Real Time, Time Series Forecasting of Inter- and Intra-State Political Conflict\textsuperscript{1}

Patrick T. Brandt  
School of Economic, Political and Policy Sciences  
University of Texas, Dallas  
pbrandt@utdallas.edu

John R. Freeman  
Department of Political Science  
University of Minnesota  
freeman@umn.edu

Philip A. Schrodt  
Department of Political Science  
The Pennsylvania State University  
schrodt@psu.edu

\textsuperscript{1}An earlier version of this paper was presented at the 50th Annual Meeting of the International Studies Association, New York. Since this paper contains \textit{ex ante} forecasts, the authors want to note that this version of the paper was written on April 17, 2010 and only uses data through the end of 2009. Part of this research is based upon work supported by the National Science Foundation under Award Nos. 0921018, 0921051, and 1004414. We would also like to thank Christian Cantir and Andrew Lucius for their research assistance. Finally, we are grateful for the helpful comments of the reviewers.
Abstract

We propose a framework for forecasting and analyzing regional and international conflicts. It generates forecasts that 1) are accurate but accounts for uncertainty, 2) are produced in (near) real time, 3) address the simultaneous actors’ behaviors, 4) incorporate prior beliefs, and 5) generate policy contingent forecasts. We combine the CAMEO event-coding framework with Markov-switching and Bayesian vector autoregression models to meet these goals. Our example produces a series for forecasts for material conflict between the Israelis and Palestinians for 2010. Our forecast is that the level of material conflict between these belligerents will increase in 2010, compared to 2009.

KEYWORDS: Forecasting, Levant, vector autoregression, Markov-switching, endogeneity
1 Introduction

Scholars and policy makers want to anticipate intra- and international conflicts. They also want to evaluate what might have occurred if certain actions had been taken in the past and (or) what might happen if governments take certain actions in a given conflict in the future. To this end, they developed tools for analyzing state failure, political instability, rebellion, repression and civil war. Organizations like the International Crisis Group, ICG (www.crisisgroup.org) publishes weekly its “Crisis Watch” to help policymakers anticipate and hopefully mediate conflicts world-wide. The Swiss Agency for Development and Cooperation funded a similar effort in recent years (FAST; www.swisspeace.org). Bueno de Mesquita (2009), Schneider et al (2010), Goldstone et al (2010) and O’Brien (2010) describe these and other recent conflict forecasting projects.

An ideal forecasting tool would have at least five interrelated features:

1. It would produce accurate forecasts. This desideratum sounds straightforward. The tool makes a prediction; the prediction either is realized or not. In fact accuracy is a more complex concept. Forecasts involve uncertainty. Even the simple forecasts from regression models contain forecast error, error derived from the random nature of the dependent variable and estimation uncertainty. A forecast therefore is better characterized by an interval rather than a point, more precisely, it is a probability distribution. As we explain
below modern forecast evaluation now focuses on distribution prediction, not point prediction (Gneiting 2008).

2. Useful forecasts need to be produced in near-real-time. A delay of weeks or months often is too late to anticipate (prevent) a conflict from occurring or escalating, and the postdictive analyses that characterize most academic studies of conflict are useful only for model development. And the forecast must be calibrated in time. Not knowing exactly when the predicted outcome will occur diminishes the usefulness of the forecast.

3. Scholars and policymakers are interested in the behavior of collections of belligerents in conflict systems. For instance, an ideal forecasting tool would allow us to anticipate both the actions of one belligerent towards its adversaries but also the reactions of those adversaries and the subsequent behavior of the original belligerent.

4. To improve its performance, the tool should incorporate our knowledge about intra- and international conflict in general and about the history of particular cases. There is a wealth of expert knowledge about this subject (a “wisdom literature”). This knowledge teaches about strategic behavior, the propensity for conflicts to exhibit phase shifts (non-linearities) and other characteristics. We should be able easily to incorporate this knowledge into our forecasting tool.

____________________________________

2Forecast errors and intervals are explained in most intermediate regression books. Yet, many political scientists continue to base their model evaluations on point forecasts. For example, see the recent issue of PS (61(4), October 2008) on competing forecasts of the 2008 presidential election or the discussion of evaluating forecasts of European Union legislative bargaining in Achen (2006).
5. The tool should enable us to make contingent forecasts. We should be able to evaluate counterfactuals about the past history of a conflict as well as to make forecasts contingent on hypothetical behaviors of the belligerents or third parties. For example, we would want to generate contingent forecasts of the impacts of Israeli and U.S. actions on the relations among Israel, Hamas, Fatah, and the U.S. (e.g., Brandt and Freeman, 2006).

Existing tools for forecasting intra- and international conflict meet some but not all of these desiderata. Table 1 catalogues some of the most well known. Space constraints do not permit us to review all these works. So we focus on three. The first set of tools has been developed by Bueno de Mesquita to predict single decisions: an Expected Utility Approach (EUA) and an applied game theoretic model. The accuracy of the former has been heralded for years both by its creator and by government policy makers (Feder 1995; Bueno de Mesquita 1997, 2002, 2010, and www.diiusa.com). Among the EUA’s strengths are its incorporation of rational choice theory and its use of experts to calibrate preferences and other parameters of the model. Contingent predictions can be made with it by altering the model specifications and deriving a (contingent) forecast. The newest game theoretic contribution—an extension of the model in Bueno de Mesquita and Lalman (1992)—has these strengths and incorporates strategic behavior on the part of decision makers with Bayesian updating. The relative performance of this new tool could be placed in multiple locations in Table 1. An example is prediction markets Wolfers and Zitzewitz (2004) describe how index contracts in these markets can be used to generate forecasts of mean levels of variables and conceivably of collections of outcomes.

Recent publications call Bueno de Mesquita the “New Nostradamus”, see www.goodmagazine.com, April 17, 2008.
model has been extensively evaluated with respect to the decisions of the EU in the Thomson, et al. (2006) data set (Bueno de Mesquita, 2010).

