Chapter Three
The Statistical Characteristics of Event Data

Most political science data derive either from random samples, or from the measurement of entire populations. Survey research projects, which certainly provide the largest amount of data in political science generally, rely on carefully-designed random sampling techniques that have been refined over the past century. The sampling properties of these surveys can be analyzed using probability theory and have well-understood characteristics. Sampling is used not just to provide information on public opinion, but also data that are treated as if they were population characteristics such as unemployment and inflation rates.

At the opposite end of the spectrum one finds data sets that aspire to provide data on the entire population, rather than a sample. When that population is very large, as with the United States census, there may still be elements of sampling, but in many instances this is almost entirely absent. The Correlates of War project, for example, in all likelihood has identified all historical instances of "war" during the 1815-1990 period that involve nation-states recognized by the European powers. Definitional differences might remain, and some of the variables such as GNP or battle casualties may be poorly *measured*, but the population itself has been identified. Data involving international trade, treaties, borders, government type, and crises between major states also involve little if any sampling.

Event data fall between these two categories: they are neither comprehensive, nor a random sample. The news media report only a tiny fraction of the events that occur in any given day, but they do not report these events at random. Reporting is affected by factor such as the type of event (a war is more likely to be reported than a street crime), by the novelty of the event (events at the outbreak of a war are more likely to be reported than the continuation), by the

actors involved (major states, and countries with high GDP, get far more coverage than small or poor states), and by the accident of a reporter being at a place when something interesting happens, and the editors putting that report into a newspaper or onto a wire service.

This chapter will explore several issues involving the statistical characteristics of event data. We will first look at the definition of what constitute an “event”, an issue on which the existing literature shows less than complete consensus. We will then systematically go through the sources of error and uncertainty that come between “events on the ground” and a coded event data set. Coding mistakes are one source of these errors, but by no means the only, nor necessarily the greatest, source of error. Based on this model, we examine the consequences of alternative means of coding and aggregating events, showing that in at least some circumstances, analyses are relatively insensitive to specific coding schemes and weighting methods. Finally, we look at how the dynamic event data approach to early warning, which will be explored in the remaining chapters, compares to the alternative approach of structural early warning.

3.1. Defining an "event"

Despite the widespread use of event data, there is no single universally accepted definition of what constitutes an "event." This section will first survey some existing definitions, then propose an alternative.

The early event data projects provided relatively succinct definitions. According to Burgess and Lawton (1976:6), for example, "events data is the term that has been coined to refer to words and deeds — i.e. verbal and physical actions and reactions — that international actors (such as states, national elites, intergovernmental organizations and NGOs) direct toward their domestic or external environments." Azar and Ben-Dak define an event as:

> some activity undertaken by an international actor (a nation-state, a major subunit of a nation-state, an international organization) . . . at a specific time and which is directed toward another actor for the purposes of conveying interest (even non-interest) in some issue. Thus an event involves (1) an *actor*, (2) a *target*, (3) a *time* period, (4) an *activity*, and (5) an *issue* about which the activity revolves. (1975:1; quoted in Laurance, 1990:112)

The COPDAB project designates events as:

> occurrences between nations which are distinct enough from the constant flow of "transactions" (trade, mail flow, travel and so on) to stand out against this background as "reportable" or "newsworthy." Thus, to qualify as an "event," an occurrence has to be actually reported in some reputable and available public source. (Azar, 1980:146; see also Davies, 1991:3)

The WEIS codebook does not contain a clear description of an event, but an early paper by McClelland provides the following definition:

> Event-interaction is meant to refer to something very discrete and simple — to the veritable building blocks of international politics, according to my conception. The content of diplomatic history is made up, in large measure, of event-interactions. They are the specific elements of streams of exchange between nations. Here are a few examples for hypothetical Nations A and B: Nation A proposes a trade negotiation, Nation B rejects the proposal, Nation A accuses B of hostile intentions, Nation B denies the accusation, Nation B deploys troops along a disputed boundary, Nation A requests that the troops be withdrawn, . . . Each act undertaken by each actor as in the illustration is regarded as an event-interaction. (1967:8)

In a similar fashion, the manual for the BCOW data set (Leng, 1987) does not provide an explicit definition of an event but its dense set of verb-oriented event categories implicitly describes the concept.

There are at least two significant problems with these older definitions from the standpoint of rigorously delineating event data. First, they refer to activities of international actors that an analyst or coder almost never observes. An analyst observes the *report* of an activity. This difference may seem subtle but it is important; it is the same as the difference between an attitude or opinion — unobservable mental states — and the answer to a questionnaire, which is observable.

Second, none of the definitions explicitly indicates what constitutes an "activity," "transaction," or "action." In some cases (e.g., BCOW), the codebook covers this quite thoroughly. But in other instances — most importantly, the ICPSR's COPDAB and WEIS codebooks — the coder is given only a set of English verbs, verb phrases, and noun phrases from

which to infer the underlying coding rules. Through the use of the phrase "direct towards," some definitions (Burgess and Lawton, 1972; Azar and Ben-Dak, 1975) also implicitly require an assessment of motive on the part of the initiator of the event. This leaves considerable room for ambiguity in interpretation. Lest this seem like a trivial point, one major source of ambiguity in event coding arises from policy statements with no explicit target audience, e.g., "Iraq announced it was raising its oil output beyond OPEC quota levels." Some coding systems do not consider this an event because of the absence of a target; others consider it an event with "the world" as the target.

At the risk of definition proliferation, we suggest the following formal definition of an event:

> An *event* is an interaction, associated with a specific point in time, that can be described in a natural language sentence that has as its subject and object an element of a set of *actors* and as its verb an element of a set of *actions*, the contents of which are transitive verbs.[1]

When applied to a specific data source, replace the words "can be" with "is." This definition encompasses most of what is currently considered to be event data but, unlike the existing definitions, it can be unambiguously implemented. The key elements are: time, natural language, actors, and actions.

Time. All event data record a time or period of time when an interaction occurred. The most common unit used is the day, although in some instances (for example, studying crisis negotiation) a finer unit such as the hour might be appropriate. Most analyses of event data aggregate to either months or years.

Natural language. Event data coders do not observe events; they observe reports of events presented in a natural language such as Arabic, Chinese, English, German, or Hindi. Empirically, events can only be defined with respect to a human language or set of languages. The event

[1] As a matter of record, while this definition appears to be designed to justify machine coding *a priori*, it was actually a consequence of some years of experiments with machine coding and the gradual realization that interactions that do not meet this criteria are likely to be ambiguous to humans as well as machines.

coding exercise converts natural language into nominal data that can be analyzed using formal methods.

Actors. Any model of political activity will be specific to certain persons, organizations, and places, all of which are specified by noun phrases in the language or languages used in the source. Many of these phrases may be synonyms referring to the same actor. For example, in U.S. political discourse, the former Union of Soviet Socialist Republics might variously be called *the USSR*, *the Soviets*, *the Soviet Union*, or by the name of its current leader (e.g., *Khrushchev*, *Brezhnev*). For most international relations research, the actors will be political entities; however, the definition is flexible and allows for the possibility of human-nature interactions if this is important for the research questions being studied.

Actions. An event coding scheme deals only with certain interactions between actors. All of these interactions can be described by transitive verbs; for example, *apologize, met with, endorsed*, *promise*, *accuse*, *threaten,* or *attack.* (Transitive verbs are those that can take a direct object and indirect object. For some events, the second actor is the direct object of the sentence ("Syria accused Israel . . ."); in other cases, it is the indirect object ("Saudi Arabia promised economic aid to Syria"). As with the nouns, multiple verbs might signify the same category of behavior, either because the words are synonyms within the language (e.g., *grant, bestow*, *contribute*, *donate*, *fund*, *present, provide*) or because the behaviors, although linguistically distinct, are politically equivalent, a characteristic that Most and Starr (1989: chapter 5) refer to as "foreign policy substitutability." These equivalence sets will vary with the specific problem or the theoretical approach and in large part determine the validity of a particular coding scheme.

One objection that could be raised to the approach of focusing on a standard list of verbs and actors is that in some unusual situations, the exact wording of an official speech will be an important point in negotiations. For example, President Richard Nixon's use of the term "People's Republic of China" rather than "Red China" signaled the beginning of the USA-PRC détente in 1970. Similarly, the United States insisted on some very precise wording in the

statements by Yasir Arafat in 1988 prior to the re-establishment of U.S. diplomatic contact with the PLO. However even a negotiation involving specific language generates a large number of standard event verbs such as *accept, reject, inform*, *confer*, and *rebuke*. Precise language may be the *object* of a negotiation, but rarely is it part of the *process* of negotiation, and even less does it affect the *reporting* of the political behavior.

This does not rule out the possibility of focusing on specific phrases that are used with specialized intent in diplomatic communication. For example, the phrase "frank discussion," when used by the U.S. State Department, almost always means that the discussions involved substantial disagreement. To the extent that these words are used consistently, they can be coded as signaling disagreement: The phrase "had frank discussions" might code to a WEIS 111 (*Turn down proposal*) rather than a WEIS 025 (*Explain policy*), despite the latter interpretation being closer to the dictionary meaning of the phrase.

The language vary dramatically between sources. Pro-government and anti-government sources might report the same occurrence using quite different words:

> Terrorists slaughtered innocent civilians in the town of Ochos Rios before being driven off by government troops.
>
> Liberation forces battled occupying forces in the town of Ochos Rios, causing several casualties before retreating.

Rhetorical flourishes — *terrorists* versus *liberation forces*, *slaughtered* versus *causing*, *driven off* versus *retreated* — often say more about the source than they do about the political behavior. In many cases, such rhetoric is used in a highly styled fashion that can actually improve coding accuracy, once the patterns are identified and vocabulary lists customized. From either phrasing of this event one can infer that things were not very quiet in Ochos Rios and that people were killed. With multiple reports of this type, one can reasonably infer some sort of insurgency.

