Forecasts and Contingencies: From Methodology to Policy

Philip A. Schrodt

Department of Political Science

University of Kansas

1541 Lilac Lane, Lawrence, KS 66044 USA

phone: +1.785.864.9024 fax: +1.785.864.5700

schrodt@ku.edu

August 2002

Paper presented at the theme panel
"Political Utility and Fundamental Research: The Problem of Pasteur's Quadrant"
at the American Political Science Association meetings, Boston, 29 August - 1 September 2002

Electronic copies of this paper are available at
`http://www.ukans.edu/~keds/papers.html` and
`http://apsaproceedings.cup.org/`

Abstract

A "folk criticism" in political science maintains that the discipline should confine its efforts to explanation and avoid venturing down the dark, dirty, and dangerous path to forecasting and prediction. I argue that not only is this position inconsistent with the experiences of other sciences, but in fact the questions involved in making robust and valid predictions invoke many core methodological issues in political analysis. Those issues include, among others, the question of the level of predictability in political behavior, the problem of case selection in small-N situations, and the various alternative models that could be used to formalize predictions. This essay focuses on the problem of forecasting in international politics, and concludes by noting some of the problems of institutional culture—bureaucratic and academic—that have inhibited greater use of systematic forecasting methods in foreign policy.

As I prepare this essay, a two-day Congressional hearing—and fully half of *The New York Times* letters-to-the-editor page (1 August 2002)—is addressing the possible consequences of a United States military action in Iraq. The voices being heard include generals, current and former civilian officials involved with defense and foreign policy, and assorted pundits who frequent CNN. Conspicuously absent from the consideration of the U.S. policy options are political scientists: only Morton Halperin might be so identified, and his presence probably comes from his experience in the State Department rather than from his work on bureaucratic politics.[1]

This situation is hardly unusual. A few days after the attacks on 11 September 2001, I received a call from a program director at an organization responsible for substantial funding of political science research who said "My bosses are asking me why they haven't been seeing any political scientists analyzing this situation. They wonder what we are doing." This absence can be explained in part by the narrow focus of the popular media on a few experts—and this political scientist, at least, was spending a large amount of time doing public presentations on the subject, albeit largely using qualitative methods (see Gerner and Schrodt 2002)—but the fact remains that in the realm of foreign policy, in contrast to issues such as election outcomes and/or future economic performance, the use of forecasting based on systematic methodologies is very much the exception rather the rule.[2] Pasteur's quadrant generally has been avoided in quantitative international relations research.

This lack of involvement in policy-relevant forecasting cannot be attributed to an absence of theoretically interesting questions relevant to policy. For example, at the present time the first exercise in the course on peacekeeping operations at the Command and General Staff College at Ft. Leavenworth is the design of a U.S.-led peacekeeping force for the West Bank and Gaza. The students, I suspect, address this almost entirely with historical analogies and other atheoretical methods. But if this question is considered theoretically, one is quickly led into a host of core questions about political instability, civil-military relations, protracted conflict, mass mobilization, social movements, and the role of hegemonic powers. The world at the beginning of the 21st century is a very interesting place.[3]

---

[1] Two political scientists—Shibley Telhami of the University of Maryland and Fouad Ajami of Johns Hopkins—testified on the topic of "regional considerations." Telhami has co-authored articles using quantitative forecasting methods (Goldstein et al 2001) and has introduced these results into public debate, for example in an op-ed by Thomas Oliphant in the *Boston Globe* 10 March 2002. However, I suspect that both Telhami and Ajami were chosen for their expertise in the language and regional politics—area studies, in a word—rather than forecasting.

[2] Perhaps in response to 9-11, the popular media have recently shown an interest in quantitative forecasting projects, most notably the one-page synopsis of four conflict forecasting projects in *Wired* (July 2002, pg. 044 [sic]), followed the next month by a nice discussion of Lewis Richardson's statistical work on war magnitudes (August 2002, pp. 081-082). The first *Wired* article in turn inspired a one-hour broadcast on 17 July 2002 on National Public Radio's *Public Interest with Kojo Nnamdi* (http://www.wamu.org/pi/shows/piarc_020715.html; accessed 2 August 2002) featuring, among others, Harvard's Gary King. Is political forecasting kewl or what?