Unfortunately these tools fall short on the first and second desiderata. They make no provision for estimation and other kinds of uncertainty in their forecasts. For example, they do not account for experts’ uncertainty about the value of the main parameters of clout, salience and resolve, or, more generally that agents’ utility functions have random elements. For these and other reasons, the predictions of the EUA and new game theoretic models are probability distributions. But they are presented as single values (points) on an issue dimension. This same problem plagues the new game theoretic model and its evaluation; it too presents density predictions as point predictions.5 Neither tool has a meaningful time component. In his earlier EUA model, Bueno de Mesquita (1997, 264) indicates that predicted outcomes are expected to occur in “reasonable amount of time.” But he acknowledges that, in practice, the EUA forecasts suffer from “off-on-time problems.” Finally, the new game theoretic model focuses on iterations

5 There are only limited references in the EUA literature and in the new game theoretic contribution models as to the processes of elicitation eliciting resolve scores from subject matter specialists. In fact, this is a common difficulty in these modeling enterprises (Achen 2006a). Elicitation necessarily gauges the assessor’s degree of certainty in the information she or he is providing the analyst. Assessors do not provide analysts with point estimates but rather with densities. Hence the predictions derived from models that use these elicited parameters also must be densities of some kind. The derivation and evaluations of point estimates in forecasting, along with measures like absolute mean error, median percentage error, etc. therefore are ill-conceived. For a discussion and application of elicitation methods in political science see Freeman and Gill (2008, 2006).
of a single game: its applicability to repeated games—especially for games representing whole conflict systems—is unclear.\(^6\)

Table 1: Some Approaches to Forecasting Conflict in International Relations

<table>
<thead>
<tr>
<th>Type</th>
<th>Ex Post</th>
<th>Ex Post Counterfactual</th>
<th>Ex Ante</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Shot Decision or Event</td>
<td>Game theoretic (Brams and Togman, 2000; Achen 2006a,b; Bueno de Mesquita, 2010; Schneider et al. 2010)</td>
<td></td>
<td>Expected Utility Approach (Organski and Lust-Okar, 1997, Bueno de Mesquita, 2002, 2010); Expert Systems (Tetlock, 2005); Role Playing (Bennett and McQuade, 1996)</td>
</tr>
<tr>
<td>Conflict System: Directed Dyadic Behavior</td>
<td>Cluster-Density Analysis (Schrodt and Gerner, 2000); BVAR (Brandt and Freeman, 2006); B-SVAR (Brandt, Colaresi and Freeman, 2008)</td>
<td>bvvar (Brandt and Freeman, 2006)</td>
<td>International Crisis Group (ICG); Swiss Peace Foundation (FAST); VAR (Pevehouse and Goldstein, 1999)</td>
</tr>
<tr>
<td>Behavioral Phase Shifts</td>
<td>Integrated Crisis Early Warning Systems (ICEWS) (O’Brien, 2009); Hidden Markov Models (Schrodt, 2000)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^6\)The related work of Organski and Lust-Okar (1997) suffers from the same off-on-time problem. They note that their predicted outcome for the Jerusalem negotiations does not indicate when the final agreement would be reached (Organski and Lust-Okar 1997, 350). Briefly, while early on in his article, Bueno de Mesquita (2010) draws a distinction between iterated and repeated games, later this distinction becomes less clear. He also acknowledges that the usefulness of his game theoretic model breaks down as one moves from the realm of single-shot, noncooperative games to repeated cooperative games.
The Political Instability Task Force, (PITF), the successor to the earlier State Failures Project (Esty, et al. 1995; Esty, et al. 1998; Bates et al. 2003; Goldstone et al. 2010) is a second major contribution to conflict forecasting. It produces predictions of political instability over several years based on strongly restricted, single equation logistic regression models with samples matched by region and time. The variables in these models are chosen on the basis of theory and expert judgment. Most of the accuracy assessments in this literature are *ex post* in nature (Goldstone, et al. 2010; King and Zeng, 2001). However it is likely that unpublished *ex ante* assessments of political instability have been produced inside the U.S. government.\(^7\)

The PITF project also falls short on several of the desiderata. To begin, forecast error is not incorporated in its accuracy assessments. No mention of estimation uncertainty is made in the model scores to which the cutpoints are applied; no error bands or confidence intervals are supplied for any of the forecasting results or comparative odds ratios (Goldstone, et al. 2010; 194, Figures 2 and 3, Tables 2 and 3). Second, the PITF models are highly aggregated in space and time. This is because the dependent variable is conceived as a *single* episode that may encompass sequences of events over periods of up to five years (*ibid*, 192). These results tell us very little about short-term conflict dynamics or how the directed behavior of one belligerent affects the behavior of other belligerents over time. Understanding this kind of behavior is critical to anticipating outbreaks of violence weeks and months ahead of time. Finally, its strong—usually

\(^7\)King and Zeng (2001) and Beck, King and Zeng (2000) criticize the State Failures Project. Their contribution is a neural net model that uses strongly restricted, single equations of the nested logit type. Their demonstration of the accuracy of the neural net model also is *ex post* in nature. The newest work reported by the PITF (Goldstone, et al. 2010) clearly is informed by these critiques and employs rare events estimators and a more sophisticated case control design.
untested—exogeneity assumptions ignore theoretical knowledge about the relationships among political and economic variables, such as the extent to which conflict diffuses across regions through time.\(^8\)

Brandt and Freeman (2006) and Brandt, Colaresi and Freeman (2008) employ a third approach. Using data from the Kansas Events Data Project (KEDS), they show how Bayesian time series models can be used to analyze conflict dynamics and to make \textit{ex post} forecasts in the Levant over the short-term. Their models are Bayesian vector autoregressions (BVAR) and Bayesian structural vector autoregressions (BSVAR). Their specifications use a modification—incorporating expert judgment of international relations scholars—of the Sims and Zha (1998) prior from macroeconomic analysis. The models produce \textit{ex post} forecasts with error bands for weekly and monthly patterns of conflict and cooperation (directed dyadic behavior) of the Israelis, Palestinians and United States. In one of these analyses, the impact of Jewish public opinion is incorporated (\textit{ibid.}). Counterfactual (\textit{ex post}) forecasts are reported in both the pieces.