Similarly, the extensive popular literature on differences in U.S. and Japanese business negotiating styles (e.g., Hall and Hall, 1987) points out that Japanese negotiators rarely explicitly reject a proposal (e.g., "We will consider your proposal." means "No.") whereas U.S. negotiators

tend to use extremely strong language (e.g., "We can't possibly work with that." means "We're getting close to an agreement."). To the extent that such phrases are used consistently—and unless they are used consistently they have no information value under any circumstances—the phrases can be accommodated in any systematic coding framework, although the lists of verbs may need to vary with the source of the statement. Because those lists are explicit and reproducible, the source of errors due to misinterpretation can be isolated and identified. In some cases, the terms used may be so ambiguous as to preclude coding altogether. However, if the source itself is ambiguous, human coding will also be unreliable and the appropriate action is to find an alternative source of information.[2]

3.2. Stochastic Elements in the Measurement of Event Data

As noted in the introduction to this chapter, event data are neither a random sample nor a description of a complete population. In the process of going from “events on the ground” to a data set that can be analyzed using statistical techniques, a number of different sources of stochastic variation are introduced. These are summarized in Figure 3.1, and each source will be discussed in detail in this section. The section will conclude with some suggestions on how some of these errors could be corrected, at least for the purpose of analyzing aggregate data.

3.2.1. Noise in Event Data

Any event stream contains noise in the form of random events that appear to be endogenous to the process being studied but which in fact have been generated by other processes or are irrelevant to the model. Another source of noise is when a source (e.g. Reuters)—faced with a slow news day—reports events that it normally would not report. The distinction between

[2] For example, one can imagine a situation where an area had a substantial amount of economically motivated banditry as well as politically motivated guerrilla activity but where the government press referred to all violent activity as the product of "bandits" as a way to delegitimize the guerrillas.

endogenous and exogenous events depends on the theory being studied. What is noise to one model may be the signal of another, just as the variation in growth of a soybean plant due to insect damage is noise in a study of fertilizers but signal to a study of insecticides.

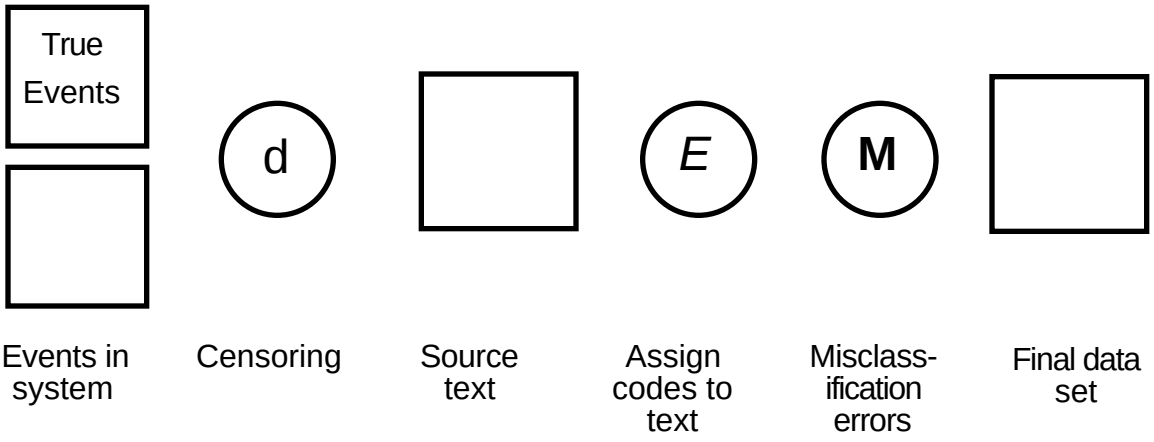

Figure 3.1. The Process of Generating Event Data

In correlational studies, noise is usually assumed to be normally distributed. In an event data study, the default model for noise would be statistical independence and a Poisson distribution since the Poisson is the only temporal probability distribution that is memory-free. Ideally, one could ascertain the Poisson intensity parameter for the distribution of noise affecting each event code and use that information in statistical studies, just as the mean and variance of normally distributed errors are used in correlational studies. As in correlational studies, the actual noise will often not be distributed according to the assumptions of the model, and might exhibit non-Poisson behavior such as cyclicity or statistical interdependence.

3.2.2. Censoring

Censoring means that an event occurs in the system and does not appear in the data. The term is meant in the statistical, rather than political, sense: while overt censoring of information is certainly a factor in event data, far more problematic are the editing and coverage biases

introduced in journalistic and historical sources; these have been discussed exhaustively in the event data literature.

Censoring occurs nonuniformly. Any event data set has a vector $\underline{d}_C$ (i.e. a vector indexed on the set of classification codes **C**)[3] that is the ratio of the frequency of codes in the observed data set to their frequency in an ideal data set where all events occurring in the system were reported. These ratios can be quite low—it is unlikely that existing data sets capture more than a few percent of all political activity except for extreme events such as the outbreak of war. To the extent that some events are more important than others in determining international behavior, censoring is probably inversely proportional to importance: the more important an event, the more likely it will be reported.

The converse of censoring is "disinformation": strings in the source text concerning events that did not actually occur. These might be introduced deliberately—for example the deception campaigns that preceded the US ground offensive against Iraq in 1991—but they are more commonly generated by rumors and second-hand information. While disinformation from deception and rumor is probably not a major component of event data, it is more problematic than pure noise because it is very non-random and is specifically designed to make the system appear as though it is operating under a different process than that actually occurring.

In particular, rumors must convey plausible patterns of human behavior or they will not be voluntarily transmitted. For example, in the week following the June 1989 Tiananmen Square massacre in Beijing, many rumors circulated about pro-democracy military forces preparing to move against the city. These rumors eventually proved to be completely groundless but were credible: when a comparable situation occurred in Romania in December 1989, military units did turn against the government. The nature of rumors and story-telling means that a credible sequence of events is generated; if information is missing from the story, it will be provided in a

[3] In Figure 1, $\underline{d}$ is placed prior to the source text because that is where most censoring occurs in the real world — events occur but are not reported. For reasons that will be clear momentarily, it is more convenient to index $\underline{d}$ on event codes rather than original events.

fashion that makes sense to the teller and listener, rather than generated randomly. Rumors, as a consequence, may be more likely to fit models of political behavior than will actual events. Note that if regularity did not exist in the international system, disinformation and strategic deception could not exist since the entity being deceived must be able to fit the information to a pre-existing model of international behavior.

3.2.3. Misclassification

Misclassification occurs when the code assigned to a text string does not correspond to the code that was intended when the coding scheme was designed. Any event data *set* has an error matrix **M** where e_{ij} gives the probability of misclassifying v_j as v_i. This error matrix incorporates errors due to both validity and reliability problems.

The coding errors generated by human and machine coding are quite different. Most human coding errors either involve relatively subtle differences in behavior (e.g. deciding whether a negative statement should be coded into WEIS's "comment", "reject" or "protest" categories) or missing the event altogether. As a consequence, in human coding **M** has a block structure—one can arrange the row and columns of **M** is a fashion where non-zero entries are most likely to occur in adjacent cells with the remainder of the matrix being zero: A trade agreement may be confused with a cultural exchange but it is unlikely to be confused with a war.

(The single most common human coding error, however, is a form of censoring rather than misclassification: human coders tend to miss some of the dyadic events in a sentence. For example the simple phrase "Representatives from France and Russia will visit Baghdad and Teheran next week" generates eight WEIS events—four dyadic visits (WEIS 032) and four dyadic hostings (WEIS 033)—and human coders tend to miss some of the combinations. Humans are also more likely than machines to miss events in compound sentences.)

Machine coding, in contrast, will occasionally generate event codes that are completely implausible (uses of force between allies) due to sentences that have unusual grammatical constructions or that use metaphors mistakenly interpreted as events. While these are relatively rare and can usually be avoided with an appropriate filter, the statistical techniques appropriate

for machine-coded data may be different from those appropriate for human-coded data. In machine coding, the structure of **M** generally varies with the event category.

For example, the widespread use of military metaphors to describe conflictual but nonviolent effects (e.g. "At the United Nations, Iraq's ambassador blasted the U.S. policy on continued sanctions" versus "Israeli warplanes blasted suspected Hizballah positions in southern Lebanon") means that the WEIS force code is often erroneously coded. The WEIS accuse and deny categories, in contrast, are almost never incorrectly miscoded, and typically the word used in the news report will be "accuse" or "deny." The major WEIS category promise is usually correctly identified, but the subcategory may by incorrectly specified if the dictionaries do not have contain the appropriate direct object (for example "promised to send emergency aid" is in our dictionaries; "promised to send trained dogs to find earthquake victims" is not).

When machine coding is used, the rate of classification errors can be substantially reduced by the use of simple filters that tell the program to skip sentences that appear too complex or do not have sufficient information to code correctly; this process of "complexity filtering" is discussed in Chapter 2. For example a sentence such as

> Prince Sadruddin Aga Khan, a veteran troubleshooter, was named on Tuesday to oversee all *U.N* humanitarian operations in *Iraq* and *Kuwait*, and on *Iraq's* borders with *Turkey* and *Iran*, where hundreds of thousands of refugees have fled in recent weeks.

contains references to five different actors and at least three verb phrases is much more likely to be incorrectly coded (by a machine or a human) than a simple "Syria accused the United States of..." sentence.

3.2.4. Schematic Error

A final source of classification error is "schematic error." This occurs when the coding system combines two sets of behavior that have distinct natural-language representations—and which should remain distinct for theoretical or analytical reasons—into a single category. The system may also do the opposite—separate two sets of behavior that could be combined—as in WEIS's notoriously overlapping warn and threaten categories.

Event structures allow multiple events to serve the same function. Coding errors within these substitution sets will have no effect on the fit of the structure, since all events within the set are equivalent. The existing coding schemes of COPDAB and WEIS implicitly use a high degree of substitution, since they map many distinct text strings into the same event code. However, the substitution mapping is done at the *coding* stage rather than the *modeling* stage.

