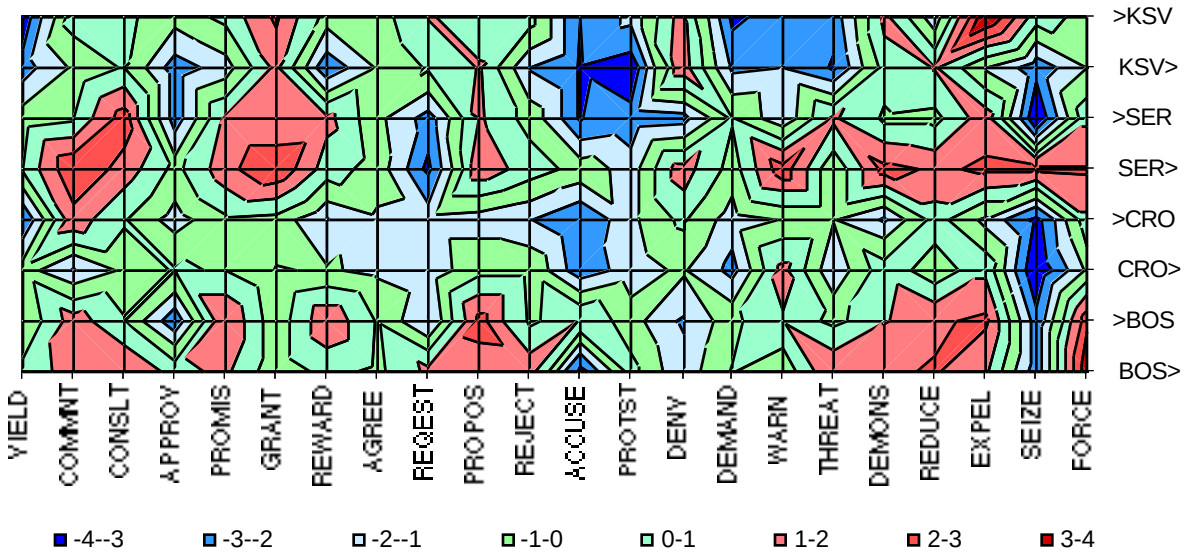
Prev_Trans	NaN		Same_Trans		-3.352		Next_Trans	3.352
	BOS>	>BOS	CRO>	>CRO	SER>	>SER	KSV>	>KSV
YIELD	---	+++	---	---	2.253	---	***	---
COMMNT	---	---	---	---	---	-2.622	---	---
CONSLT	---	---	---	---	+++	---	***	-2.016
APPROV	---	---	---	---	---	-2.594	---	-2.175
PROMIS	---	---	---	---	---	+++	---	---
GRANT	---	---	2.027	***	***	---	---	---
REWARD	---	---	---	---	-2.051	---	---	-2.064
AGREE	---	+++	---	---	---	---	---	---
REQEST	***	---	---	+++	+++	***	---	-2.723
PROPOS	---	-2.259	+++	---	+++	+++	---	+++
REJECT	---	---	---	---	---	---	+++	-2.051
ACCUSE	---	***	---	---	2.193	---	+++	+++
PROTST	---	+++	---	+++	+++	---	---	---
DENY	***	---	---	---	---	-3.119	---	---
DEMAND	---	---	---	---	---	---	***	***
WARN	---	---	---	---	---	---	---	---
THREAT	---	---	---	---	---	---	---	---
DEMONS	+++	-2.259	---	***	---	---	---	---
REDUCE	---	***	---	---	---	---	***	+++
EXPEL	---	***	---	***	3.409	---	---	---
SEIZE	---	---	***	---	---	---	---	---
FORCE	2.239	---	---	---	---	---	+++	-3.112
NONEVT	-2.402							

State 1

Prev_Trans	3.853		Same_Trans		-5.535		Next_Trans	4.039
	BOS>	>BOS	CRO>	>CRO	SER>	>SER	KSV>	>KSV
YIELD	---	+++	---	---	---	---	+++	---
COMMNT	---	---	---	---	---	---	+++	+++
CONSLT	+++	---	---	---	---	---	+++	+++
APPROV	+++	---	+++	---	---	+++	---	+++
PROMIS	+++	+++	---	---	---	---	---	+++
GRANT	---	2.079	---	---	4.514	---	---	---
REWARD	3.516	3.680	---	---	***	---	+++	+++
AGREE	+++	***	---	---	---	---	---	+++
REQEST	+++	---	---	---	+++	---	---	+++
PROPOS	---	2.575	+++	---	2.623	---	---	+++
REJECT	---	---	---	+++	---	***	---	---
ACCUSE	2.450	---	-2.098	---	---	---	---	+++
PROTST	---	---	+++	---	+++	---	***	+++
DENY	+++	---	---	---	---	***	---	---
DEMAND	2.157	---	---	---	---	---	---	+++
WARN	---	---	---	---	---	---	---	---
THREAT	+++	---	---	---	---	+++	---	---
DEMONS	---	---	2.420	+++	---	---	---	+++
REDUCE	---	2.646	---	---	***	---	---	---
EXPEL	---	---	---	---	---	+++	---	---
SEIZE	---	---	+++	+++	---	---	+++	+++
FORCE	2.441	+++	---	---	---	---	-2.487	+++
NONEVT	-2.083							