The work of Brandt, et al. suffers from two problems. First, Brandt, et al.’s accuracy assessments are based on the mean forecasts from their models; they rely on measures like residual mean square error (RMSE) of the respective point forecasts to evaluate their models. This is in spite of the fact that their approach produces—via computational simulation or Gibbs

\(^8\) Illustrative of an unreasonable exogeneity assumption is the practice of treating neighboring conflict in four or more bordering states as an independent variable. The practice of assuming economic variables are exogenous when they are clearly affected by episodes of political instability—in other words, when causation runs both ways from and to economic variables—may explain why Goldstone, et al. (2010) find few relationships between them and political instability since the respective equations suffer from endogeneity bias.
sampling—the full probability distribution for their forecasts. As we note below, Chatfield (2001) and others show that RMSE and related criteria are sensitive to outliers. Gneiting (2008) and other statisticians recommend using the full predictive density for forecast evaluation. In addition, all of Brandt et al.’s work is ex post in nature. They have not produced any ex ante forecasts with their Bayesian time series models.

We propose to build on the approach of Brandt, et al. because it applies the advances in forecast evaluation from statistics, produces temporally disaggregated forecasts in real-time, allows for the analysis of whole conflict systems, and incorporates expert judgment in the form of Bayesian priors. Specifically, we propose a technology that accurately forecasts the weekly behavior of medium-sized conflict systems in real-time. This approach’s structural features and Bayesian prior, and the basic statistical models and estimation strategies will be based on existing theoretical and empirical research on conflict dynamics. We focus especially on those theories and empirical studies that illuminate and analyze phase shifts in the behavior of belligerents, shifts that can be traced to the multiple equilibria of games of incomplete information and to the invasion of dynamic versions of such games by certain strategies (e.g., Diehl, 2006), to path dependent sequences of cooperative and conflictual events (Schrod and Gerner, 2000; Huth and Allee, 2002), to multiple equilibria in strategies played by audiences and elites in two level games in (non-) democracies (Riou, 1998; Huth and Allee, 2002b; Rousseau, 2005), and to psychological triggers that produce different types of cooperative and conflictual behavior (Keashly and Fisher, 1996; Senese and Vasquez, 2008). This theoretical work is supplemented empirically by an extensive set of datasets that also employ crisis phase concepts: the most notable of these are the early Butterworth-Scranton conflict management dataset (Butterworth and Scranton, 1976), CASCON (Bloomfield and Leiss, 1969; Bloomfield and Moulton, 1997;

These theoretical and empirical phase shifts are consistent non-linearities in the conflict processes. Our methodology represents them with a Markov-switching process in Bayesian time series models. In terms of measurement, we employ event data series from the KEDS project. The time series are constructed from archival and real-time text feeds by automated text parsing. The choice of the actual forecasting model for any given conflict or region will be based on forecast density evaluation using probability integral transform methods and other new tools developed by statisticians.

This new technology will produce theoretically grounded, policy-relevant forecasts, including contingent forecasts. All the results and tools will be available in the public domain. In this first stage of our project, real-time ex ante forecasts for three major conflict systems will be produced. More conflict systems will be included in future stages. Finally, we will actively monitor and compare our forecasting methods to other statistical forecasting models.

2 Research Design

Our design consists of four parts: production of event time series for selected cases, incorporation of statistical advances on forecast evaluation and phase shift analysis, comparison of our forecasting models to others, and website design and management. We explain these parts in the

---

9 In 2010 this project migrated from the University of Kansas to Pennsylvania State University, and we have updated the URLs to reflect this. However, the data and software we are using here were largely developed at Kansas, so we will refer to it using the KEDS name.
next sections. Figure 1 outlines our overall data generation, model building, and forecasting approach.

2.1 Data and Case Selection

Event data are used in our project. These are data on the actions between two parties to a conflict, nominal codes indicating who did what to whom and when. These data will be produced by the state-of-the-art TABARI automated coding software\textsuperscript{10} using the CAMEO event and actor coding system.\textsuperscript{11} TABARI can analyze text that is downloaded from data services such as Lexis-Nexis or Factiva, or in real-time—at for essentially no cost beyond that of a web connection and web-enabled computer—via real simple syndication feeds (RSS). RSS aggregators such as Google News\textsuperscript{12} and the European Media Monitor\textsuperscript{13} each monitor over 4,000 sources, including both international media sources such as the \textit{New York Times}, \textit{Christian Science Monitor}, \textit{BBC}, \textit{Agence France Press}, and \textit{Reuters}, as well as thousands of local sources. The availability of these aggregators, and RSS feeds more generally, provide an unprecedented ability to monitor global political behavior in real-time.

\textsuperscript{10}http://eventdata.psu.edu/software.html

\textsuperscript{11}http://eventdata.psu.edu/data.dir/cameo.html

\textsuperscript{12}http://news.google.com/

\textsuperscript{13}http://emm.jrc.it/overview.html
The automated coding program TABARI works in conjunction with the CAMEO event coding system, which has extended the scope of the Cold War-era WEIS and COPDAB inter-state coding systems to provide detailed coding of intra-state conflict and to incorporate contemporary, post-Cold War theoretical concerns and concepts. TABARI also provides a detailed and systematic framework for coding a wide variety of political actors, whether international, supranational, transnational, or internal. From the coded text we can extract interval measures of inter-state and inter-group cooperation and conflict, count data on various types of behaviors, and nominal data such as event sequences or patterns. Because text data are available and updated frequently, and TABARI’s coding is fully automated, we will have a continuous flow of information on a wide range of variables and actors. Because it is informed by social
science theory in how it catalogs actors and actions via dictionaries, the TABARI/CAMEO routines will be much more efficient and accurate than routines used by unsupervised data miners or by manual coding.

Retrospective automated event data coding from a single source (typically *Reuters* or *Agence France Presse*) is now a well-established procedure, and almost all studies of political conflict in refereed political science literature over the past fifteen years have used this type of data. The challenge is moving this to multiple sources in real-time. Specifically the three key challenges are

1. Automated detection and classification of new political actors, as well as the re-classification of existing actors (e.g., a political party switching from government to opposition). To a much lesser extent, event coding systems need to be updated as new political vocabulary—such as the emergence of the phrase “ethnic cleansing” in the early 1990s—comes into use.