If one were dealing with a set of models where the substitution sets were always the same—if a pair of events v_1 and v_2 were found in one substitution set they would always be found in a substitution set whenever one or the other were present—then the problem of determining the details of a coding scheme would be solved. Again, existing event sets implicitly do this already. However, the substitution sets probably vary across models. For example, the event

[move an aircraft carrier from the Mediterranean to the Gulf]

and the event

[move an aircraft carrier from the Pacific Ocean to the Gulf]

are roughly equivalent as far as US relations with Iraq are concerned, but have very different implications for relations between the US and Korea. The specific event structure being analyzed—the *context* of the event—is important.

Schematic error also occurs when certain event types are ignored altogether. WEIS and the international scale of COPDAB were designed for the coding of inter-state events, and implicitly focused on the militarized conflicts of the Cold War. Trade was considered "routine" and not coded, but even trade disputes received little attention. Several categories of politically salient activity that have become very important in the late- and post-Cold War period—for example refugee flows and human rights violations—do not have separate codes, nor, as the IDEA project has pointed out, are several types of general political activity such as the adjudication of a dispute.

3.2.5. Statistical Corrections for Coding and Censoring Errors

The overall error structure of an event data set can be specified by combining the elements of noise, censoring and misclassification. Let $\underline{r}$ be the true frequency of the codes—the frequency generated from an ideal uncensored data source coded using *E* without misclassification errors. Let $\underline{n}$ be the frequency of the noise (both random noise and disinformation); let

$$\mathbf{D} = \text{diag}(\underline{d}_C)$$

that is, the matrix with the elements of $\underline{d}_C$ on the diagonal. Then $\underline{x}$, the observed frequency of events in the data set, is

$$\underline{x} = \mathbf{MD}\ (\underline{r} + \underline{n})$$

Using this relationship, it is possible to make several statistical corrections to an event set if one is interested only in the aggregate *frequency* of events; frequency is the key concern is most correlational and descriptive studies.

Assume that the misclassification matrix **M** can be estimated and let $\underline{t} = (\underline{r} + \underline{n})$ be the true event frequency vector. If there is no censoring, then the observed frequency vector $\underline{x}$ is simply

$$\underline{x} = \mathbf{M}\,\underline{t}$$

One can correct for misclassification and get an improved estimate of $\underline{t}$ by adjusting $\underline{x}$ using

$$\underline{x}^* = \mathbf{M}^{-1}\underline{x}$$

Under certain circumstances, **M** is very straightforward to compute. Suppose one has two coding systems, one very slow but accurate (e.g. coding by a principal investigator or well-trained and well-motivated graduate student coders), and the other fast but less accurate (e.g. machine coding or poorly trained, supervised or motivated work study students). Assuming that both coding processes are consistent, then **M** can be estimated by comparing two coding results on a suitably large and representative set of texts.

If the censoring vector $\underline{d}$ were also known, the correction can be extended further:

$$\underline{x}^* = (\mathbf{MD})^{-1}\underline{x}$$

$\underline{d}$ is less likely to be known with any degree of confidence than **M**, although efforts could be made to approximate it by comparing multiple sources.

The standard technique for estimating the size of an unknown population is to sample the population twice and compare the number of cases captured in both samples[4]. Let N be the true population size, n_1 and n_2 the sizes of two independent random samples from N, and m the number of cases that occur in both random samples. The probability of a case being in sample 1 is $p_1 = n_1 N$; probability of a case being in sample 2 is $p_2 = n_2 N$, so

$$m = p_1 p_2 N \qquad N = \frac{n_1 n_2}{m}$$

Once N is known, the number of cases being censored can be estimated by comparing the number of events found in a source to the estimated population size N.

The weakness in this approach is the requirement that the two samples be random. All text sources of event data are biased to report certain events while ignoring others, rather than randomly sampling from all possible events. However, comparing two sources that are attempting to provide equivalent coverage—for example the *New York Times* and the *Los Angeles Times*—would provide a rough estimate of the censoring probabilities. This technique could also be used to ascertain which event categories are more frequently censored.

3.3. Interval vs. Discrete Approaches to Event Data Analysis

As noted in Chapter 1, Charles McClelland's early work assumed that event data would be used for the systematic study of *sequences* of behaviors. McClelland (1970:6) notes, in the quotation that opens Chapter 1, that event data could form a bridge between the then-prevalent general systems theories of international behavior, and the understanding of political behavior through textual history: “a starting point [for event data research] is provided as readily by the ordering principle of classical diplomatic history as by the basic concepts of general system analysis.”

[4] This method is typically used to estimate fish or insect populations; it has also been used to estimate the undercount in the U.S. Census.

In McClelland's assessment, however, this transition away from the general systems approach—which was rooted in continuous-time dynamics and interval-level variables—failed. After some years of work with event data focusing on several crises, he concluded:

> It proved relatively easy to discern event patterns and sequences intuitively. We found we could follow the successions of action and response in flow diagram form. Stages of crisis and the linkage of event types to temporary *status quo* situations also were amenable to investigation. We were defeated, however, in the attempt to categorize and measure event sequences. This was an early expectation that was disappointed by the data which showed too few significant sequences to support quantitative or systematic treatment. (McClelland, 1970:33)

As a consequence of this failure, McClelland's "World News Index" project, published in the mid-1970s, used interval-level variables in its measures. With the hindsight of two decades, the failure of a discrete event approach appears due to a paucity of data and processing capability. McClelland writes of analyzing hundreds or at most thousands of events; a contemporary event data researcher has available tens of thousands of events and computer power sufficient to work with millions.

After this early definition of international politics as event sequences, the field of quantitative IR moved rapidly to analyzing events with interval-level techniques. This change was probably due to the general shift in quantitative IR in the late 1960s away from historical approaches towards theories based on the model of the physical sciences and economics. By the 1970s a Kuhnian split was underway in international relations with the traditionalist and behavioralist camps proudly speaking totally different languages, whereas when McClelland's 1961 *World Politics* article was written, this split was not apparent. For example, Rummel (1972) proposes a science of international politics similar to meteorology, using interval-level metaphors such as field theory and interval-level techniques such as factor analysis. Azar, while using WEIS as the model for COPDAB, abandoned McClelland's nominal categories in favor of an interval-level measure and approached coding as a scaling problem. Azar and Sloan (1975) consists entirely of interval-level data and Azar emphasizes

> quantitative aggregations, called here 'analytic data', [which] are summaries of the weighted frequencies of interactions. They describe the amount of conflict or cooperation exchanged between or within nation-states over some unit of time. (Azar 1980:150)

This conversion of discrete entities to interval-level data is somewhat puzzling from a statistical standpoint. The reason probably is due as much to paradigmatic developments in quantitative international relations as in the nature of the data. During the 1970s, data-based studies of international behavior saw the ascendancy of correlational analyses, particularly regression. The mathematics behind these techniques had been fully developed by econometricians and could be easily applied to international relations data using SPSS and other statistical packages; comparable tools were not available for sequence analysis. The successful formal theories were continuous-variable models such as the Richardson arms race model and DYNAMO-like global models. Rational choice models preserved some discrete variables, particularly in game theory, but even these models used continuous variables in expected utility calculations[5].

The emphasis on crisis, initiated by McClelland and expanded in a large event-based crisis management literature (e.g. Hoople, Andriole and Freedy 1984, Azar et al. 1977) also contributed to the emphasis on interval-level variables. Implicit in most of the crisis models is either a simple distinction between crisis and non-crisis or a unidimensional ordinal set of "steps" to crisis (e.g. Hoople, 1984). Crisis forecasting is reduced to a problem of monitoring some activity—usually some aggregated measure—to ascertain when the system is going to change states. The crisis literature also frequently emphases the concept of the "intensity" of events, another interval measure.

The advantage of this approach is that a wide variety of methods are readily available. The clear disadvantage is that the process of reducing behavior to a single dimension through scaling

[5] Another factor favoring continuous formulations may be the ease with which one can draw a line representing a continuous variable and call it a "pattern" (for example Brody, 1972) whereas discrete sequences are more difficult to visualize and explain (McClelland, 1961).

loses a great deal of information and introduces a large number of free parameters (the issue of the choice of weights will be discussed in detail below). For example in principle (although almost never in practice), a month characterized by a large amount of conflict in the first two weeks (negative numbers on most scales), followed by a large amount of reconciliation in the last two weeks (positive numbers) could aggregate to value close to zero, which is the same value that would occur in a month where nothing happened.

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

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

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

Despite these disadvantages, one of the most common techniques used in event data analysis is to aggregate events over time—typically by week, month or year—using a numerical scale. This changes the data set from an irregular, nominal-level time series to a regular, interval-level time series. An event data scale assigns a numerical value to each event category found in the

coding scheme. Table 3.1 shows subsets of the Azar-Sloan (1976) and Goldstein (1992) scales, which apply to the COPDAB and WEIS coding schemes respectively; the full scales can be found in the event coding appendices of this volume.

Table 3.1. Examples of Event Scales

Azar-Sloan Scale

COPDAB Category (Azar 1982)	COPDAB Code	Scale value
Military, economic and strategic support	3	31
Mild verbal support; exchanges of minor officials	7	6
Diplomatic-economic hostile actions	11	29
Full scale war	15	102

Goldstein Scale

WEIS Category	WEIS Code	Scale value
Praise	41	3.4
Promise Policy Support	51	4.5
Extend Military Aid	72	8.3
Criticize	121	-2.2
Ultimatum	174	-6.9
Military Engagement	223	-10.0

The Azar-Sloan and Goldstein scales are the systems most commonly used in the literature, but other systems exist; their development is summarized in Goldstein (1992:373-374). The Goldstein scale has been used in a number of recent studies, including Huxtable and Pevehouse 1996, Goldstein and Pevehouse 1997; Schrodt and Gerner 1997, 1998, Bond et al. 1997; Kinsella 1995, 1998; Reuveny and Kang 1996a, 1996b. Prior to this, Vincent's (1990) scale for aggregating WEIS events was used in a number of studies, most notably Goldstein and Freeman's (1990) book-length study of superpower interactions.