[3] It is also interesting to note that almost all of the core approaches in "scientific" international relations, including the statistical analysis of large-scale data sets (Richardson 1960a), differential equation modeling (Richardson 1960b), expected utility and game theory, content analysis, and event data analysis arose in response to specific policy concerns such as World War I and II, the Cuban Missile Crisis, and the nuclear arms race (see Burgess and Lawton 1972, Rapoport 1974, Zinnes 1976, Merritt, Muncaster, and Zinnes 1993). While the actual policy *impact* of these studies may have been quite limited (O'Neill 1994), the *motivation* was at least as much driven by an interest in policy as by an interest in basic science. The alienation that arose between the academic and policy communities during the Vietnam War marked the end of this tendency.

My position on this issue is two-fold. First, contrary to some received wisdom, forecasting is a legitimate scientific endeavor, and claims to the contrary are misguided at best and self-defeating at worst. Second, the methodological issues involved in forecasting address many of the core questions in contemporary political analysis and serious efforts devoted to forecasting would probably substantially increase the quality of our science. My remarks will focus on international relations and foreign policy, although I suspect they apply almost equally to the field of comparative politics, and possibly to some studies of United States politics.

Political Forecasting

There is a diffuse—but widespread—"folk criticism" which maintains that political scientists studying the international system should confine their efforts to explanation and avoid venturing down the dark, dirty, and dangerous path to forecasting and prediction. Systematic predictive work in the field is rarely published—the forecasting work of Bueno de Mesquita and isolated articles such as Pevehouse and Goldstein (1999) are unusual exceptions[1]—and comparative studies of forecasting accuracy such as those found in electoral studies are also non-existent.[2] Papers and articles that attempt to forecast will be, likely as not, dismissed by the discussant/referee with a brusque "That's only a forecast." [3]

This philosophical position has left me completely mystified from the first time I encountered it in graduate school. In the natural sciences, successful forecasts are the epitome of validation of a theory, and some successful predictions—for example Edmond Halley's forecast of the return of the eponymous comet, or Sir Arthur Eddington's 1919 confirmation of Einstein's prediction of the deviation of starlight during a total eclipse—are considered landmarks in the history of science. In the social sciences, one finds industrialized countries spending hundreds of millions of dollars on data collection and econometric modeling in order to provide economic forecasting. The accuracy (and influence) of opinion polls is now sufficiently high that their publication in the days prior to an election is now regulated in many democracies.

---

[1] Bueno de Mesquita has probably been the strongest advocate of the legitimacy of forecasting in the academic international relations community; a thorough review of his work and arguments can be found in Ray and Russett (1996). For example, Bueno de Mesquita (2000: Chapter 17) is the only introductory text I know of that includes an extended discussion of forecasting. Bueno de Mesquita's models use expected utility theory in contrast to the statistical and pattern recognition approaches I will be emphasizing.

The other niche for policy-oriented forecasting is the "early warning" literature. Choucri and Robinson (1979), Hoople, Andriole and Freedy (1984), Laurence (1990) and Singer and Wallace (1979) are useful surveys of the first-generation of this work in the 1970s; Davies and Gurr (1998), Rupesinghe and Kuroda (1992), and Schmeidl and Adelman (1998) cover the "second-generation" work from the 1990s.

[2] See Wlezein 2001 as well as the related articles in the March 2001 issue of *PS* (http://www.apsanet.org/PS/march01; accessed 1 August 2002) for discussions of election forecasting. Gurr and Lichbach (1986) is one of the few examples of a similar exercise dealing with internal conflict from a comparative perspective.