Figure 14a
DIFFERENCE-OF-MEANS TESTS BETWEEN HIGH AND LOW MODELS, P3 MODEL STATE 1

P3 - STATE 1 - ALL



P3 - STATE 1 - BEST

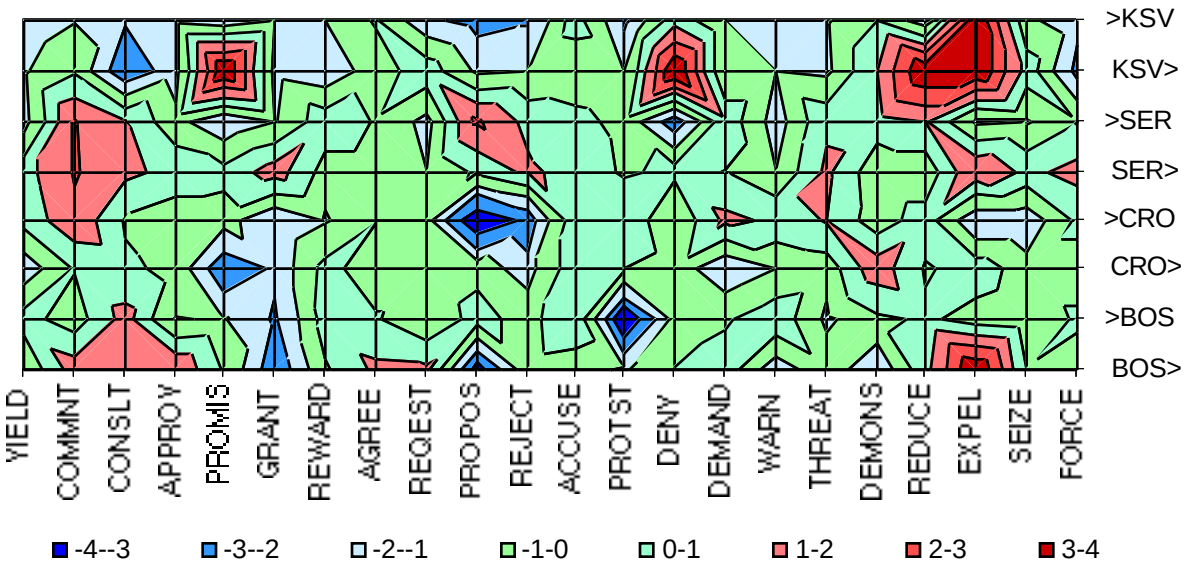
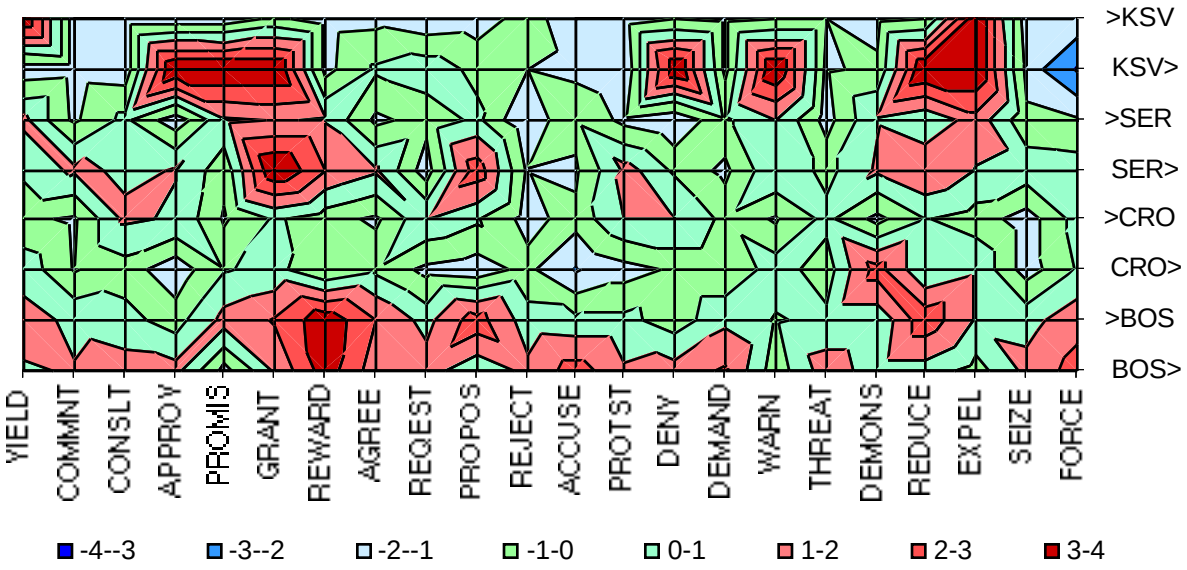