All of these scales have been constructed by querying panels of experts about the relative intensity of various event categories. In most cases, the weights have been assigned on a single

cooperation-conflict dimension. Beyond this, there is no consistency in the scaling—for example the Goldstein weights are roughly proportional to the logarithm of the Azar-Sloan weights.

The weighting schemes appear to work fairly well. While in principle the uni-dimensional cooperation-to-conflict scaling should be problematic—for example the USA-Canada or USA-Japan relationships are characterized by high levels of both cooperation and political conflict—in practice this hasn't prevented the scaled data from being used successfully in a variety of studies. This may be due in part to the fact that event data have been primarily employed to study highly conflictual situations such as the Cold War (Ashley 1980; Goldstein and Freeman 1990; Dixon 1986) and the Middle East (Azar 1972; Azar et al. 1979; Schrodt and Gerner 1997, 1998) where "cooperation" is largely expressed as a reduction of conflict.

Nonetheless, there are a couple of clear problems with scaling. First, aggregating events is controversial: the "folk criticism"[6] of the Azar-Sloan scale is "3 riots equals a nuclear war." This debate goes back to the earliest event data discussions (e.g. Azar and Ben-Dak 1975; Azar, Brody and McClelland 1972) and has continued over time: see exchanges between Howell (1983) and McClelland (1983) or Vincent (1990) and Dixon (1990). Second, the assignment of weights by panels of experts is arbitrary, atheoretical and detached from any specific empirical context. For example, why should the same set of weights should apply to a dispute such as Israel-Lebanon, where military exchanges are very common, and USA-USSR, where military exchanges were virtually nonexistent? Finally, it is unclear how much effect the choice of a particular weighting scheme has on the results of the study—minor differences in weights could lead to major differences in results.

The next two subsections report the results of three sets of experiments we have done on the effects of changing the scaling and aggregation in event data. These experiments were done in the context of substantive research, rather than as basic experiments on event data in general, but

[6] As in "folk theorem": we've heard this phrase many times over the years but have no idea who originated it. The Azar-Sloan value for "inciting of riots" (COPDAB category 12) is equal to 44; "full-scale war" is 102.

consistently show that event data seem to be relatively *insensitive* to these changes. This would help explain why coding and scaling systems such as the Goldstein scale and the WEIS categories have been effective for political analysis despite their rather *ad hoc* nature, and also provides some indication of the relative importance of censoring, coding, and schematic errors discussed in Section 3.1.

3.3.1. The Effects of Simplifying Scales

In this subsection, we will examine the effect of simplified weighting schemes on the delineation of phases of political behavior in the Middle East (1979-1995) is done in Chapter 4, using the Goldstein (1992) scale as the reference point.[7] The weights will be progressively simplified to give decreasing levels of differentiation between the event categories. These simplified scales are shown in Table 3.2; in the analysis the data were reduced to monthly aggregations using six different systems that derive progressively less information from the daily events.. By comparing the results of these various weighting schemes we can evaluate the extent that the results of an analysis are dependent on the choice of a particular weighting scheme.

Table 3.2. Alternative Scales

Goldstein:	Goldstein weights
difference:	cooperative events = 1; conflictual events = -1.
total:	all events = 1.
conflict:	cooperative event = 0; conflictual events = 1.
cooperation:	cooperative event = 1; conflictual events = 0.
report:	1 if any event was reported in the month, 0 otherwise

[7] The details of the analytical method will not be repeated here; these are discussed in the various sections of Chapters 4 and 6.

Table 3.3 reports the effects of these alternative weights on a discriminant analysis for the assignment of crisis phase (see Chapter 4, section 4.3). The results are rather striking: There is almost no difference between the weighting systems: all behave almost identically in various measures of the discriminant analysis. In fact, the Goldstein scale is actually the *least* effective weighting system for differentiating between the *a priori* clusters of behavior. The most effective measures seem to be the number of cooperative events, and the total number of events, though the differences in classification accuracy are not large.

Table 3.3. Discriminant Analysis Results

Weighting scheme	%correct(1)	variance explained(2)	canonical corr.(3)	Wilks' λ	signif	# factors
Goldstein	85.6%	76.3%	0.85	0.008	<.001	6
difference	89.7%	74.7%	0.85	0.007	<.001	7
total	94.4%	83.0%	0.93	0.001	<.001	6
conflict	88.2%	76.9%	0.86	0.007	<.001	6
cooperation	92.3%	82.2%	0.91	0.002	<.001	7
report	89.2%	73.6%	0.87	0.008	<.001	7
random date(4)	61.0%	69.5%	0.66	0.131	.37	0
random dyad(5)	57.4%	68.8%	0.67	0.119	.18	0

Notes:

1. Percentage of the cases classified correctly.

2. Variance explained by the first three factors of the discriminant analysis.

3. Canonical correlation for the first discriminant function. The canonical correlation is equivalent to the Pearson product moment of the value of the discriminant function regressed on a set of dummy variables representing each of the categories in the data set

4. This set used the Goldstein weights but randomized the order of the dates in the data set. (4)(5)4. is, the data had no correspondence to the actual political phases, but the5. isd the information in eac to the patterns observed in a given month.

5. This set randomly ordered the information in each dyads so that the marginal characteristics of the data set remain the same, but the data are randomly order and have no relation to either the actual phase or to each other.

The discriminant analysis on the two random data sets produce quite different results from the analysis of the actual events. The random data have almost a 70% accuracy rate in phase classification, probably because of the contrast between the very high dimension (54) of the variable space and the relatively small number of clusters being classified (7). However, the Wilks' λ measurement correctly adjusts for this—despite the high classification accuracy, λ shows that the none of the discriminant functions are statistically significant.[8]

A second set of tests was done with using the LML clustering measure discussed in Chapter 4, Section 4.5. Figure 3.2 shows the cluster boundaries generated by each of the weighting vectors, using the LML difference level Δ=0.15 (except the Goldstein weight, where Δ=0.30[9]). Table 3.4 shows the number of cases where the boundaries determined by the simplified vectors are within ±3 months of the *a priori* boundaries discussed in Chapter 4, Table 4.1, and the Goldstein boundaries.

These results are generally similar to those found in the discriminant analysis: in general the simplified weights do as well or better than the Goldstein scale in matching the *a priori* boundaries, and there is little difference in the results obtained with the various simplified weighting schemes. The "total" vector—where all event types are weighted equally—also has a high level of correspondence with the divisions found by the other methods; the "difference" variation on this does as well in matching the *a priori* transitions but (unsurprisingly) corresponds more closely to the Goldstein divisions than does the constant vector. In contrast to the discriminant analysis, the vector with the least correspondence to the other transitions

[8] This is an instructive example of why inferential statistical methods, rather than gross classification accuracy, is needed to evaluate these models.

[9] The higher value of Δ in the Goldstein weights is a result of those weights being between -10.0 and 10.0, whereas the remaining weights are between -1 and 1. Because the high number of zero values in the vectors more or less forces the regression line through zero, the data are effectively not invariant to a change of scale despite the use of a correlation measure, and the higher variance of the Goldstein scale leads to higher correlation values.

considers only the cooperative events; in particular this has almost no correspondence with the *a priori* transitions.

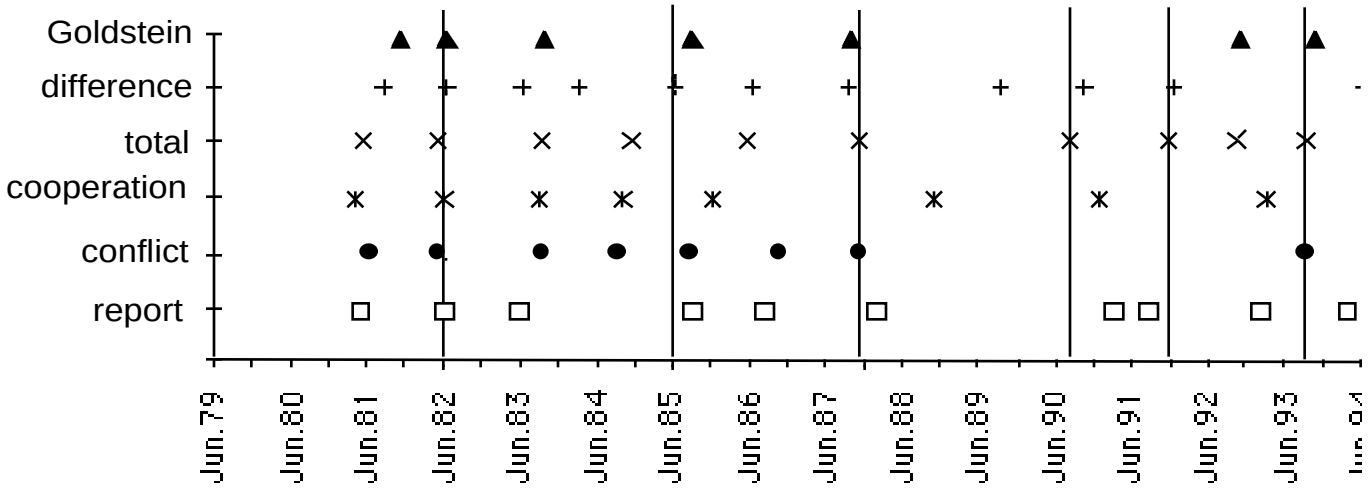

Figure 3.2. Cluster boundaries under various weighting systems

Table 3.4. Correspondence between the boundaries generated by default vectors and other techniques

	Genetic (N=12)	Goldstein (N=11)	*a priori* (N=6)
difference	6	8	5
total	8	6	5
cooperation	5	3	1
conflict	7	6	4
report	9	5	4
Goldstein	9	--	4

3.3.2. Optimizing the Scale

The Goldstein scale was not designed for the purpose of clustering political behavior, and it is possible that some other set of weights might be superior for this purpose. To assess this possibility, optimal weights for the LML clustering procedure were estimated using a genetic algorithm (GA)—described in Appendix 3.A—that maximized the following clustering measure:

$$F^C = \frac{\text{average distance between adjacent clusters}}{\text{average distance within clusters}}$$

where "distance" is defined by the correlation metric and the "average distance" is calculated as the average distance between points. F^C is similar to the F-ratio maximized in discriminant analysis except that only the distance between adjacent clusters is considered and measure uses the total distance between points, rather than group variances.