2. Effective filtering of historical and non-political stories (e.g., sports and business stories that use military metaphors) from a multiple-source stream, as well as the efficient detection of duplicate reports of the same event.

3. Calibration of event reports across multiple sources and regions: it is well known that some areas of the world are covered less extensively than others (e.g., Africa is covered less than East Asia), and as one shifts to multiple news courses, editorial policies such as the mix between political and economic stories also affect the data. Based on some recent experiments, item response theory (IRT) and other scaling methods appear to have great potential in resolving this issue (Schrodt, 2007). But additional work is required to apply them.
In addition to these specific issues, we also will deal with some general issues involved in scaling conflict systems, and some additional refinements to the CAMEO coding systems as we take that out of the Middle East, the Balkans and Africa—the geographical domains where most of the development has been done—into other parts of the world. The open-source TABARI/CAMEO system recently was employed in the U.S. Department of Defense ICEWS project (O’Brien 2010) to code 25 gigabytes of Asian news reports involving more than 6.7 million stories and 253 million lines of text from 70 news sources, so we are fairly confident that the system is sufficiently robust to be able to generate the data that we need.

In this, the first stage of project, we will focus on three cases. The first is the Levant, specifically the Israel-Hamas-Fatah and Israel-Lebanon-Hizbollah cases. The KEDS project has a retrospective dataset on this conflict that provides thirty years of coverage14, more than ample data for estimating and assessing retrospective models. Furthermore, the news coverage of the Levant is continuous and dense. Schrodt and his collaborators have used this data in several retrospective predictive studies (Hudson, Schrodt and Whitmer, 2008; Schrodt 2006; Schrodt and Gerner 2004; Schrodt 2000), and Schrodt also has done field work in the region. Brandt and Freeman already have applied some Bayesian time series models to this case ex post (Brandt and Freeman, 2006; Brandt, Colaresi, and Freeman, 2008). Finally there are a number of competing forecasts to which we can compare the success of our technology. Note however that while we have extensive experience with the collection and analysis of these, our collective past analyses of KEDS data for the Levant have not produced real-time forecast densities like what we are proposing here. Our goal is to extend this to real-time forecast density estimation.

14http://eventdata.psu.edu/data.dir/levant.html
In addition, to demonstrate the generality of our new approach, we also will forecast China-Taiwan-US relations and India-Pakistan-US relations. As just noted, TABARI/CAMEO very recently has been used to analyze a large body of Asian news text. So we have experience with the respective data sources.

2.2 Bayesian Models for Forecasting and Modeling Nonlinearities in International Relations

2.2.1 Bayesian Forecasting in International Relations

To forecast systems of behavior in time we need a multi-equation time series model. Consider the following:

\[ \sum_{i=0}^{p} y_{t-i} A_i = d + \epsilon_t, t = 1,2,\ldots,T \]  

(1)

where \( y_t \) is is a \( 1 \times m \) dimensional vector say for the directed dyadic behavior of a three actor system, \( A_i \) is a \( m \times m \) matrix of coefficients for a set of lag polynomials representing the memory in the system back to lag \( p \), \( d \) is a \( 1 \times m \) vector of constants or deterministic variables such as electoral calendar counters, and \( \epsilon_t \) is a \( 1 \times m \) dimensional error vector assumed to be uncorrelated with the \( y_{t-s} \) for all \( s > 0 \), serially uncorrelated, and with individual variances equal to unity. The reduced form of equation 1 is

\[ y_t = c + y_{t-1} B_1 + \ldots + y_{t-p} B_p + u_t, \quad t = 1,2,\ldots,T \]  

(2)
where $y_t$ again is an $1 \times m$ vector of observations at time $t$, $B_l$ is the $m \times m$ coefficient matrix for the $l$th lag, $p$ is the maximum number of lags, and $u_t$ are the reduced form residuals.\(^{15}\) Many scholars have used similar reduced form models to study conflict dynamics in the Levant and Bosnia (e.g., Goldstein, Pevehouse, Gerner and Telhami, 2001; Pevehouse and Goldstein, 1999; Goldstein and Pevehouse, 1997), as well as to study policy counterfactuals (Goldstein and Freeman, 1990, 1991). But to our knowledge they have not been used to produce \textit{ex ante} statistical forecasts.\(^{16}\)

There are at least two major problems with such models. First they are overparameterized. For a four actor illustration (e.g., the U.S., Israel, Hamas, Fatah) $y_t$ would contain the $m=12$ directed, dyadic behaviors in the conflict system. If the VAR representation of this conflict uses $p=6$ lagged values and $d$ is a vector of constants, the VAR model has 876 parameters. Clearly this produces a tremendous amount of estimation uncertainty and therefore a lack of precision in the model’s forecasts.\(^{17}\) Overparameterization also can produce mistaken inferences about

\(^{15}\)The reduced form in equation (2) is related to the simultaneous equation model in equation (1) by

\begin{align*}
c &= dA_0^{-1} \quad B_l &= -A_lA_0^{-1} \quad l = 1, 2, \ldots, p \quad u_t &= \epsilon A_0^{-1}.
\end{align*}

\(^{16}\)For example, Pevehouse and Goldstein’s 1999 article, “Serbian Defiance or Compliance in Kosovo? Statistical Analysis and Real-Time Predictions” is based solely on causality tests in a reduced form model of the respective conflict.

\(^{17}\)Perhaps this is why Goldstein and Pevehouse and Goldstein et al. do not produce any impulse responses or forecasts for their models (they just report causality test results). goldstein1990 also do not produce confidence intervals (error bands) for their policy counterfactuals.
persistence, more specifically, about stationarity (Sims and Zha, 1998). But knife-edge tests for non-stationarity (unit roots) also can be erroneous. These tests can produce pretest bias in multivariate time series analysis (Freeman et al., 1998).