Figure 14b.
DIFFERENCE-OF-MEANS TESTS BETWEEN HIGH AND LOW MODELS, N3 MODEL STATE 1

N3 - STATE 1 - ALL



N3 - STATE 1 - BEST

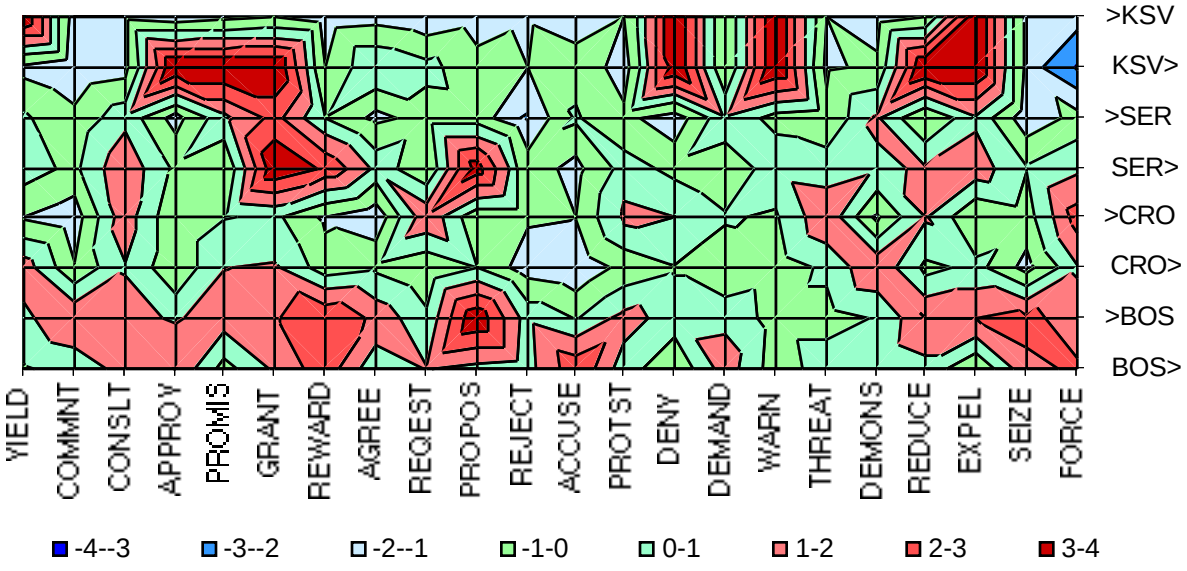
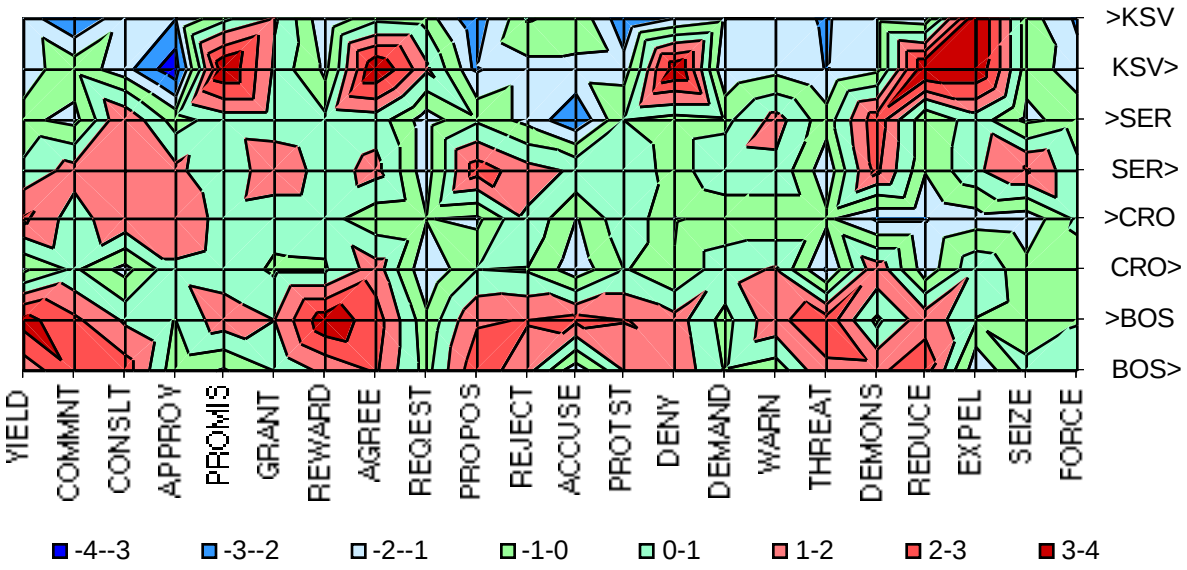