[3] My efforts to locate a canonical source for the philosophical basis of this criticism have been unsuccessful so far. A posting to a listserv for quantitative international relations researchers elicited numerous helpful comments—and complete agreement that the issue is real, and not just a strawman—but no definitive articles. Ray and Russett (1996) primarily address Gaddis (1992b) and the unrelated criticisms of post-modernists—"Derrida's quadrant", the lower-left cell of the Stokes typology. Are we perhaps dealing with a phantom here, the philosophical equivalent of Bismarck's story (Morgenthau 1973:130) of the Russian soldier standing guard in St. Petersburg at a nondescript location where Catherine the Great had admired a wildflower a century earlier? Further suggestions on the relevant literature would be appreciated.

But forecasting in international relations is generally—though not universally— discouraged. This is not to say that foreign policy is formulated in the absence of any forecasts. To the contrary, foreign policy is subject to endless rounds both unconditional and contingent forecasts. But in contrast to the fields of economics and elections, the forecasts that become part of the public debate are done almost entirely using unsystematic methods.[1]

For purposes of this discussion, a forecast is simply a statement that a particular event—ideally one that has been specified unambiguously—will occur (or might occur, with some unambiguous probability) at a point in the future.[2] In economic and electoral forecasting, the "event" is usually the value of a continuous variable (for example an unemployment rate or the percentage of voters choosing candidate X). In international relations, forecasts usually concern a discrete occurrence chosen from a relatively small set of possible events (e.g. the US will or will not successfully occupy Baghdad; Saddam Hussein will or will not remain in power).

There are two different types of forecasting with very different properties. *Unconditional* forecasts seek simply to predict the future under a *ceteris paribus* condition. The exercise is one of simple extrapolation: if things continue on the current track, then X, Y and Z will occur. This is the classical crystal ball problem; the most appropriate natural science analogy would be meteorology. For reasons discussed below, unconditional forecasts, while hardly a trivial exercise, are the simpler and less scientifically interesting of the two categories.

Of greater interest to both the policy and the scientific communities is the *contingent forecast*—the "what if" question. The appropriate natural sciences analogy here is chemistry (or applied physics, a.k.a."engineering"), where variables are continually manipulated to establish true causal relationships. This is a much more difficult problem because contingent forecasts will only be correct if the model has identified true causal mechanisms; correlation is not sufficient. For example, if a model indicates that infant mortality is a strong correlate of state stability, but in fact infant mortality is simply a surrogate indicator for a cluster of other variables that are the "true" causes, then efforts to reduce infant mortality, however meritorious in their own right, will have only limited effects on state stability. This confusing of correlation and causality is probably one of the main reasons that forecasting has gotten such a bad name.

---

[1] Due to problems of secrecy, it is difficult to determine just how much systematic forecasting work using statistical or algorithmic methods is being done by the U.S. government. Having watched this for a number of years—and having on numerous occasions alerted individuals in one part of the government to work being done in some other part of the government—my sense is that there is a persistent under-current of work being done, but there is nothing remotely resembling a coherent plan for the development of such models. The work that is done has very little input from social scientists, most of the projects don't get very far and—in contrast to the regular use of systematic modeling in economics and many fields of domestic policy research (e.g. educational reform, welfare reform)—rarely surfaces in the public debate. Given the penchant for leaks of classified material, it seems unlikely that a vast enterprise of quantitative modeling lies, hidden for decades, within the foreign policy, intelligence, and defense policy establishment.

[2] I will be focusing almost exclusively on forecasting over time, since that is the application of greatest interest to the policy community. However, it is worth noting that in the natural science, many predictions do not involve time but rather deal with additional instances of a theoretically-predicted pattern. For example contemporary geological theory predicts that petroleum deposits will occur under salt domes in sedimentary formations, and that it will not occur in granite intrusions; the opposite is true for gold. An emerging astronomical theory suggests that all galaxies have at their center a massive black hole, and this "prediction" has led astronomers to look for additional evidence of such objects. Neither of these are strictly temporal forecasts—the oil and gold were in place millions of years before the prediction was made, and the black holes formed billions of years in the past—but both provide empirical evidence in support of a theory.