Bayesian multivariate time series analysis addresses these and other problems. By using expert knowledge in a Bayesian prior on the coefficients in equations 1 and 2, we are able to reduce the estimation uncertainty and obtain more precise dynamic inferences and forecasts. This prior applies to the weight of behavior at distant lags and the degree to which we believe the series might be non-stationary and possibly cointegrated. In the Bayesian model these restrictions are inexact whereas in the frequentist model they are exact and knife-edge. Brandt and Freeman (2006, 2009) explain how a prior for the study of intra- and international conflict (cooperation) can be constructed and applied to equations 1 and 2 (see also Brandt, Colaresi, and Freeman, 2008). This work also explains how concepts like log marginal data densities (Chib, 1995) can be used to calculate Bayes factors (Kass and Raftery, 1995) to assess the in-sample fit of these Bayesian time series models. Finally, all these developments are based on posterior Markov chain Monte Carlo (MCMC) samples. So the technology needed to produce predictive distributions for these models is available.18

2.2.2 Markov-switching Bayesian time series models

Our theories (e.g., Lebovic, 1994; Lund, 1996; Gurr and Harff, 1996) teach us that human conflicts are subject to phase shifts or fundamental non-linearities. This is another reason why normality assumptions break down in time series forecasting: non-linearities caused by conflict

18Brandt has implemented these models in an open source R package called MSBVAR,
http://yule.utdallas.edu/code.html
phase shifts are bound to produce fat tailed (leptokuric) forecasting distributions. Unlike financial analysts who model this behavior with “heavy tail distributions,” we propose to model the non-linearities with Markov-switching processes. These models allow us to capture the phase changes and non-normality of the data using Markov and mixture models (fruhwirth06). Specifically, we will include in our evaluation of contending forecasting models, Markov-switching, Bayesian multivariate time series models.\(^{19}\)

A Markov-switching version of the Bayesian VAR model (in equations 1 and 2) can be represented as a set of \(h\) state VAR models. The structural Markov-switching Bayesian VAR (MS-BVAR) version is

\[
\sum_{l=0}^{p} y_{t-l} A_l(s_i) = d(s_i) + \varepsilon_i(s_i), \quad t=1,2,\ldots,T, \tag{3}
\]

where \(s_i=j\) is an \(h\)-dimensional vector indexing the state of the process where \(j\) is the integer label for the state, with a \(h \times h\) Markov transition matrix \(Q\). The rows of \(Q\) give the probability of transitioning from state \(s_{t-1}\) to \(s_t\), or \(Pr(s_t=k|s_{t-1}=j)\) for states \(k\) and \(j\). Note that while the models in equations (1) and (2) generically have \(m^2p+m\) regression parameters, the model in equation (3) has \(h(m^2p+m)\) parameters. As noted earlier, for a 12 equation, 6 lag model the model in equation (2) has 876 parameters. So if we allow a Markov-switching process with 2 states (say high and

\(^{19}\)On the prevalence of fat tails in forecast distributions in economics and finance see such works as Granger (2005). Note that our models will allow us to capture more permanent structural breaks in conflict as well. These breaks are simply switches with high degrees of permanence.
low conflict intensity / volatility) there will be 1752 parameters. So a Bayesian approach is even more sensible when allowing for non-linearities and phase shifts, since it allows us to deal with the proliferation of parameters.

A prior for this model has to include beliefs about the \( A_i \) and \( d \) parameters as well as the Markov process \( Q \). We already have been able to adapt the random walk prior in Sims, Waggoner, and Zha, 2008 and Sims and Zha, 1998 to serve as a baseline set of beliefs in an analysis of international conflict event data (Brandt and Appleby, 2007). The sensitivity of this prior can be evaluated using methods like those discussed in Brandt and Freeman (2009).

Forecasting with these models is a straightforward application of data augmentation via Gibbs sampling. So we can generate dyadic, phase or state specific \textit{ex ante} and \textit{ex post} counterfactual forecasts. The MCMC output of the Bayesian model in equation (3) also can be used to estimate the number of states or phases in the model (Sims, Waggoner, and Zha, 2008; Frühwirth-Schnatter, 2006). We will also be able to evaluate these forecasts in the manner discussed below: forecast densities that account for the possible state-dependent mixtures of forecast distributions. This allows us to capture both fat-tailed and other non-linearities in a sensible manner. Finally, the forecasts from a Markov-switching VAR model like equation (3) are a weighted combination of the forecasts for each state or phase. This means that the forecasts account for both the inherent variability of the time series themselves, but also the parameters and the phases. Thus,

\[ h(m^2p+m) \] parameters spread over \( 2^m(h!) \) possible posterior modes. Sampling from these modes is computationally intensive but feasible (Frühwirth-Schnatter, 2001; Scott, 2002).

\[ \text{The MS-BVAR model outlined here can capture changepoint, SETAR, TAR, and other switching and non-linear time series models as special cases (Frühwirth-Schnatter, 2006; Krolzig, 1997).} \]
the uncertainty of changepoints or phase shifts is captured directly in the estimation and forecasting process enabling additional analyses of how regime changes impact the ability to predict conflict (cooperation).

### 2.3 Forecast Densities and Evaluations: Making the Transition from Point to Distribution Prediction

How then do we build the most *scientifically sound and useful* Bayesian time series model for forecasting intra and international conflict (cooperation) out-of-sample (*ex ante*) in real-time? The answer is that we need to make the transition from point prediction to distribution prediction and to apply the technique of “recalibration” (Gneiting, 2008; Timmerman, 2000; Tay and Wallis, 2000). Statistical forecasting in political science still uses point prediction to evaluate model performance. In fact, we usually use a residual mean square error (RMSE) criterion for this purpose. First, there are many such criteria for point forecasts (Wallis, 1995; Diebold and Lopez, 1996; Chatfield, 2001). And RMSE is one of the weakest criteria because it is sensitive to outliers. But more important, point prediction does not tell us anything about the (un-)certainty of a forecast. And reliance on normality assumptions to construct forecast confidence intervals is prone to mistaken inferences since they amount to strong, often inaccurate assumptions about the symmetry and about the size of the tails of a forecast distribution.²²

One alternative is to use prediction intervals. Put simply, this is the practice of constructing the upper and lower limits between which the future values of a series are expected to lie with

---

²²For a succinct summary and critique of these normality assumptions see Tay and Wallis (2000, esp.235-236).
some prescribed probability (Chatfield, 2001; Granger, 2005) Tools like fan charts are used to represent these intervals. Since the late 1990s the Bank of England has presented its forecasts for inflation and growth in this way (Gneiting, 2008). For various reasons these intervals are often difficult to apply. One is that until recently many software packages did not produce them. More fundamentally, they are apt to be too narrow on average (Chatfield, 2001).