Figure 14c.
DIFFERENCE-OF-MEANS TESTS BETWEEN HIGH AND LOW MODELS, P3 MODEL STATE 4

P3 - STATE 4 - ALL



P3 - STATE 4 - BEST

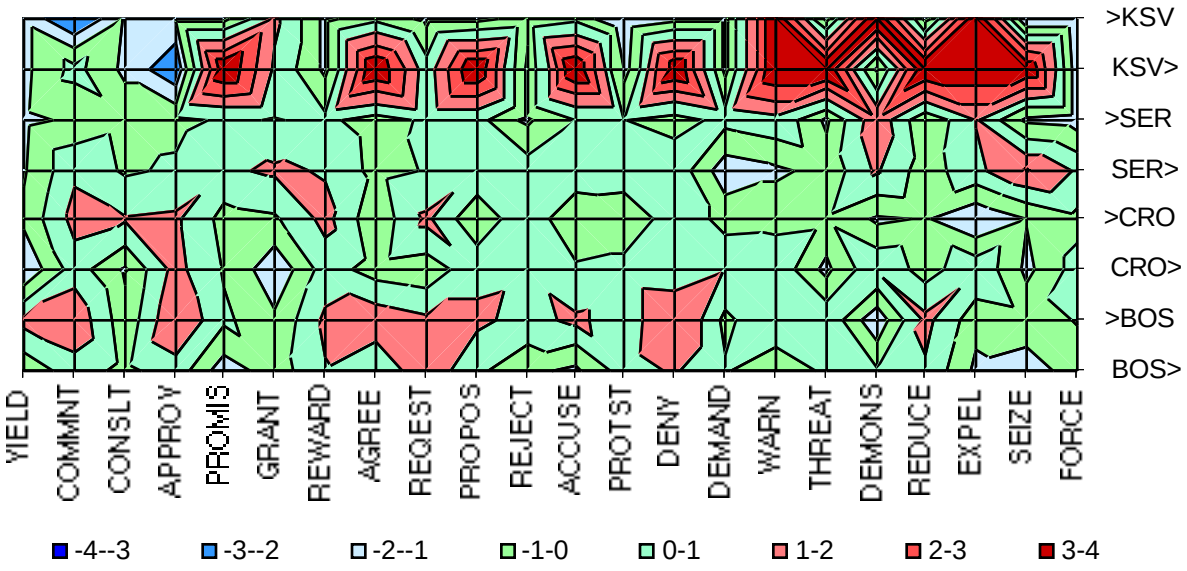
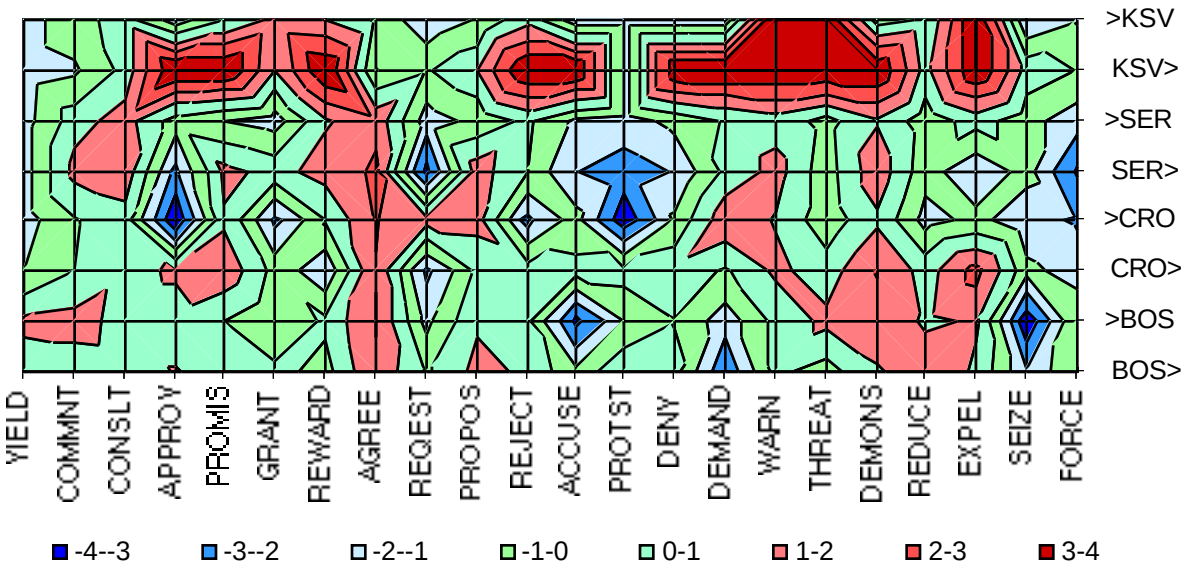
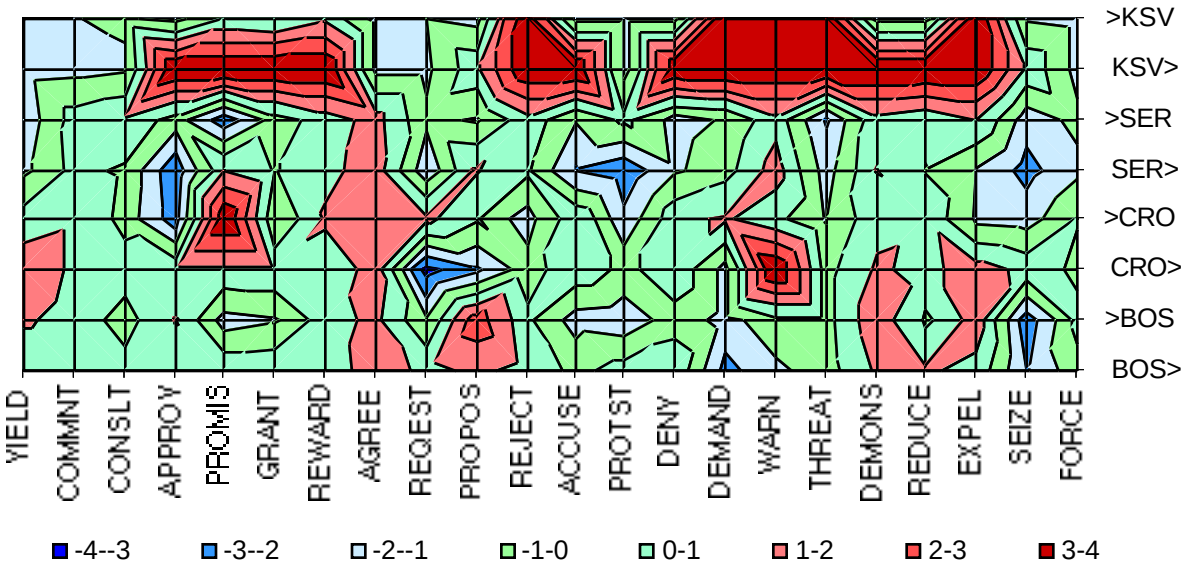


Figure 14d. DIFFERENCE-OF-MEANS TESTS BETWEEN HIGH AND LOW MODELS, N3 MODEL STATE 4

N3 - STATE 4 - ALL



N3 - STATE 4 - BEST



Comparison of 1-month and 6-month forecast models

T–tests were run for the difference between the 1-month and 6-month forecasting models. There appear to be few if any systematic differences—in fact in comparisons of the later states, the number of significant t-values barely rises above the level expected by chance. This result is particularly strong in the low model estimates in .best models, which would suggest that the most effective low models are simply modeling the background activity in non-conflict periods, with appropriate adjustment to maximize the distinctions between this and the high-conflict periods.

In the high models, there is a consistent pattern is that there are more differences in the early states of the models. This would suggest that more of the discrimination may be found in the early states, which is also consistent with how the HMM models themselves work: The best fit for a sequence might involve simply cycling in the early states of the model, which should result in more precise estimation of those states. This is also consistent with the interpretation that these forecasts are dealing with general shifts in behavior (which involve relatively diffuse changes in the event sequences) rather than a more finely-differentiated development of a six-stage crisis that would involve a larger number of distinct Markov states.

Three Experiments in Simplifying the Model

As the discussion above makes clear, one major drawback of the HMM approach is the large number of parameters. Because these parameters do not appear to exhibit any obvious structure—widely different values produce roughly comparable results—it seems likely that many of them are redundant. I therefore undertook two experiments to test the effects on accuracy of substantially reducing the number of parameters, first by looking only at the activity of the dominant actor in the system—Serbia—and then by reducing the detail in the event codes. I also looked at the effects of changing the classification variable from predicting conflict in a single week to predicting conflict in a 1- or 2-month period.

Reducing the Number of Dyads: Serbia Only

Serbia is clearly the dominant actor in this crisis. During the 1990s, conflict in the former Yugoslavia shifted focus from Croatia to Bosnia to Kosovo, but Serbia (and ethnic Serbs) were involved at all of these stages. This suggested an alternative approach: just monitor the activity of Serbia. Consequently I re-estimated a simpler model that involves only 45 codes:

[any source] -> Serbia Serbia -> [any target]

plus the 00 nonevent code. The remainder of the forecasting design was the same as before.

The focus on Serbia was also suggested by the fact that the model did not do very well at all on predicting the outbreak of violence in Kosovo. This could be due to the fact that while the original model included events involving Kosovo, there was in fact no real precedent for violence in Kosovo, and the reason that the model did not predict problems in Kosovo might have been due to it "looking" for the wrong thing.