The GA works reliably and most of the experiments produced similar sets of cluster boundaries. The boundaries are, however, dependent on the value of Δ in the LML criterion: Higher values of Δ consistently produce higher F^C values, but fewer clusters, because of the stricter threshold for establishing a new cluster.

The results of a number of different experiments with the genetic algorithm are shown in Figure 3.3. The thick lines above the X-axis are a histogram of the cluster boundaries identified by 100 GA runs for Δ=0.15; F^C in these runs ranged from 1.70 to 1.51 compared to F^C = 1.08 for the Goldstein weights.[10] The solid squares on or near some of these lines show the number of cluster boundaries found for 25 GA runs with Δ=0.25. The solid triangles just below the X-axis show the location of the cluster boundaries generated by the Goldstein weights for Δ=0.30. As indicated in Table 2, the relationship between the simplified weighting schemes and the divisions

[10] For Δ=0.25 the range of F^C for 25 GA experiments is 2.35 to 1.90; for the Goldstein scale F^C = 1.30. By comparison, the F^C values for the other weighting systems are difference = 1.18; total = 1.46; cooperation = 1.26; conflict = 1.29; report = 1.42.

found by the GA are generally similar to the relationship between those divisions and the Goldstein divisions.

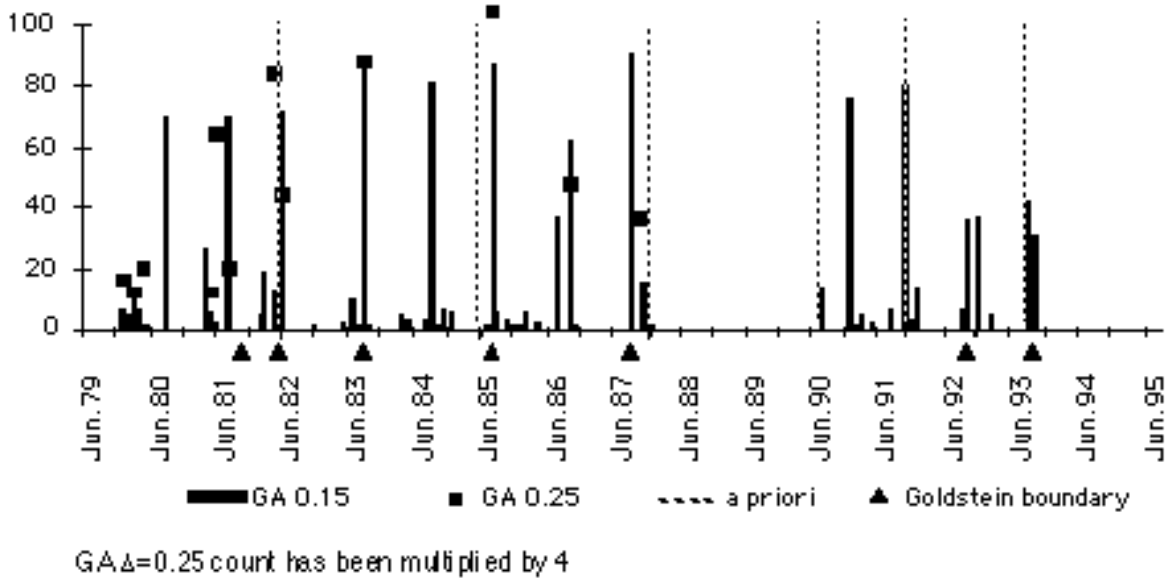

Figure 3.3. Cluster boundaries under genetic algorithm optimization

While the cluster boundaries determined by the GA were quite consistent across the various experiments this was not true of the scale weights generating those boundaries: in fact there was no consistent patterns whatsoever in those values. The distribution of the 4,950 correlations between the weight vectors generated by the experiments is generally Normal with a mean of 0.077 and standard deviation of 0.307; the number of significant correlations do not exceed the number expected by chance. The average correlation of the GA weights with the Goldstein weights is r=-0.008 with a standard deviation of 0.188, and there is virtually no difference in the average correlation within the best-fitting 50% of the vectors (r=-0.006) and within the remaining 50% (r=-0.009).

Examination of the intermediate results produced by the GA showed that despite this diversity, the GA is working correctly to generate and select vectors that increase the value of F^c; there are simply a lot of different ways to do this. The lack of convergence of the weight vectors is probably due to a problem comparable to the effect that collinearity in a linear model has on

increasing the variance of the estimates of the parameters of a regression equations: Because the correlation distance is invariant with respect to a linear transformation of the weights, very different sets of weights can be used to produce essentially the same distances. This is consistent with the fact that there is relatively little variance in the delineation of the clusters but huge variance in the values of the weights producing those clusters. Consequently, producing an "ideal" set of weights via an estimation procedure—expert panel, genetic algorithm, or otherwise—is probably a hopeless task.

A genetic algorithm was also used to maximize the separation of the clusters, given the *a priori* cluster boundaries in Table 4.1 in Chapter 4. The purpose of this experiment was to see whether it would be possible to find a better set of weights than those provided by Goldstein for that set of cluster transitions.[11] Twenty experiments were done, with the GA allowed to run for 128 generations.

The results of this experiment differed substantially from the experiments where both the weights and cluster boundaries are allowed to vary. In particular, there is a significant correlation at the 0.01 level between the weight vectors in about 30% of the cases. None of those weight vectors, however, correlate significantly with the vector of Goldstein weights. Furthermore, the

[11] We also tried to estimate optimal weights in a linear fashion using discriminant analysis, where the independent variables were the WEIS event counts (by 2-digit category) totaled across all of the dyads by month. While this technique removes the information on which dyads were interacting—to differentiate the dyads and event categories would involve 1188 variables and we have only 208 data points—this method could provide a rough estimate for the event weightings appropriate for the full system.

The results were generally disappointing. The classification accuracy with all variables was only 73%, and the first three discriminant functions explained only 75% of the variance. There was no discernible pattern to the weights or functions. When stepwise discriminant was used the accuracy dropped to 60% but the variables chosen tended to be those with a high density of events: 01 (Yield), 02 (Comment), 03 (Meet), 11 (Reject), 12 (Accuse), 13 (Protest), 21 (Seize) and 22 (Force). The discriminant weights mirrored those of the Goldstein scale to some extent: 01 = 0.65, 02 = 0.70, 03 = 0.23, 11 = -0.48, 12 = -0.09, 13 = 0.22, 21 = -0.32 and 22 = -0.83.

value of F^C for the optimized vectors is 1.58 while the F^C value for the constant vector is 1.54, so the optimization provides very little improvement in the separation of the clusters.

The variance of the weight in many of the optimized vectors is quite small: Half have a standard deviation between 0.5 and 1.0; the remainder have a standard deviation between 5.0 and 7.0 (the Goldstein weights have a standard deviation of 4.34). This bi-modal distribution is entirely a function of whether or not the vector contains both positive and negative weights: the low-variance vectors have only positive weights.

The high correlation between these optimal vectors and the constant vector suggested one additional experiment: computing the distance between points by correlating the frequencies of the 2-digit WEIS events without applying *any* weighting (in principle this method could also be applied to 3-digit categories). The LML measure was computed as before, with the only difference being that the correlation was computed on vectors containing counts of the twenty-two 2-digit WEIS events for each dyad-month, so each correlation used 22 x 56 = 1232 points rather than the 56 points of the vectors containing the Goldstein scores.

Figure 3.4 shows the LML_t curves from this analysis, and Figure 3.5 shows the cluster boundaries. In general, the event count measure produces results quite similar to those of the Goldstein measure—the correlation (r) between the Goldstein LML_t and event count LML_t is 0.63 —particularly in terms of matching the *a priori* cluster boundaries. (The Δ=0.15 threshold is used for the event count boundaries, which in part explains the higher number of clusters for that measure.) Once again this suggests that the clustering method is not strongly dependent on the Goldstein weights, and the frequency of coded events alone is the primary factor that differentiates the major political features of the data. In another experiment that parallels the "report" scale used in subsection 3.3.1, LML curves were computed for a data set that replaced all event counts that were greater than zero with a value of 1—in other words, a set that measured only the presence of events rather than their quantity. This produces credible cluster breaks—for example it correlates at the 0.78 level with the LML curve from the true data set

when the constant vector is used—though unsurprisingly the variation of the curve is attenuated compared to that produced by the actual data.