The favored approach in statistics is density evaluation, or what is called “probability forecasting” (Rosenblatt, 1952; Dawid, 1984). A density forecast is “a complete description of the uncertainty associated with a prediction, and stands in contrast to a point forecast, which by itself, contains no description of the associated uncertainty” (Tay and Wallis, 2000, 235). Our plan then is to use our Markov-switching Bayesian multivariate time series model to produce entire forecast densities for all our variables for conflict and cooperation between belligerents and to evaluate these densities for their accuracy through time. On the basis of the results we will recalibrate our model gradually improving its performance. We will make all of the data forecasts publicly available via a project website. We will also compare our forecasts to those generated by other methods and parties such as the recently announced political forecasting website—The Call—by the Eurasia Group in collaboration with Foreign Policy magazine (http://eurasia.foreignpolicy.com/node) and any additional such sites that come on-line during the period of the grant.

_________________________

Space does not allow for a full presentation of interval forecasting or for a discussion of the methods that have been developed to evaluate interval forecasts. (Christoffersen, 1998; Taylor, 1999; Tay and Wallis, 2000).
Several tools are used in this process. The primary tool is the probability integral transform (PIT). The PIT is defined for a series of realized values and forecasts. Let \( \{y_t\}_{t=1}^m \) be a series of realizations from some conditional density \( \{f(y_t | \Omega_{t-1})\}_{t=1}^m \) where \( \Omega_{t-1} \) is the (past) information set at time \( t \). Then if the one-step ahead density forecasts \( \{p_{t-1}(y_t)\}_{t=1}^m \) equal \( \{f(y_t | \Omega_{t-1})\}_{t=1}^m \) (and assuming a non-zero Jacobian), the PIT of \( \{y_t\}_{t=1}^m \) with respect to \( \{p_{t-1}(y_t)\}_{t=1}^m \) is i.i.d. \( U(0,1) \), or

\[
\{z_t\}_{t=1}^m = \left\{ \int_{-\infty}^{x_t} p_{t-1}(u) \, du \right\}_{t=1}^m \sim U(0,1)
\]

If the forecast density is correct, then the \( z_t \) is an independent uniform \( U(0,1) \) variate and the model is well-calibrated. This result extends to the case of time-dependent density forecasts and to multivariate forecasts. Deviations in \( z_t \) from uniform iid are indications that the forecasts do not capture some aspect of the data generating process. One uses graphical tools—comparison of the plot of the \( z_t \) with the 45 degree line that is the uniform’s distribution function, histograms, Kolmogorov-Smirnov and Cramer-von Mises tests, and correlograms of the levels and powers of the PIT to make this determination. Still newer tools that complement PIT histograms are calibration plots and sharpness diagrams. Using these tools we will be able to rank the

---

24 The remainder of this paragraph’s definition of the PIT is taken from Diebold, Hahn and Tay (1999, 661-2). See Diebold, Gunther and Tay (1998, 867-8) for a proof of why the PIT has the properties reported here.

25 In a recent paper, Gneiting, Balabdaoui and Raftery (2008) review work (Hamill, 2001) that shows the PIT is only necessary not sufficient for forecasts to be ideal. They propose a “paradigm of maximizing sharpness of the predictive distribution subject to calibration.” The tools of calibration plots and sharpness
performance of alternate forecasting models. By recalibrating the best of them, we will produce better forecasts of the Levant, South Asia and East Asian cases. Finally, if forecasting models can be found that capture the true data generating process, from a decision theoretic perspective, all analysts will prefer it regardless of their loss functions (Diebold, Gunther and Tay 1998; Diebold, Tay and Wallis, 1998). Hence our best models will aid decision-making.

In the first stage of our project, we will evaluate competing forecasting models for intra- and international conflict (cooperation) for our three cases and evaluate them with the PIT and the tools mentioned above. The competing models will include theoretically relevant actor clusters and variables like election forces. To benchmark their performance we will include naïve models like univariate autoregressive moving average specifications. But Bayesian reduced form models like that in equation (2) and its Markov-switching variant in equation (3) will be the main objects of evaluation. Therefore we will need to apply multivariate techniques for density forecasting (Kling and Bessler, 1989; Diebold, Hahn and Tay, 1999; Clements and Smith, 2000). Thus we will produce forecasts that meet our earlier desiderata. As with any scientific enterprise, we expect to learn as much from where our models go wrong as from where they are correct. Erroneous predictions are key to determining the gaps and weaknesses in our existing theoretical concepts, variables, and measurement methods. The advantage of systematic forecasting such as the system we are proposing is that we can re-assess the models based on the sets of assumptions based on different theoretical perspectives, particularly as these feed into the model through the diagrams are proposed and illustrated in their article. We hope eventually to apply these most advanced tools in our project.

26For example, Kling and Bessler (1989) illustrate how a Bayesian multivariate model with a Litterman prior can be recalibrated in time to produce better forecasts of four variables for the U.S. macroeconomy.
priors. With the high-speed automated coding provided by the TABARI system, we can also experiment with the theoretical constructs embedded in the event coding ontologies. For example by further differentiating sub-state actors, or differentiating types of events that are currently lumped together into a single coding category, but which can be differentiated in the texts from which they were originally derived, we can determine how definitions of actors and events affect our ability to produce high quality forecasts.

3 Illustrating Phase Baseline Forecasts for the Levant

To illustrate our approach, we generate a set of *ex ante* forecasts for material conflict among the Israelis and Palestinians. This is only an example of what we propose here: we plan to expand this example to include more equations (actors), mediation and cooperation event series, less aggregated data (weekly instead of monthly), and more complex dynamics. With these caveats, the goal here is to point out the advantages of MS-BVAR and BVAR models for forecasting conflict processes.

The case study for this analysis is data on the Israeli-Palestinian conflict. We employ monthly CAMEO-coded data from *Agence France Presse* (hereafter, AFP) from 1996:1-2009:12.\(^27\) We aggregated the data for U.S., Israeli and Palestinian material conflict events for each month.\(^28\) We use these data because they capture the second intifada and periods of recent high and low conflict. For the six dyadic material conflict series (ISR2PAL, PAL2ISR, USA2ISR, ISR2USA, ...