Tables 3 and 4 show the summary statistics on the accuracy of the simplified model. In terms of overall accuracy, the results for the U and P models are almost identical to the results in the 4-dyad model, and the N model actually improves by about 10%! (The improvement in the N model is due largely to increased accuracy during the implementation of the Dayton accords). These results are also mirrored in the analysis of the high and low-week errors in Table 4: The U and P models gain between 5% and 20% observed accuracy in the high conflict week, while losing only about 2% observed accuracy in low-conflict weeks. The forecast accuracy of the U and P models remains about the same. In the N model, the observed accuracy for high-conflict weeks increases slightly (about 5% for the 3 and 6 month forecasts), and increases about 10% for low-conflict weeks. Forecast accuracy increases about 7% for high-conflict weeks and stays the same for low-conflict weeks. In almost none of these cases is there a serious loss of accuracy when moving to the simpler model.[18]

Figure 7 shows the 5-week centered moving average of the 3-N and 3-P Serbia-only models; in general these show patterns (including mirror-imaging) that are similar to those found in the 4-dyad models. However, the direct comparison between the Serbia-only and 4-dyad 3-N models—shown in Figure 8—reveals several interesting differences. First, as I had hoped, the Serbia-only model is much quicker at picking up the cessation of hostilities following the Dayton Accords—the two lines parallel each other during the period from July-1991 to October-1997, with the Serbia-only model being about 30% more accurate. Second, there are two deep spikes of low accuracy in the Serbia-only model that are not found in the 4-dyad model: these occur in late 1991 and late 1998. In both cases, the inaccuracy is due to false-positives: the model is saying there will be conflict, but it is not reflected in the data. Both of these periods are followed by major outbreaks of violence.

This analysis suggests that this simple Serbia-only model may be more accurate than the more complex 4-dyad model, despite the fact that the earlier model contains much more information. There are at least three reasons that this might be true. First, simple models of social behavior are often more accurate, because the measurement errors in an more complex model add more signal

---

[18] The Serbia-only model also shows a greater likelihood of consistently being less accurate at longer horizons, which would be consistent with that model having reduced amounts of noise.

than noise.[19] (This parallels the experience of the State Failures Project, which started by looking at several hundred variables, and found that all they really needed were half a dozen.) Second, "watch the bad guys" is intuitively plausible—Serbia has been the initiator of much of the violence in this region, and when not the initiator, usually the target, so monitoring Serbia alone is going to get most of the required information. Such an approach might not work in an area characterized by a truly multi-actor conflict such as central Africa, or the Afghan or Lebanese civil wars. Finally, the underlying theory of sequence matching suggests that models should not be overly specific—the whole point of the exercise is to generalize. In this case, the generalization is across all Serbian behavior, irrespective of target.

Table 3
Summary of Estimated Models: Serbia Only

	Best Models			All Models		
Model	**# Models**	**# Obsrv**	**% Correct**	**# Models**	**# Obsrv**	**%Correct**
1-U	8	3256	80.1%	32	13024	76.9%
3-U	7	2786	81.5%	32	12736	76.3%
6-U	2	770	79.9%	32	12320	74.4%
1-P	7	2849	81.1%	32	13024	76.8%
3-P	6	2388	81.2%	32	12736	76.3%
6-P	2	770	79.9%	32	12320	75.0%
1-N	31	12617	64.1%	32	13024	63.6%
3-N	30	11940	59.7%	32	12736	58.9%
6-N	32	12320	60.2%	32	12320	60.2%

[19] In the case of event data, coverage bias is probably the most important source of error: reporters go to where the action is, and when the action is in Sarajevo, few reports will come out of Kosovo.

Table 4a
High Conflict Weeks: Serbia Only

	Best Models		All Models	
Model	Observed	Forecast	Observed	Forecast
1-U	38.7%	51.5%	43.1%	43.3%
3-U	40.3%	58.1%	49.7%	43.9%
6-U	35.5%	55.1%	43.5%	41.1%
1-P	37.9%	55.3%	44.5%	43.3%
3-P	38.8%	57.4%	46.9%	43.7%
6-P	33.1%	55.6%	42.8%	42.1%
1-N	92.8%	35.5%	93.0%	35.2%
3-N	91.3%	33.1%	91.7%	32.7%
6-N	90.6%	34.1%	90.6%	34.1%

Table 4b
Low Conflict Weeks: Serbia Only

	Best Models		All Models	
Model	Observed	Forecast	Observed	Forecast
1-U	90.6%	85.3%	85.5%	85.4%
3-U	92.3%	85.4%	83.3%	86.3%
6-U	92.1%	83.9%	82.9%	84.2%
1-P	92.2%	85.3%	85.0%	85.7%
3-P	92.4%	85.1%	84.1%	85.7%
6-P	92.7%	83.5%	83.8%	84.2%
1-N	56.8%	96.9%	56.1%	96.9%
3-N	51.3%	95.7%	50.2%	95.8%
6-N	51.8%	95.3%	51.8%	95.3%

Figure 7.