As George (1993) has pointed out in great detail, most of the political analysis of interest to the policy community involves contingent forecasts. This is necessary because policy-making is a feed-*forward* problem (Casti 1989: 355-357)—policies are determined based on their projected future impacts. In the policy world, the complications do not arise merely because the answers to the game isn't in the back of the book. In the absence of extensive forecasting exercises, the pay-offs of the game aren't filled in, and in many cases, the options aren't even known. While one could probably function as a viable political actor using heuristic methods that do not require feed-forward—for example, tit-for-tat (Axelrod 1984) or standard operating procedures incrementally adapted from past failures (Cyert and March 1963)—most policy involves extensive efforts at forecasting, particularly on policy issues that are highly salient at a particular time, such as the aforementioned hearings on Iraq.

As I will argue in more detail below, contingent forecasts in particular invoke a number of core questions in basic political methodology. George's observation would further suggest that the most interesting problems for the policy community also involve some of the most interesting (and difficult) problems in applied political analysis.

Weaknesses in Forecasting

Before proceeding to the discussion of why forecasting is an interesting problem, I should address briefly some of the obvious weaknesses in the enterprise, as well as one concern I do not consider to be legitimate. Most of these apply primarily to unconditional forecasts rather than contingent forecasts.

The single greatest criticism of unconditional forecasting is that it leads to "data-mining": Take a very large number of variables, cram them into a generic model, crunch the numbers and then accept the results irrespective of whether they make any theoretical sense. The data-mining approach is tempting because it is (a) easy; (b) looks impressive; and (c) actually works in applications where one is interested only in unconditional forecasts to the exclusion of explanatory theory or manipulation of the underlying variables. For example, a life insurance company using data-mining might determine that individuals who attend stock car races are more likely to die of lung cancer. The linkage is cultural rather than causal—individuals who enjoy stock car races appear more likely to be smokers, a relationship possibly linked to the extensive sponsorship of stock car races by tobacco companies—but for an insurance firm simply interested in setting insurance rates based on projected mortality, stock car racing would be a useful indicator. However, if a government was interested in reducing the rate of lung cancer—"Kids, if a friend asks you to attend a NASCAR event, JUST SAY NO!"—this variable would be ineffective, because of the absence of the causal linkage.[1]

[1] An interesting example of a useful predictive study that is partly causal and partly correlational is a model recently developed by the British Meteorological Office and the Department of Health that forecasts demand for health services as much as two weeks in advance using temperature data (*Economist* 3 August 2002, pg.66). The model is sufficiently accurate that one hospital saved £400,000 in operating costs over a 4-week period by using it. In this instance, the statistical relationship is not strictly causal, since temperature causes illness to rise or decline via a complex set of behavioral and physiological changes that would be impractical to model in detail. Furthermore, the objective of the Health Service is purely one of prediction, not intervention (although the *Economist* does note that this study reinforces Mother's advice to wear a scarf on cold days). On the other hand, because it is possible to establish theoretically-justified linkages between temperature and illness—and clearly the illness is not causing the temperature change—one does have elements of causality here.

A dramatic example of the failure of the data-mining approach (and almost everything else) can be found in the most recent large-scale unclassified U.S. government effort at quantitative political forecasting in comparative politics, the State Failures Project (SFP; Esty et al 1995, 1998). Ironically, the SFP started out in the correct direction by involving some of the top people in the field who could have applied their very substantial theoretical knowledge to the analysis. Yet in the end—probably due to inflexible government contracting—SFP degenerated into a simple data-mining project. Vast amounts of money were [apparently] spent on acquiring data on hundreds of variables (99% of which were not used in the final model); these were thrown into a couple of very general models which then selected, atheoretically, a small set of common-sense indicators that are almost certainly correlational, not causal. As a consequence, the SFP model performs abysmally in out-of-sample tests (King and Zeng 2001).