\(^{27}\) We also did this analysis using data from the Reuters’ wire service. The results from those data are consistent with what is reported here.

\(^{28}\) These counts of material conflict events most correlate strongly with the Goldstein-scaled WEIS data from earlier KEDS coding.
PAL2USA, USA2PAL) we fit two models. The first is a Bayesian VAR with one lag using the prior specified in Brandt and Freeman (2006). The second is an MS-BVAR model with one lag where the prior for each regime or state is the same symmetric prior used in the BVAR model. The diffuse prior for the regimes is Dirichlet.\(^{29}\)

\[^{29}\text{For the MS-BVAR model we find the initial posterior mode and the regime probabilities via the EM algorithm. We then draw 20000 values from this posterior mode to construct the subsequent inferences and forecasts. The BVAR model posterior has a known analytic posterior density. We sample 20000 draws directly from this density to construct the BVAR forecasts. See Brandt and Freeman (2006) for details on the latter.}\]
Figure 1 shows the probabilities of the two-regime MS-BVAR model for the sample. Note that the high conflict periods (the black regimes) are those that correspond to episodes of very high regional material conflict. The regime switches come at some clearly defined points in the conflict: from late 1999 until early 2005 the regime with the higher probability is generally the high conflict (black) regime. This corresponds with the second intifada. The period in 2005 and 2006 corresponds to the series of Qassam rocket attacks by Palestinians on Israel. The period in 2008 when the red or low conflict regime has higher probability corresponds to the Israeli-Hamas truce followed by stepped up rocket attacks. Thus, there is evidence for changes in parameters throughout the recent years of the conflict. The first regime occurs for about 60% of the sample period and one where the conflict dynamics are nearly non-stationary, has an equilibrium level of twice as many material conflict events than the lower conflict regime, and a variance that is more than 50% larger than the lower conflict regime. The second regime is one where the conflict stays near a status quo level with lower volatility.30

Most importantly, the Israeli-Palestinian conflict’s current state is the high conflict regime. The in-sample predicted regimes moved back and forth from the high to low conflict regimes throughout 2009. The December 2009 regime is the high material conflict regime with a

\[ Q = \begin{pmatrix} 0.84(0.09) & 0.16(0.09) \\ 0.24(0.10) & 0.76(0.10) \end{pmatrix} \]

where the standard errors are given in parentheses.

\[ \text{Further evidence of the MS process is the posterior transition matrix. If the transition matrix is equal across its rows then this is just a mixture process. This is not the case here: the data generation process is consistent with Markov-switching.} \]

30
predicted probability of 0.75. This persists in forecasts for 2010. Figure 3 presents 12 months of forecasts (covering all of 2010) for the material conflict series. Note that these are a weighted combination (based on the probabilities of each regime, forecast 12 months into the future). Thus, this is a Markov-switching distribution of the high and low conflict regimes. An important caveat to these current forecasts is that they are probably weakened by the fact that in this model we are not differentiating between Gaza and the West Bank. At present, those two systems seem entirely decoupled, as even very high levels of violence in Gaza have had almost no effect (other than a few demonstrations) on violence in the West Bank. We should be able to correct this in the future by looking separately at Fatah and Hamas actions.
The median forecasts across the BVAR and MS-BVAR models are very similar. For both the BVAR and MSBVAR models, there is an upward trend in the number of material conflict events forecasted for 2010. The BVAR model forecasts are for a higher level of conflict than what is seen in the MSBVAR models. This is contrary to some current journalistic accounts that see little prospect for change in this conflict in 2010 (Ephron, 2010). The differences across the two models come in how we would characterize the forecast density coverage of the predictions. In general, the BVAR forecast credible intervals that assume constant parameters are too small: they are almost always within the forecast credible intervals of the MS-BVAR forecasts (which allow for the non-zero probability for the high conflict regime). Thus Figure 3 indicates that the risk of additional material conflict between the Israelis and Palestinians is different when we consider the MS-BVAR versus the BVAR forecast models. We believe that this serves as initial evidence for our claims that there are parameter shifts in the data that should be modeled.

This is a rather circumscribed example: in later iterations we plan on expanding this in the manner discussed at the start of this section. The inclusion of additional dyadic conflict and cooperation series (both material and verbal, per the CAMEO scheme) should provide us with more evidence for identifying parameter shifts and regime switching like that uncovered here.

4 Conclusions

Our plan is to expand this forecasting exercise in several directions. First, we plan on covering multiple conflicts around the globe – in addition to the Levant, this project will expand to cover India and Pakistan and China and Taiwan. We have already made efforts to establish the collections of actors needed to code the data for these latter two sets of conflict. Second, we have to further scale up the number of actors, dyads, and measures included in our forecasts.
The previous paragraph belies the technical complexity of our proposal. The estimation and summarization of forecast densities for even a simple BVAR model are computationally intensive: simulating the 100,000 posterior draws of nine month forecasts for six and seven equation (directed-dyad) models in Brandt, Colaresi, and Freeman (2008) required over one hour of CPU time.\(^{31}\) At present the sampling and forecasting for the 20000 draws what we have reported here takes 12 minutes on a current server.\(^{32}\) While this is an impressive decline in computation time, the burdens of estimating more advanced Markov-switching versions of these models is several orders of magnitude larger because we need to investigate forecasting models with different numbers of states and VAR lags (see below). Further, as we proposed in section 2.1, we will be forecasting multiple conflicts: Israel-Fatah-Hamas, Israel-Lebanon-Hizbollah, India-Pakistan, and China-Taiwan. In each of these cases there are many or more actors than in the \textit{ex post} forecasts of Brandt, Colaresi, and Freeman (2008), since the U.S. and other regional actors would also be part of the models. So covering the four conflicts would take at least 4-6 hours for the sample number of periods, just to produce one set of MS-BVAR forecasts which is serious computational time across the different model orders, conflicts, and forecast horizons. Past experience with the BVAR and MS-BVAR models has shown that each additional actor added to a conflict increases the Bayesian posterior simulation time exponentially (since adding one actor increases the scale of the problem from \(2^{m-m} \) to \(2^{m+1-(m+1)}\)). Further, each additional Markov-switching state increases the number of posterior modes that must be sampled from \(h!\) to \((h+1)!\) Even for small values of \(m\) and \(h\), the initially estimated times grows by a factor of

\(^{31}\)Time measured on an Apple G5 dual core 2.0Ghz PPC 970MP processor.