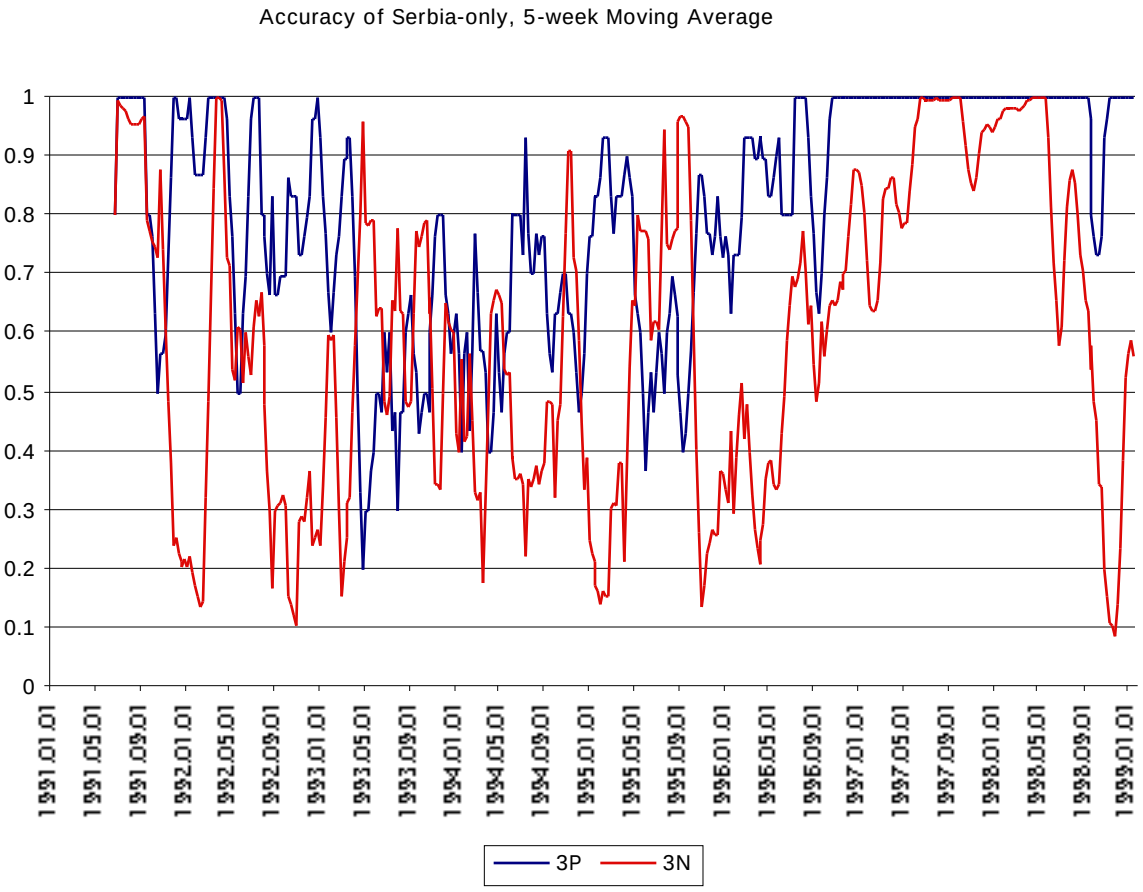
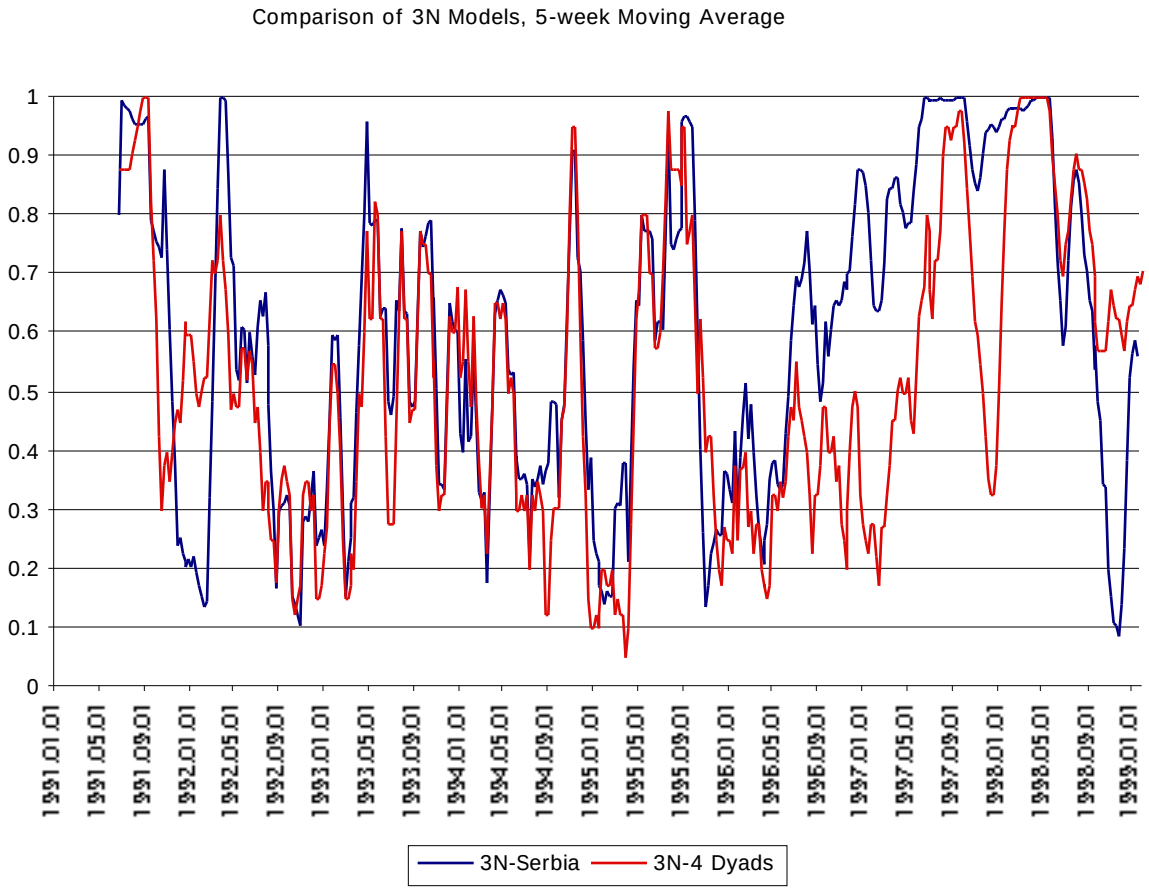

Accuracy of Serbia-only, 5-week Moving Average
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1991.01.01
1991.05.01
1991.09.01
1992.01.01
1992.05.01
1992.09.01
1993.01.01
1993.05.01
1993.09.01
1994.01.01
1994.05.01
1994.09.01
1995.01.01
1995.05.01
1995.09.01
1996.01.01
1996.05.01
1996.09.01
1997.01.01
1997.05.01
1997.09.01
1998.01.01
1998.05.01
1998.09.01
1999.01.01
3P
3N

Figure 8.