Yet forecasting does not have to involve data-mining, just as one does not need to use tobacco products in order to enjoy stock car racing. In out-of-sample tests, models that utilize theoretically justified causal relationships should do as well—and probably a whole lot better—than those derived by data-mining. Data-mining, like the use of tobacco products, is a habit to be broken, not a necessity.[1]

While unsystematic policy forecasting is very common, it is not necessarily very accurate, a point I will address in more detail below. The most dramatic example of this problem in recent years was the failure of Western political analysts to forecast the end of the Cold War, a topic that has been explored in depth by Gaddis (1992a, 1992b) among others. This failure occurred despite the efforts of thousands of analysts who had access to an information-collection effort costing tens of billions of dollars per year, despite observations by journalists and others outside the official community (e.g. Cockburn 1983) that the Soviet system was near collapse, and despite an ideological pre-supposition that asserted that communism was fundamentally flawed and inefficient. The policy community should not have missed this one—the collapse simply extrapolated trends journalists had already documented, and occurred for reasons consistent with the theoretical critique from liberal political theory.[2]

But political, institutional and ideological pressures prevailed to blind the official policy community to the obvious. A fatally weakened Soviet Union would hardly serve to justify massive inflation of the defense budget; individuals who had spent their entire lives studying the nuances of the Soviet Communist Party could hardly imagine life without it, and the "Soviet threat" was far too convenient an excuse for justifying existing foreign policy. This forecasting mistake the cost the U.S. perhaps $500-billion dollars in unnecessary defense spending and

[1] Singer and Wallace (1979; pp. 7-12) provide an interesting discussion of this problem in their introduction to a series of articles covering first-generation crisis forecasting models. They are very conscious of the dangers of data-mining—Singer, despite the assertions of his detractors, has made this clear for his entire career—but also defend the necessity of purely inductive work given the immature state of the field. I think they were right about this, *at the time*. But the work in Singer and Wallace (1979) is a quarter-century old and much (though not all) of that statistical brush-clearing now has been done. In 1995, we already knew that infant mortality rate was an excellent surrogate indicator for some combination of development and political stability; we didn't need to sort through 700 other variables to figure this out.

[2] I am using the term "liberal" in the sense of political theory, not the sense of contemporary United States political discourse.

subjected Eastern Europe to perhaps an additional five to seven years of Communist control.[1] In short, before we go too far in criticizing the failures of systematic forecasting, it is well to remember that the unsystematic efforts are scarcely without flaw. In the words of Edward Abbey (1975: 338) "One man alone can be pretty dumb sometimes, but for real bona fide stupidity there ain't nothing can beat teamwork."

An additional general criticism is sometimes made that I do not consider legitimate: the data available to academic research are not sufficiently good for forecasting. Often implicit in this criticism is an assumption that adequate data exist somewhere, doubtlessly stamped "Top Secret."

But this is unlikely, particularly for data required to make forecasts of political behavior at a six-month horizon. The extent to which governments held a monopoly on policy-relevant information was limited even in the past (see Moynihan 1998) and has declined precipitously in the past fifteen years with the advent of global, 24/7 media coverage. Add to this our ability to acquire data in a timely fashion over the Web, reducing a process that once took weeks and involved the movement of punch cards or tapes to a process that takes seconds and a couple of mouse clicks, and the emerging ability to code information directly from the Web (Schrodt and Gerner 1994; King and Lowe 2001), and adverse access to timely data should be only a modest constraint.

This issue is largely an empirical one: if we can construct models that forecast accurately in out-of-sample tests, the data are good enough. If they don't, the data might be inadequate, though the problem may just as well lie in the underlying theory or choice of model. Data may be insufficient for forecasting some issues some of the time, but they are not inadequate for forecasting all issues at any time, and the information available to the academic community differs only marginally from that available to the policy community.

Methodological Challenges

Data-mining and telling the boss what she or he wants to hear are not particularly interesting methodological issues. These aspects of forecasting may go far in explaining the evident disrepute directed at the forecasting exercise. Yet if forecasting is taken seriously, it addresses many core theoretical and methodological issues in political science, as well as seriously challenging some of our deeply held ways of doing things. (I regard this second characteristic as positive rather than negative; others may disagree.) My focus is on international behavior and foreign policy, but I would guess that many of these observations would apply to the field of political science generally.