\(^{32}\)Time measured on an Apple dual quad-core 2.26 Ghz server.
4(h!)(2^m−m). So covering our basic three-actor models \(m=6\) for a two state model \(h=2\) would grow greatly to simulate the forecast densities over the conflicts. Our goal is to further work on these issues to reduce the computational burdens and time so that this becomes a feasible exercise.

Finally, once we expand our forecasting capabilities, we will work on further calibration and refinement methods. The plan is that these will be part of an on-going comparison of the \textit{ex ante} forecasts to the actual data. Once we have forecasts generated we will be able to compare the \textit{ex ante} forecasts to the actual data.

A possible reservation to this project may be that it is either already being done in the intelligence community, or that it should be done there rather than by us. We have several responses to this. First, to the best of our knowledge, none of what we are proposing duplicates existing efforts. Hollywood portrayals of an intelligence community endowed with vast technological capabilities for analysis are fiction, not reality, and at present the ability to collect information far outpaces the ability to analyze it. Furthermore, the public record of recent quantitative research sponsored by the U.S. security communities—the CIA’s PITF (Goldstone, et al., 2010) and DARPA’s ICEWS (O’Brien 2010)—involves methodologies similar to (though still distinct from) those we are using, and it seems quite unlikely that these projects would be funded if superior classified models were already in use.

Second, while forty years ago there appeared to be little point in doing work that did not use classified information because this was, in principle (if not always in reality, as the experience in Vietnam indicated) vastly superior to the unclassified work, that situation has changed dramatically with the information sources available for strategic analysis on the Internet. Since we are trying to make forecasts at policy-relevant lead times of six months to five years, not
trying to figure out where Osama bin-Laden is sleeping tonight, the additional information provided by classified sources—Hollywood notwithstanding—is not only minimal, but often dysfunctional (Lowenthal, 2009; Johnson and Wirtz, 2004).

Third, there is an increasing interest in technical political forecasting in the NGO community. Two recent examples of this would be the SwissPeace FAST project\(^{33}\) which operated from 1998-2008, and the current Armed Conflict Location and Event Data (ACLED) project of the Peace Research Institute Oslo.\(^{34}\) Both the ability to generate real-time event data at a very low cost and some of the quantitative forecasting methods will be of interest to NGOs involved in either conflict monitoring or conflict forecasting. If successful, our system would be particularly useful in assisting NGOs who are called upon to provide material services such as food, medical support and refugee housing in areas experiencing violent political conflict. Acquisition and transportation of these resources typically requires three to six months of advance planning, particularly when large-scale relief supplies that must be shipped by sea are required. We are not suggesting that our models will be the sole factor in these plans, but—given the known problems with the accuracy of qualitative forecasting (Tetlock, 2005)—they can be another input. Quantitative models are already used for the advance planning of reactions to natural disasters: for example, the models of the U.S. Agency for International Development Famine Early

\(^{33}\)http://www.swisspeace.ch/typo3/en/peace-conflict-research/early-warning/index.html; for several years Schrodt was involved in generating event data for the Israel-Palestine conflict component of FAST.

\(^{34}\)http://www.prio.no/CSCW/Datasets/Armed-Conflict/Armed-Conflict-Location-and-Event-Data/
Warning System Network, FEWS-NET\textsuperscript{35}, the U.S. National Oceanic and Atmospheric Administration (NOAA) Coastal Services hazards assessment toolkits\textsuperscript{36} and the United Nations Food and Agriculture Organization’s Global Information and Early Warning System.\textsuperscript{37} Our project would extend these to human-initiated political disasters.

Fifth, there is little publicly available software for the methods we are proposing to use to forecast conflict (cooperation). We will continue the practice begun with the development of the MSBVAR and TABARI software packages of making our code available and open source. The expanded availability of software and data that flows from this project will benefit applied international relations, time series analysis, econometrics, etc.

Finally, the methods we are developing could be extended to many domains beyond political analysis. The event/network methods we develop for the study of political conflict could quite easily, for example, be used for the study of economic competition or small-group dynamics. That synergy will not occur if tool development such as this remains only behind the screen of security classification. Further, the application of BVAR, MS-BVAR, and other time series models will be advanced. As we have already done with developing modified priors for these models, we believe that there will be novel identification methods for the states and phases that can be used by researchers in other fields.

\textsuperscript{35}\url{http://www.fews.net}

\textsuperscript{36}\url{http://www.csc.noaa.gov/bins/resources/hazards.html}

\textsuperscript{37}\url{http://www.fao.org/GIEWS/english/index.htm}
References

Achen, Christopher H. 2006a Institutional realism and bargaining models. In The European

Union Decides, ed. Robert Thomson, Frans N. Stokman, Christopher H. Achen, and


Achen, Christopher H. 2006b. Evaluating political decision-making models. In The European

Union Decides, ed. Robert Thomson, Frans N. Stokman, Christopher H. Achen and


Applications International Corporation.

Beck, Nathaniel, Gary King and Langche Zeng. 2000. “Improving Quantitative Studies of


Bennett, P.G. and P. McQuade. 1996. “Experimental Dramas: Prototyping a Multiuser


to Policy. New York, NY: St. Martin’s Press.


In Being Useful: Policy Relevance and International Relations Theory, ed. N. Miroslav J.


Figure captions

Figure 1: Data generation, model specification and forecasting

Figure 2: Regime Probabilities: Monthly Levant Material Conflict Data, 1996-2010. Black (red) line is regime one (two).

Figure 3: Material conflict forecasts for Israeli and Palestinian dyads based on CAMEO material conflict data, 2010:1-2010:12. Black / solid (red / dashed) lines are the median BVAR (MS-BVAR) forecasts with 68% and 50% credible intervals.