Modifying the and Coding Scheme

A second experiment involves simplifying the coding system. Earlier work on conflict in the Middle East (Gerner and Schrodt 1998) showed that in a cluster analysis, it was possible to substantially reduce the number of coding categories substantially without much loss in predictive power (in fact predictive power might even be gained by eliminating sources of noise). I therefore re-estimated the models using the following five-category system:

0. Non-event
1. Verbal cooperation (WEIS categories 02, 03, 04, 05, 08, 09,10)
2. Material cooperation (WEIS categories 01, 06, 07)
3. Verbal conflict (WEIS categories 11, 12, 13, 14, 15, 16, 17)
4. Material conflict (WEIS categories 18, 19, 20, 21, 22)

This reduces the total number of codes in the 8-dyad Balkans model from 177 to 33, which also substantially reduces the number of low frequency code categories. (This is also likely to reduce the effect of coding variance and coding error somewhat: Several of the "verbal conflict" codes in WEIS are ambiguous even for human coders, and the automated coding probably generates some misclassification in those categories.)

The results of this experiment are given in Table 13a. As before, the models were evaluated with 1, 3 and 6-month forecast lags, and with the P and N weighting scheme. 32 Monte-Carlo genetic algorithm estimates were done for each set of experimental parameters. For purposes of comparison, Table 13b presents the statistics for the original model in the same format.

In general, the results of the new analysis are comparable to those of the original analysis. As before, the drop-off in accuracy with the increasing forecasting lag is small—about 4% from the 1-month to 6-month forecast lag—but consistently there is a small decrease. The overall accuracy measure decreases about 4% for the P models and *increases* about 8% for the N models. The largest difference in the results occurs with respect to the accuracy of the high-conflict predictions in the P models—these average about 18% better in the percentage of the observed high weeks that were correctly forecast, albeit at the cost of an 8% decrease in the corresponding percentage of the observed low weeks that were correctly forecast. The N model shows an 11% increase in the percentage of the observed low weeks that were correctly forecast and a 5% increase in the percentage of forecasts of high conflict that actually had high conflict. All of the remaining statistics differ from the original model by less than 3%.

This analysis clearly supports the results found in Gerner and Schrodt (1998)—the use of simplified event coding systems at worst involves only a small penalty in terms of predictive accuracy, and at best can actually improve the accuracy, probably through the reduction of noise. This is particularly important when automated coding is being used, since automated coding is generally less capable of making subtle distinctions between event categories, but generally is quite good at making large distinctions such as the difference between cooperative and conflictual behavior.[20]

---

[20] It should be noted that both this test and the earlier Gerner and Schrodt (1998) test use machine-coded data, so this effect might be due to the errors found in that type of coding. However, this seems somewhat unlikely given the magnitude of the effect, the fact that the overall error rate in machine coding is comparable to that of human coding, and the fact that many of the categories that are ambigous in machine coding are also ambiguous to human coders.

Table 13a. Accuracy for 5-Category Coding System

Experiment	%accuracy	%high correct	%low correct	%high forecast	%low forecast
P1	74.4	46.2	81.5	38.9	85.6
P3	71.7	44.1	78.9	35.4	84.4
P6	71.4	44.2	78.8	36.4	83.8
N1	61.9	90.7	54.6	33.7	95.8
N3	57.8	87.0	50.2	31.4	93.6
N6	56.8	85.9	48.8	31.5	92.7

Table 13b. Accuracy for 1-Week Prediction Periods and 23-Category Coding System

Experiment	%accuracy	%high correct	%low correct	%high forecast	%low forecast
P1	77.6	29.3	89.5	40.8	83.7
P3	76.0	29.0	87.9	37.9	82.9
P6	76.9	25.9	90.6	42.6	82.0
N1	54.2	92.7	45.3	28.1	96.4
N3	49.0	88.1	39.6	25.9	93.3
N6	47.7	88.5	37.4	26.3	92.8

Modifying the Prediction Framework

The final experiment involves modifying the criteria for making a prediction. As noted at a number of points above, the objective of predicting the level of conflict in a single week is unrealistically precise—most predictions by human political analysts are made over more general periods of time. Consequently, the classification variable was modified to indicate whether a high-conflict week occurred during a 4- or 8-week period following the forecasting-lag period. The threshold for high-conflict was set at greater than 26 WEIS category **22** events for the 4-week

period and greater than 57 WEIS category **22** events for the 8-week period; approximately 30% of the periods fall into the high-conflict category.

These results are reported in Table 14, the 4 and 8 prefixes in the "Experiment" code refer to conflict occurring in 4- and 8-week periods respectively. These runs were done with the 5-code, rather than the earlier 22-code, system; and there were 32 Monte Carlo experiments for each experimental condition.

Surprisingly, the overall accuracy changes very little from the results of the 5-code experiment reported in Table 13a; almost all of the accuracy percentages are within 3% of the earlier results. However, there are substantial changes in the *pattern* of the accuracy. As with the change in the 5-code system, this change benefits the accuracy of the forecasts of the high-conflict months. In the P model, the number of high conflict periods correctly predicted increases by about 13% for the 1-month and 3-month forecast lags, and the percentage of high conflict periods where high conflict is actually observed also increases by about 13%. In the N model, the percentage of high conflict periods where high conflict is actually observed also increases by about 10%, but there is almost no change in the the number of high conflict periods correctly predicted. Because the overall accuracy changes little, these increases in accuracy for the high-conflict periods are compensated by decreases in accuracy for the low-conflict periods, though those percentage changes are smaller—typically around 4%—because the number of low-conflict periods is larger.

These results were somewhat unexpected: I had expected that predictions for the longer periods would be more accurate because they would not be subject to errors due to the random week-to-week variations within a period of high conflict. In a sense, this is what occurs in the higher prediction accuracy for high-conflict periods. However, that has relatively little impact on the overall accuracy, which is dominated by the low-conflict periods and low-conflict is highly autocorrelated. The results are consistent with the model is picking up very general characteristics of the behavior that change only slowly over time, rather than looking at indicators for specific weeks, despite that choice of the classification variable; this will be discussed in greater detail below.