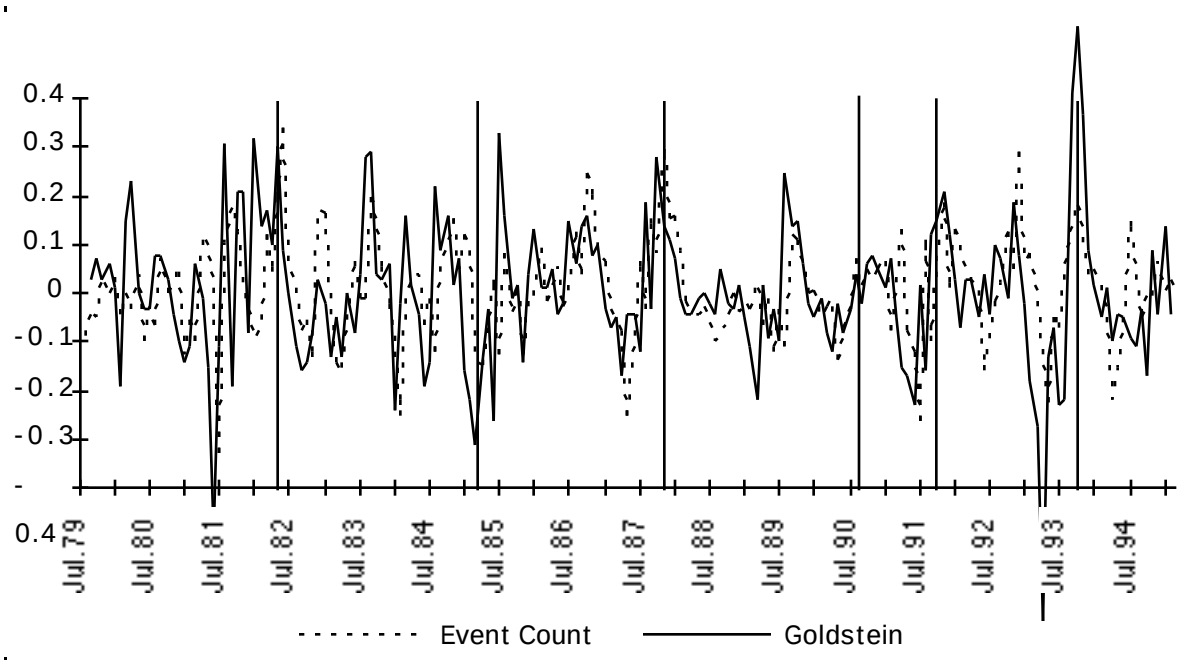

Figure 3.4. Comparison of LML for the event count and Goldstein weights

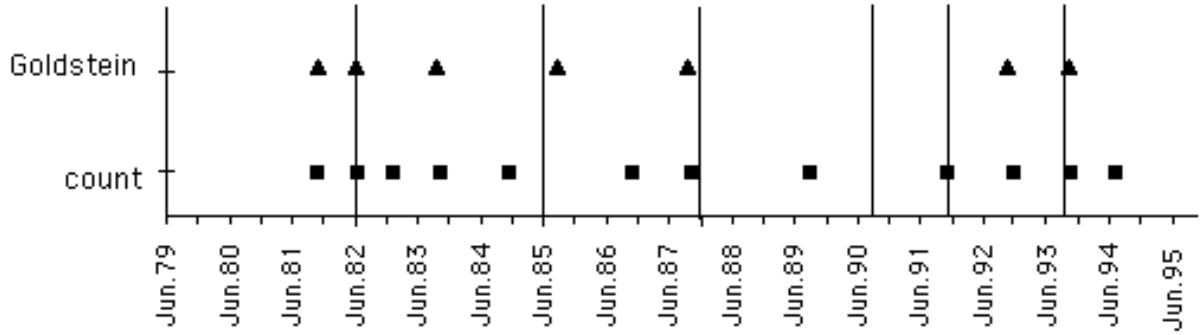

Figure 3.5. Cluster breaks for the event count and Goldstein weights

3.3.3. Simplifying the Nominal Coding Scheme

The final experiment involves simplifying the coding system at the level of categories. This will use the hidden Markov model technique discussed in Chapter 6, which uses disaggregated

nominal variables rather than scaled data. The substantive problem involved forecasting conflict in the former Yugoslavia for the period 1990-1999.

The original motivation for this experiment was, in fact, pragmatic. Hidden Markov models (HMM) involve a large number of parameters and are estimated with a numerical optimization technique that produces a large number of local maxima. The models had originally been estimated using the 22 WEIS "cue categories" (plus a non-event code), but the resulting parameter estimates had come nowhere near converging to a single set of estimates, and were proving very difficult to interpret.

Based on the earlier experiments with the discriminant and LML analyses on the Middle East, it seemed possible that this lack of convergence might be due to the model trying to derive too much detail from the data. For example, if any incidences of violence in the region resulted in events distributed across the WEIS "violence" categories "expel", "seize" and "force", and likewise cooperation involved a combination of meetings, promises, agreements and requests, then attempting to draw subtle distinctions between these categories would be futile. While estimation of an HMM uses a non-linear method, one likely result of this co-occurrence of event types would be an effect similar to that of co-linearity in linear regression: the standard errors of the parameter estimates would increase substantially, and models with very similar degrees of fit might have quite different parameters, as was being found. The models were therefore re-estimated 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 reduced the total number of parameters in the model by about a factor of 5. It 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 3.5; the rows refer to various combinations of optimization weight (P vs. N) estimated at 1, 3 and 6-month forecast lags (see Chapter 6). 32 Monte-Carlo genetic algorithm estimates were done for each set of experimental parameters. For purposes of comparison, Table 3.5a presents the results for the 23 event-category model; Table 3.5b shows the results for the 5 event-category model in the same format.

Table 3.5a. Accuracy for 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

Table 3.5b. 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

In general, the results of the 5-category analysis are comparable to those of the 23-cateogry analysis. In both schemes, the drop-off in accuracy with the increasing forecasting lag is small—about 4% from the 1-month to 6-month forecast lag—though consistently there is a small decrease. The overall accuracy 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 week 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% increases 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 hypothesis that 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.

(Meanwhile, the intuitive insight that motivated this experiment did not prove to be correct: The simplified coding system reduced the variance of the HMM parameter estimates somewhat, but not to the point where these converged to a single, readily-interpreted set of estimates. This could either be due to the collinearity remaining in the data—which looked at the activity of four actors in this subsystem—or to intrinsic indetermancy of HMM parameter estimates.)

3.3.4. Analysis

Based on these experiments, it appears that most of the information that can be used to differentiate political behavior is found in the event counts themselves, rather than in the detailed classification or weighting of events. This could be due to at least two factors.

First, almost all event data contain a large number of time points where no activity has been reported; these non-events are unaffected by any change in the weighting or classification schemes. Around 50% of the dyad-months in the Levant data set have zero values, as are about 50% of the dyad-days in the most intensely-active dyad in the Balkans data set, Serbia=>Bosnia.

Second, the existence of *any* activity in a dyad may be a signal that Reuters reporters or editors think that the political behavior in the dyad is important: this is particularly true with respect to verbal activities where Reuters has often has an option of reporting the activity (in contrast to physical actions such as demonstrations and military clashes, which are reported more frequently).

This in turn means that results of event data analyses are likely to be very robust, rather than being dependent on any particular idiosyncratic choice of scale or coding scheme. The experiments here are by no means definitive, since they have examined only two geographical regions and three analytical techniques. The Levant and Balkans are definitely atypical of dyads in general—they have greater levels of conflict, and are much more thoroughly covered by the news media—but they may be representative of the conflict-prone regions that are most frequently analyzed using event data. Similarily, the correlational methods used in discriminant analysis and the LML clustering are typical of most statistical analysis.[12] It therefore seems unlikely that the development of a complex and highly differentiated scale to be a magic bullet

[12] These tests have all used machine-coded data, so the insensitivity might be due to the errors found in automated 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.

Along a similar line, a journal referee for an earlier version of this research suggested that the results should not be accepted until we replicated them on all dyads in the international system. While we concur that this would be interesting, there are roughly 40,000 such dyads, the optimized weight and HMM experiments each required about 24-hours of computer time, and we were disinclined to spend the next two centuries doing that exercise. Readers might, however, find it useful to do comparable experiments with whatever region and analytical method they are working on.

that will suddenly reveal features of the data that were completely invisible to simpler techniques. While one can not necessarily conclude that "Less is more" in event data, there is also little evidence that "More is more."

The downside of this dependence on event reports is that we know event coverage to be inconsistent across geographical regions and sources. In one of the earliest studies of regional source effects, Doran, Pendley and Antunes (1973) found substantial differences in reports of domestic violence in Central America depending on whether regional or North American sources were used. More recently, Davenport and Galaich (1998) found very substantial differences across geographical regions in reports of human rights violations found in the *New York Times*. When Schrodt, Huxtable and Gerner (1996) compared Reuters-based data on the Levant and West Africa, missing data was more of a problem in the latter region than in the former, and a discriminant analysis using the Goldstein scale classified *a priori* behavioral phrases with only about 75% accuracy in West Africa versus the 90% accuracy in the Levant.

This inconsistency does not necessarily imply that event data cannot be used for political analysis—to the contrary, as increasingly sophisticated techniques are employed in the analysis of event data, the results are becoming ever stronger and more consistent. However, if those results are strongly dependent on the *frequency* of the reports rather than on their *content*, comparisons across geographical regions and across different sources may be more difficult than had been anticipated when the event data exercise began. Any source (or combination of sources) of events—whether a global source such as Reuters, Associated Press or Agence France Presse, a hegemonic source such as the *New York Times* or *Times of London*, or the various capital-city regional sources found in COPDAB—presents only a tiny subset of the "events" that occur in a day. The question for the event data analyst is not whether that subset is comprehensive—it never will be—but whether it is useful for the research task at hand.

Finally, this lack of sensitivity to event weights and categorization has an important implication for the use of machine-coded data. While machine-coding is more consistent over time than human-coded data, machine-coding is less sensitive to nuances of reported political

behavior, and it is possible that those nuances could be very important. This question has, in fact, driven much of the debate about event data coding from the earliest developments of the technique.

Our analysis does not support the conclusion that nuance is important: Because similar results can be obtained from huge differences in the weighting of event categories and high level of aggregation in the nominal coding of events, there is little evidence that subtle differences in the coding of events would have a major difference on statistical measures based on that data. Consequently, inexpensive machine-coded data is likely to contain most of the relevant information that a vastly more expensive human-coded data set would contain. In particular, the most common machine-coding error is confusing the object of an action with an indirect object or a location. Coding programs will not, however, create a actor that is not mentioned in the text. For example, if a series of events involves Israel, Syria, Lebanon and the Palestinians, some actions of Israel towards Syria might be incorrectly coded as applying to Lebanon or the Palestinians., but it would never create an extraneous Egypt=>Jordan event from these texts. If most of the information found in event data comes from "who" rather than "what," relatively simple coding systems that can be efficiently coded with automated techniques will prove quite functional.

3.4. Statistical Approaches to Early Warning

The discussion so far has focused on general characteristics of how event data is collected and coded. In this section, we will consider some issues dealing with how event data is *used* in political analysis. We will consider specifically on the problem of statistical early warning, which has been the focus of most of our work, as well as much of the work that motivated the initial development of event data.