[1] I reject here the argument popular among U.S. conservatives that the defense spending increases of the Reagan era caused the Soviet Union to collapse. It was already collapsing, and in the opinion of this inexpert observer, the effect of the Reagan military spending and rhetoric was to provide the Communist Party with an external threat that actually delayed the *perestroika* (theirs, not ours…) and *glastnost* reforms that paved the way for the collapse of Communist control. Meanwhile, illustrating that, contrary to Cyert and March, learning is not necessarily adaptive, the contemporary "axis of evil" policy of the United States is repeating exactly this mistake with respect to maintaining the clerical regime in Iran against an active Iranian democratic reform movement.

How accurate are unsystematic political predictions?

Just how accurate *is* traditional, intuitive, unsystematic political forecasting? The accuracy of a statistical model is usually known—often with multiple measures—and reported. But we have remarkably few comparable assessments of human political forecasting. One can easily point to big obvious failures—Pearl Harbor, flawed assessments of government success in Vietnam in the years prior to the 1968 Tet Offensive, or the aforementioned failure to predict the end of the Cold War. But are these just balanced by equally dramatic cases where the predictions were correct and things worked out as anticipated: the Marshall Plan, the Cuban Missile Crisis, U.S. peacekeeping efforts in the former Yugoslavia?

On a day-to-day basis, or over long periods of time, what is the accuracy? More specifically, what is the accuracy of the analysts, as opposed to the policy implementations, which may have been implemented over the objections of analysts? What types of behavior can we predict well; what types of behavior can we predict poorly? What is the trade-off in the technique between false positives and false negatives? When the appropriate studies are available in the literature, I can answer these questions quite easily for statistical models. I have little idea what comparable figures are for humans.

The little available evidence suggests that it might not be very good. An on-going study by Philip Tetlock—which involves over 200 experts and 5,000 predictions on a number of foreign policy issues—found as of 1998 that "experts were only slightly more accurate that one would expect by chance" (Tetlock 1999: 351).

But this raises an interesting measurement issue: what is the null model against which human forecasts should be compared? In Tetlock's statement, "chance" is presumably a 50% probability of being right or wrong; hence a chimpanzee with a Ouija board could do about as well as the experts, much as chimpanzees tossing darts to select stocks have been shown to out-perform most mutual fund managers. The plausibility of that 50% probability rests on Tetlock's selection of questions, all of which appeared to be "close calls" that could go either way (for example, the return to power of the Communist Party in Russia in the early 1990s, or the collapse of the European Union due to its monetary crisis). Had Tetlock posed the questions "The Earth will unite to repel an invasion of sentient slime molds from Proxima Centauri" or "The United States will collapse in the early 1990s" and expert opinion split about 50-50, the face validity of Tetlock's selection of "experts" would be called into question. This, however, raises the question of how Tetlock chose those issues, and I would guess this involved a lot of intuitive and qualitative prediction.

The null-model problem is further complicated by the fact that simple autoregressive prediction— whatever happened yesterday will happen today—produces extraordinarily accurate forecasts when evaluated at a high frequency and with short time horizons. This is not the problem Tetlock is pursuing—he is asking experts to make single predictions at long time horizons—but what *is* the appropriate way to frame this problem?

A similar problem arises in the assessment of one of the few formal models known to have been used and evaluated by the Central Intelligence Agency, the "Policon" model[1] derived from the

[1] Now "Decision Insights, Inc."

work of Bueno de Mesquita. In a declassified study, Feder (1995) supports a claim of accuracy in excess of 90% (see also http://www.diiusa.com/record.html) and strongly endorses the system

> Forecasts and analyses using Policon have proved to be significantly more precise and detailed than traditional analyses. Additionally, a number of predictions based on Policon have contradicted those made by the intelligence community, nearly always represented by the analysts who provided the input data. In every case, the Policon forecasts proved to be correct. (pg. 57 in original document; from http://www.diiusa.com/cia_pg18.html, accessed 15 August 2002)

But the importance of the 90% accuracy—while apparently superior to those of human analysts—still cannot be evaluated without answering the "compared to what" question, as well as having a systematic evaluation of the system.