Table 14. Accuracy for 4- and 8-Week Prediction Periods and 5-Category Coding System

Experiment	%accuracy	%high correct	%low correct	%high forecast	%low forecast
4P1	70.7	59.6	75.5	51.2	81.3
4P3	69.1	55.0	75.4	49.9	79.1
4P6	67.1	47.3	75.6	45.4	77.1
8P1	72.8	59.4	78.2	52.3	82.8
8P3	69.6	58.9	74.0	48.4	81.3
8P6	69.7	51.6	77.5	49.2	79.1
4N1	62.1	89.2	50.4	43.7	91.6
4N3	56.8	87.4	43.2	40.6	88.6
4N6	55.3	85.4	42.5	38.9	87.2
8N1	64.9	89.6	55.0	44.5	93.0
8N3	59.1	86.9	47.6	40.7	89.8
8N6	58.5	86.3	46.8	40.7	89.0

Conclusion

The overall conclusion of this analysis is that hidden Markov models are a robust, though hardly flawless, method for forecasting political conflict, at least when applied to area such as the former Yugoslavia where substantial information about political events is available. From the standpoint of pure prediction, the models are credible.

Unfortunately, these models are less useful from the standpoint of *inference*; in particular, it is very difficult to figure out what information the HMM is using to make these prediction. At various points in the discussion above, I have noted some clear—and generally plausible—patterns in the estimated probabilities. However, in general it has proven quite difficult to make much sense out of these. This is not to say that this exercise has been useless—without looking for patterns, there was no way to know they were not present—but a series of experiments trying to find patterns in these coefficients has not produced any obvious results.

There are a couple of likely reasons for this. First, these forecasting models have relatively long time horizons. The common-sense signs of short-term escalatory behavior such as demands, threats and small-scale incidents of violence will not necessarily be found at three and six month horizons or, in this data, even at one month. (One sees a little bit of this, for example in behaviors coming from Serbia, or towards Bosnia and Kosovo, but not a lot).

What the models seem to be picking up instead are relatively diffuse indicators, and often as not, simply increased attention to the area by the international media (as well as the international community. For example, one would expect that if a NATO commander or UN representative called attention to some issue, Reuters would almost always report it, whereas if the mayor of a village made the same comment, this might be ignored. Some of these indicators may be indirect: For example I initially viewed the emphasis in the high-conflict models on Kosovo to be a statistical artifact, but this could also be reflecting that fact that after 1991 or so, the international community consistently responded to aggressive moves by Serbia by warning Serbia not to do anything in Kosovo, and these activities are probably picked up in the data set. The presence or absence of non-events is clearly very important, a result that was expected—the fact that the international media are reporting on an area is by itself a useful indicator.

The interpretation of the coefficients is further complicated by the fact that we are dealing with probabilities—which by definition sum to 1.0—so an increase in one probability necessarily leads to a decrease in another. Add to this the propensity of interactions in event data to be symmetrical, and the fact that the codes are aggregated not only within the four actors in the Balkans, but also interactions between those actors and the international system as a whole, and the determination of the parameters gets very complex. In particular, it is quite different from the more familiar task of interpreting regression coefficients, which (in the absence of significant collinearity) are more or less independent of each other in value.

This is further complicated by the problem of the indeterminancy of the estimates produced by the Baum-Welch algorithm. This indeterminancy does not seem to be dealt with in detail in the HMM literature—most HMM applications are solely concerned with prediction, not inference—but where it is mentioned, the experience that I have had estimating these models appears to be typical. I have done a series of additional unreported experiments to attempt to find ways to reduce the variance of the estimates—increasing the number of templates used, varying the parameters of the genetic algorithm, changing the convergence conditions of the Baum-Welch algorithm, and setting the initial observation probabilities in the vicinity of the marginal probabilities of events in the data set as a whole—and none of these had a major impact. There is some limited evidence that the variance in the *accuracy* of the models estimated with the 5-code system is less than that of the 22-code system, but the variance in the parameters is still quite high.

Finally, the sheer number of parameter estimates generated by these HMMs complicates the problem of interpretation. The HMM models may be similar to neural networks in this regard: the diffuse coefficient structure is the model's way of dealing with the high degree of uncertainty in the underlying data, and the complexity of tradeoffs between the parameter values makes them almost impossible to interpret in a simple fashion. Doing a full interpretation would ideally involve five dimensions of comparison—WEIS category, dyad, Markov state, weighting scheme and high/low model—which is three dimensions more than most people can deal with. I've focused here on two weights, one or two Markov states, and a two-dimensional actor-by-code comparison, but that obviously leaves plenty of other possibilities that have not been explored.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Feller, William. 1968. *An Introduction to Probability Theory and Its Applications.* New York: Wiley.

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

Gerner, D. J. and P. A. Schrodt. 1998. "The Effects of Media Coverage on Crisis Assessment and Early Warning in the Middle East." In *Early Warning and Early Response*, ed. Susanne Schmeidl and Howard Adelman. New York: Columbia University Press-Columbia International Affairs Online.

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

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

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

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

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

Huxtable, P. A. 1997. "Uncertainty and Foreign Policy-making: Conflict and Cooperation in West Africa." Ph.D. dissertation, University of Kansas.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Schrodt, P. A. 1985. The Role of Stochastic Models in International Relations Research. In *Theories, Models and Simulation in International Relations*, ed. M. D. Ward. Boulder: Rienner

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

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

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

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

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

Schrodt, P. A. 2000. "Pattern Recognition of International Crises using Hidden Markov Models." In *Political Complexity: Nonlinear Models of Politics*, ed. Diana Richards. Pp. 296-328. Ann Arbor: University of Michigan Press.

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

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

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

Schrodt, P. A., and D. J. Gerner. 1997. An Event Data Set for the Arabian/Persian Gulf Region 1979-1997. Paper presented at the International Studies Association, Minneapolis, March 1998.(this paper be downloaded from http://www.ukans.edu/~keds/papers.html)

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

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

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

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