Statistical approaches to early warning can be classified into two broad categories: structural and dynamic.[13] The *structural* category consists of studies that use events (or more typically, a specific category of event such as a civil or international war) as a dependent variable and explain these using a large number of exogenous independent variables. In the domain of domestic instability, this approach is exemplified by the work of Gurr and his associates, most recently in the "State Failure Project" [SFP] (Esty et al. 1995, 1998); Gurr and Lichbach (1986) and Gurr and Harff (1996) provide surveys of these methods more generally. In the field of international instability, the structural approach is illustrated by the work of Bueno de Mesquita and his associates, and more generally by the Correlates of War project; Wayman and Diehl (1994), Gochman and Sabrosky (1990) and Midlarsky (1993) provide general surveys. These approaches have tended to use standard multivariate linear regression models, although recently the research has branched out to other techniques; for example, the SFP uses logistic regression, neural networks and some simple time series methods.

In contrast to the structural approach, in *dynamic* early warning models event data measures are used as both the independent and dependent variables. Most of the event data projects of the late 1970s classified dyads with respect to the likelihood of a crisis based on a set of event-based empirical indicators. For instance, the Early Warning and Monitoring System (EWAMS), developed with funding from the U.S. Defense Advanced Research Projects Agency (DARPA; see Hopple 1984; Laurance 1990), evaluated three WEIS-based indicators (conflict, tension, and uncertainty) to determine an alert status for any dyad. Azar et al. (1977) use a similar approach based on whether behaviors measured with the COPDAB event scale fall outside a range of "normal" interactions for the dyad.

[13] This discussion will not consider the large literature on non-statistical (qualitative) approaches to forecasting Contemporary surveys of qualitative approaches can be found in Rupesinghe & Kuroda (1992), Gurr & Harff (1994), and Adelman & Schmeidl (1995). We also will not deal with the topic of long-range forecasting using formal methods, which is primarily done using simulation. Ward (1985) and Hughes (1993) provide surveys of that literature.

Researchers justify the dynamic approach—which is at odds with most statistical modeling in political science in using only lagged endogenous variables—in three ways. The first rationale is that many of the structural variables that are theoretically important for determining the likelihood of conflict do not change at a rate sufficient for use in an early warning indicator; in fact many are essentially fixed (e.g. ethnic and linguistic heterogeneity; historical frequency of conflict; natural resource base). Data on variables that are changing—for example unemployment rates, economic and population growth rates—are often reported only on an annual basis, and the quality of these reports tends to be low in areas under political stress.

The second justification for the dynamic approach is that it reduces the information required by the model. The data collection effort of the first phase of SFP, for example, measured 75 independent variables (Esty et al 1995:9); this requires a large amount of information from a vast number of sources.[14] In contrast, the event data collections used in dynamic models focus on reported political interactions that can be collected systematically in real-time, which increases the predictive utility of the model.

The final justification for dynamic modeling involves the nature of political events themselves: the approach assumes that the effects of exogenous variables used in the structural models will be reflected in the pattern of events prior to a major change in the political system. As illustrated in Figure 3.6, the dynamic approach effectively uses the lagged values of the events as a substitute for the structural variables.

[14] However, the final models developed in the project found that most of the forecasting power could be accounted for with only three variables—infant mortality, trade openness, and democracy (Esty et al 1998:viii). Phase II of the SFP involves some limited analysis of dynamic variables, and suggests expanding this approach in future studies.

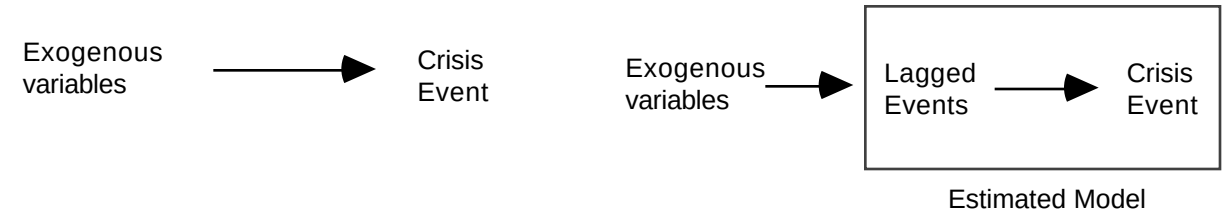

Figure 3.6. Comparison of the structural and dynamic approaches to early warning

To take a concrete illustration, Gurr (1995: 7) notes "We think, for example, that ethnic heterogeneity probably is most significant for state failure when it coincides with lack of democracy and low regime durability." Consequently, the SFP includes measures for those three variables: ethnolinguistic diversity, regime democracy, and regime durability.

A dynamic approach, in contrast, would not measure these aspects of a political system directly, but would instead assume that each would be reflected in the types of events picked up by the international media. The presence of democracy, for instance, would be reflected not only in periodic elections but in a large number of reports of disagreements between the government and the elected opposition. A low level of regime durability would be reflected in coups and attempted coups. To the extent that ethnicity was an important political factor, it would be reflected in ethnically-oriented political rallies, outbreaks of violent ethnic conflict and similar events. A suitably-designed event coding scheme should detect the presence or absence of these events and make the appropriate forecast, without directly measuring the underlying variables.

At a *theoretical* level, the dynamic-events approach accepts the importance of exogenous structural variables: *Ceteris paribus*, countries with a high level of ethnic heterogeneity will have a different propensity for conflict than those with a low level; democracies are likely to be different than autocracies, and so forth. The difference between the early warning approaches is a matter of *measurement*: the structural modeling approach seeks to measure these variables directly, whereas the dynamic approach assumes that to the extent that the variables are relevant

for early warning problems, they can be measured indirectly through the patterns of events the variables generate.

This is an optimistic, but not wholly implausible, assumption. For example, in the Reuters-based data with which we have been working, there is a clear contrast between Israel and Syria with respect to the presence of a democratic opposition and between Lebanon and Egypt with respect to the importance of ethnicity: The ethnic conflict in Lebanon is one of the most conspicuous features of the data set. Our impression is that the increase in democracy in Jordan, and the fluctuations in the Egyptian government's acceptance of a democratic opposition, would also be reflected in the activities reported in Reuters, although we have not attempted to analyze this.

An econometric analogy to this is found in the distinction between "technical" and "fundamental" analysis of stock prices. A fundamental analysis attempts to predict price changes on the basis of underlying factors such as marketing, management, raw material prices, and macroeconomic trends. Technical analysis, in contrast, assumes that these factors will be reflected in the patterns of the movements of the price of a stock (or set of stocks) and therefore analysis of those prices alone will provide sufficient information for forecasting. Fundamental analysis corresponds to the structural approach to modeling political events; technical analysis to the dynamic.

Until relatively recently, technical analysis generally had a bad reputation, consisting as it did largely of statistically-dubious patterns based on small samples, wishful thinking, and gurus whose fortunes were based more on the sale of books than on trading stock. With the increase in computing power in the 1980s, the situation changed, and "programmed trading systems" can now process sufficiently large amounts of information to generate profits (and periodically throw the market into chaos) working solely with information endogenous to the market itself. The increased information processing capacity in the 1990s in contrast to that available in the 1970s may have a similar effect on event data analysis.

Because of the labor-intensive character of human event coding, the primitive statistical methods available at the time, and institutional factors (Daly & Andriole 1980; Andriole &

Hopple 1984; Laurance 1990), the event-based early warning research was largely discontinued during the 1980s. Nonetheless, a small set of dynamic modeling efforts continued. These employed increasingly-advanced econometric time-series methods that modeled an interval-level measure of events as an autoregressive time series with disturbances. Goldstein & Freeman (1990) provide a book-length example of this approach; Ward (1982), Dixon (1986), Ward & Rajmaira (1992), Lebovic (1994) and Goldstein and Pevehouse (1996) illustrate the continued development of dynamic models of events, although these studies generally used event data to explore political interactions rather than for forecasting.

Unfortunately, standard econometric time series methods have only limited utility in the problem of early warning. In general, time series analysis seeks to determine a function

$$y_{t+k} = f(y_t, y_{t-1}..., \mathbf{X}_t, \mathbf{X}_{t-1}...) \quad \text{for some } k>0$$

In English, the fundamental problem of time series is to determine the future values of a variable y given some present and past values of that variable and (possibly) the present and past values of a vector of exogenous variables **X**. Due to the importance (and potential financial rewards) of accurate economic forecasts, there is a massive literature on time series estimation in econometrics (see Hamilton 1994).

In contrast, the problem of statistical early warning consists of finding a time T such that

$$y_t - y_s > \xi \qquad t > T > s$$

for some indicator variable y. In English, this means that the variable y has consistently higher or lower values after time T than it had prior to time T, which would occur in aggregated event data following a qualitative shift in the type of political behavior in which a dyad was engaged.

An additional distinction is that econometric time series generally are highly autoregressive (e.g., GNP, unemployment, prices of consumer goods, and inflation rates) or at least have an autoregressive component combined with generally random noise (e.g., stock prices; exchange rates). The GNP or unemployment rate of a major industrialized economy has tremendous

inertia. For instance, while the stock market crash of October 1929 was sudden, the high unemployment rates of the Great Depression required two or three years to fully develop. Furthermore, most econometric time series are measured continuously rather than episodically, so missing data is less of an issue.

Early warning, however, focuses on shifts in the time-series that are *not* autoregressive, even although the series taken as a whole might be autoregressive. An autoregressive model of war-and-peace will be very accurate, as illustrated by the presumably apocryphal story about the European political analyst who said "Every day from 1910 to 1970 I predicted that Europe would remain at peace when at peace, and remain at war when in war, and I was only wrong four times." This type of model is not, however, very useful. (More technically, such a measure succeeds according to a frequency-based measure but fails according to an *entropy*-based measure, which places higher weight on the prediction of low-probability events.) The econometric problem most comparable to political early warning is forecasting sudden economic shifts such as those observed in exchange rate fluctuations (e.g., the collapse of the Mexican peso or the European Exchange Rate Mechanism). These problems are similar to political early warning in the sense that they are primarily psychological and do not reflect a major change in the underlying physical reality: the economic fundamentals of the Mexican or European economies did not change dramatically during the days of the exchange-rate crises, but the perceptions of the future values of the relevant currencies did change.