Systems using feed-forward can predict badly yet still persist. In all likelihood, over the past three centuries we have seen a gradual (not necessarily monotonic) improvement in the accuracy of forecasting as theory and the availability of data improve. For example, in the realm of political economy, the theories of Adam Smith and David Ricardo were correctives to earlier mercantilist theories. Mercantilism was a lousy theory, and contributed substantially to the economic collapse of Spain and Portugal in the 18th century, presumably not the outcome desired by their ministers of finance. The enhancement of liberal economic theories by Keynes in response to post-World War I dislocations and the Great Depression did much the same to stabilize industrial capitalism in the 20th century. The contributions of John Locke in formulating separation-of-powers for emerging liberal democratic systems and Karl Marx (if not Marxists) in elucidating socio-political stresses in industrialized capitalism may have played similar roles in politics. Qualitative prediction accuracy is a moving target, but it would be very useful to get some assessments of it.

Perhaps the greatest question in qualitative forecasting is the extent to which analogical reasoning (Neustadt and May 1986; Khong 1992) is an effective predictor. Reasoning by historical analogy is probably the most common method used in intuitive political analysis, but just how good is it? Analogical reasoning in politics may simply be epiphenomenal: we use the technique because our brains have been over-tuned by evolution for the task of pattern recognition, and find patterns where they don't exist. A second possibility is that pattern recognition works reflexively—pattern recognition is an effective strategy in a world of pattern recognizers, and the resulting world is sufficiently stable to allow the production of Neustadt, May, Khong and the APSA. A final possibility is that reasoning by historical analogy "works" in some objective sense.

Because historical analogies are selective rather than random, the problem of analogical reasoning is closely related to the issue of case selection, which I address below, The method is difficult to deal with in formal models because pattern recognition of social behavior has proven to be a largely intractable problem in artificial intelligence research (despite considerable invested effort: see Kolodner 1988, 1993), due at least in part to the differences between the dominance of associative recall in human memory compared with the sequential recall of computers. But the problem is real, it is important, it is an issue of methodology, and it could be studied.

To what extent is political behavior intrinsically predictable?

Andrew Lo, an MIT economist, provided a candid definition of "physics envy" in a recent issue of *The Economist:*

> We would love to have 3 laws that explain 99% of economic behavior. Instead we have 99 laws that explain maybe 3% of economic behavior.[1]

This assessment is probably unduly pessimistic—at the micro-economic level at least, I would guess that simple laws such as supply-and-demand explain far more than 3% of behavior—but nonetheless instructive. The gold standard of science remains the deterministic universe of Newtonian physics, and consequently anything short of determinism is considered partially flawed.

While this perception is widespread, it is not an accurate description of even the natural sciences. The parts of physics that are deterministic are the exception rather than the rule in the natural sciences: Geology, meteorology and ecology all use statistical rather than deterministic methods, and even though biochemistry is deterministic at its core, the relevant equations are too complex to predict even such basic problems as protein folding. Sticklers can further argue that through quantum mechanisms at the micro level and chaos theory at the macro level, even physics is not deterministic at very small, or very large, time scales. Nonetheless, at an intermediate level encompassing most of the phenomena we experience in everyday life, physics (and much of chemistry) is largely deterministic. Researchers in those fields can reliably measure constants to an accuracy of millionths or even billionths. Political scientists can't, nor will we ever be able to.

This ghost of determinism haunts political science. Consider the democratic peace hypothesis: Is this invalidated by the actions of Finland in World War II, or those of the short-lived Roman Republic in 1849? To someone seeking a Newtonian or Leplacian world of complete determinism, these two idiosyncratic cases are falsifying evidence. To the statistician, well…weirdness happens in even the most regular of relationships. While I suspect that the statistical view is shared by almost all contemporary political methodologists, there are still wisps of determinism that need to be exorcised.