Despite these complications, it should be noted that in two very important respects prediction is an *easier* problem than the typical econometric estimation problem. First, forecasting models have right-and-wrong answers, or at least their accuracy can be evaluated probablistically. Coefficient estimation problems, in contrast, do not have answers: one can always specify an error structure, prior probability or alternative model structure that places the estimated emphasis on different variables, and there is no empirical method of deciding between these specifications. Second—and closely related to the first issue—forecasting problems are not affected by collinearity, which is the bane of coefficient estimation in the social sciences because

every behavior tends to be linked to every other behavior. Coefficient estimates with low standard errors are clearly useful for obtaining a theoretical understanding of a situation, but they are not essential for the pragmatic purposes of forecasting (Wonnacott & Wonnacott 1979:81). For this reason, it is not surprising that models with very diffuse coefficient structures—for example neural networks and VAR—increasingly are found in early warning research.

Most of the examples of early warning used in this book employ some form of sequence analysis. As discussed in Chapter 4, the sequence analysis approach has a long history in political science—at the most fundamental level, it is simply a systematic rendition of the "case study" or "lessons of history" technique that has been used by decision-makers since time immemorial (see May 1973, Mefford 1985, Neustadt & May 1986, Vertzberger 1990, Khong 1992) . History is considered relevant to decision-makers because they assume that when a particular set of events and circumstances observed in the past is observed again, the resulting events from that prior case can also be expected to apply in the new case, all other things being equal.

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

By itself, however, belief in the importance of historical examples is insufficient to create empirical regularities because of "Van Crevald's Law"[15]: A conspicuously successful strategic innovation is unlikely to succeed twice precisely because it was successful the first time. More generally, work of the Santa Fe Institute on the so-called the "El Farol Problem" (see Casti 1997) has demonstrated that systems of adaptive utility maximizers generally do not exhibit regularized

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

behavior *because* they look at history. In computer simulations, such agents tend to show quasi-chaotic behavior that is *not* predictable. If the political world consists solely of rational adaptive agents, there is little point in trying to make predictions based on past behaviors.[16] There are undoubtedly some forms of international behavior (for example international exchange-rate behavior) for which this is true.

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

The SOPs are themselves adaptive—they are designed to effectively solve problems and many are acquired through historical experience. But in a situation of the protracted interaction, two organizations with SOPs are *coadaptive*: each responds in part to the environment created by the other.[17] In most circumstances, this eventually brings their SOPs into a Nash equilibrium within the space of possible SOPs where neither can change strategies unilaterally without a loss of utility. This is more likely to occur when the same organizations have been interacting over a period of time, and when the payoff environment has been relatively stable. This is found,

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

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

notably, in the situation of protracted conflicts and enduring rivalries. These are situations characterized by exactly the competitive SOP "lock-in" that we've outlined above—antagonists fight, on repeated occasions, over the same issues, often over the same territory, and without resolution.

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

3.7. Conclusion

One of the anonymous referees who reviewed the validity studies in Schrodt and Gerner (1994) observed "Because human coding of events is so miserable, why should you even *try to* duplicate it with a machine?!". Good point: if WEIS can only be coded with 83% inter-coder reliability, perhaps we should be looking for an alternative system that could be coded at, say, 95% reliability.

McClelland (1983) notes that WEIS was intended to be the first cut at developing an event data coding scheme. Instead WEIS became the final word and the *de facto* U.S. government standard. WEIS has obvious problems such as the ambiguous warning/threat distinction, and similar events in distinct cue categories (e.g. 013 "admit wrongdoing; retract statement" and 061 "express regret; apologize"), and in addition harbor other application-specific cases of ambiguity that become obvious only when one is training coders.

The original event data researchers were intent on getting the maximum amount of information out of a newspaper story. Information was relatively scarce; and the investment of time required first to train a coder and then have the coder locate and read a story was large in comparison to the time require to extract some additional information. We also suspect that the extent to which

the coder's cognitive biases would cause systematic errors was underestimated: The political science research community in the 1960s was not exactly known for a diversity of perspectives, and insights from cognitive psychology on the effects that preconceptions had on interpretation did not become widely incorporated into the discipline for another decade.

This balance has changed with the availability of machine coding. The initial costs in generating a new set of event data involve developing dictionaries, but once these are complete, the incremental cost of coding additional stories is effectively zero. Consequently we might be better off throwing away information that cannot be coded reliably and focusing instead on sentences that can be reliability categorized. For example there is probably little point in attempting to code the affective content of carefully worded official statements whose underlying premise is "Diplomacy is the art of telling someone to go to hell in such a fashion that they look forward to the trip." These are the sort of sentences that even humans can't reliably code, and we should not expect machines to reliably code them . Coding systems that are designed—and tested—for reliability might result in event sequences that give a clearer signal of political events.

At the very least, any project that intends to invest a substantial amount of effort in dictionary development (whether in the refinement of the event coding vocabulary or in the refinement of the coding scheme itself) should probably first ascertain the extent to which those refinements will make a difference given the analytical technique being used. For example, the "report" weighting scheme discussed in subsection 3.3.1 can be tested on a set of data generated using a standard *.verbs* dictionary, and the resulting data set should give at least a rough idea of how much explanatory power is found in the existence of reports alone. (The detection of reports is still sensitive to the *.actors* dictionary, and the actor vocabulary should almost always be modified to accurately code any new region or time period.)

One of us recently reviewed a project that had spent a great deal of time and money meticulously human-coding events using a detailed framework whose coding manual was over one-hundred pages in length. At the analytical phase, however, the data, were subsequently into simple verbal-physical/cooperation-conflict categories similar to those used in section 3.3.3.

They were then used to estimate a neural network model that was insensitive to missing values and coding error, and had a very diffuse parameter structure that virtually precluded the analysis of the effect of the individual event category. This is the data analytical equivalent of buying a very expensive red wine and adding it to a tomato sauce that will be simmered for three hours. Maybe your guests will notice the difference, but in all probability, they won't, and an ordinary table wine would have produced indistinguishable results. Similar economies are possible with machine-coded event data.

Appendix 3.A: A Genetic Algorithm for Estimating Optimal Event Weights

Genetic algorithms (Holland 1975; Grefenstette 1987; Goldberg 1989) are a general purpose optimization method that is particularly effective in situations where there are a large number of local maxima. Because the LML clustering algorithm determines the cluster boundaries as well as determining the weights, the problem is non-linear and required the use of a numerical optimization method rather than an analytical optimization method such as discriminant analysis.

A cluster break was any point that met the following conditions:

1. $LML_t > \Delta$

2. No cluster boundaries in the previous 8 months (i.e., minimum cluster size of 8 months)

I experimented with several values of Δ in the range Δ=0.15 to Δ=0.30. The number of clusters found is inversely proportional to the value of Δ and the $LML_t \approx 0.20$ threshold is comparable to the level found to produce cluster boundaries corresponding to the *a priori* clusters when the Goldstein weights are used. A minimum cluster size is necessary because a sharp change in behavior will produce several consecutive months where LML_t is high.

The genetic algorithm is straightforward: the optimization operates on a vector of weights for the twenty-two WEIS 2-digit categories:

$$\mathbf{w} = [w_1,...,w_{22}]$$

For a given set of weights, an aggregated monthly score is computed for each dyad

$$XY_t = \mathbf{w} \bullet \mathbf{c_t} = \sum_{i=1}^{22} w_i \, c_{it} \qquad \text{where}$$

c_{it} = number of events in WEIS 2-digit category i directed from X to Y in month t

Once these scores are calculated, the LML measure is computed, the boundaries between clusters are determined using the $LML_t > \Delta$ threshold and minimum size rules discussed above, and the fitness measure

$$F^c = \frac{\text{average distance between adjacent clusters}}{\text{average distance within clusters}}$$

is computed with "distance" $\| x_i - x_j \|$ is defined by the correlation metric and the "average distance" is calculated as the average distance between points:

$$\text{Between cluster distance} = \frac{1}{N_1 N_2} \sum_{i} \sum_{C_{1j}} \sum_{C_2} \| x_i - x_j \|$$

$$\text{Within cluster distance} = \frac{2}{N_1(N_1-1)} \sum_{i} \sum_{C_1} \sum_{j>i} \| x_i - x_j \|$$

where N_i = number of points in cluster *i*. The measurement of the points in adjacent clusters rather than comparing the distance of a cluster to all other clusters is done to allow the possibility of the system returning to an equilibrium behavior, so that clusters that are separated in time might occupy the same space.

The genetic algorithm uses 32 **w** vectors that are initially set randomly to numbers between -10.0 and +10.0, the same range as the Goldstein weights. After the fitness of each vector is computed (a "generation" in the genetic algorithm), the vectors are sorted according to the value of F^c and the 16 vectors with the lowest fitness are replaced with new vectors created by recombination and mutation of the top 16 vectors.[18] The probability of a vector becoming a "parent" is proportional to the relative fitness of the vector (in other words, vectors with higher fitness are more likely to be used to produce new vectors). Mutation involves adding a random number between -1 and +1 to the weight, and mutation is done on 50% of the weights in the new vectors.

Most of the results we report are based on runs where the system ran for 48 generations: this was usually sufficient to find a vector that showed little or no change. The system was

[18] One new vector was generated by taking the average weight of the top 16 vectors, on the logic that weights that were not important in the distance calculations (notably those for codes that occur infrequently in the data set) would go to zero as the random weights canceled out. These average vectors were tagged so that their survival in future generations could be tracked. They were rarely selected in the first set of experiments where both the weights and cluster breaks were allowed to vary, but were frequently selected in the second set of experiments where the weights were optimized for a given set of cluster divisions.

occasionally allowed to run for 128 generations; this produced essentially the same results as the shorter runs. This system was implemented in a C program; the source code is available from the authors.