But having accepted that there are likely to be random elements in all models of social behavior, we still have the issue of determining the level and source of the randomness. At the present time, I see two key issues here.

First, we need to re-admit the importance of estimating the magnitude of the error term. This was banished from the discipline about two decades ago (see, most notably, King 1986), in part due to the general movement towards emphasizing explanation over prediction; in part as a reaction to earlier statistical models that emphasized maximizing the R^2, often using primitive data-mining methods. But the variance of the error term *is* an estimable parameter of a model, and one needs to pay some attention to it.

Of greater interest is the issue of where the randomness comes from. In a great many cases, the answer is boring—our models are incomplete, and our data are badly measured. But in some residual behaviors, the answer is far more interesting: randomness arises from the nature of the

[1] *Economist,* 363, 8273 (18 May 2002) "Survey of International Finance" pg. 17

process itself, and therefore cannot be eliminated no matter how good our models, nor how accurate our data.

A well-known example of intrinsic randomness arises from zero-sum games with mixed strategy solutions. Rational and informed players in such situations—which are clearly common in politics—will behave randomly. A more complex source of randomness has recently been found in the field of agent-based modeling (Axelrod 1997): Arthur's (1994) "El Farol problem" produces a situation where the strategies of individual players (and consequently the global behavior of the system) continually change, much like the patterns of a chaotic system, without ever reaching an equilibrium. Because the El Farol problem involves a coordination game, it is also likely to have analogues in political processes.

One size does not fit all. Going back to the field of economics, consider three levels of predictability. In almost any marketplace in the Third World, illiterate peasant farmers who have never read a word of microeconomic theory can set prices such that by the end of the day their stalls are empty. Likewise, in every urban area and major crossroads in the Third World, one finds jitney drivers who have organized a decentralized transportation network whose efficiency, if not safety, would be the envy of any architect of mass transit. These systems achieve a predictable optimum despite being composed of imperfect individual parts with access to only partial information.[1]

At a more sophisticated macro-economic level, central banks in all of the developed democracies have effectively eliminated bank panics—the scourge of the late 19th century—and hyperinflation. Arguably they have also substantially smoothed the fluctuations of business cycles. But nowhere have central banks managed to *eliminate* business cycles, despite heroic efforts to do so (accompanied by the occasional premature claims of success, as we saw in the late 1990s). Likewise, Kenneth Rogoff of the International Monetary Fund summarizes our knowledge of exchange rate forecasting:

> Movements in exchange rates ... are frustratingly difficult to explain, much less predict; policymakers must be conscious of this. As a young economist at the Federal Reserve Board in the 1980s, I was asked to investigate whether various new-fangled exchange-rate models could help predict rates. My colleague, Richard Meese, and I came back with the then-radical, but now-conventional, finding that no structural model can reliably explain movements after the fact, much less predict them. (*Economist* 364: 8284 (3 August 2002) pg. 63)

The appropriate models are not there, yet. If industrialized economies operate through a mechanism similar to the El Farol model, the models may never be; or alternatively in a century our descendents may look back upon the 20th century business cycle and exchange rate instability much as we look at the 19th century bank panics or 18th century mercantilism. In the

[1] Granted, the information available in a farmers market or to jitney drivers is more accurate and timely than that available to policy makers concerned with foreign policy (or, billions of dollars of technology investment notwithstanding, the information available to the manager of a local Wal-Mart). But it is still incomplete: the farmer does not know that someone's sister has unexpectedly arrived from out of town and therefore a more elaborate meal will be served that evening, nor did the jitney driver who brought the sister know of her plans in advance.

meantime, we need to know how much error we are dealing with, and whenever possible, why that error is present.

For what time horizon *is behavior predictable?*

An issue closely related to the problem of intrinsic randomness is the relationship between predictive accuracy and time horizon. The naïve view—based on experience with models whose prediction error bounds expand as a function of time due to the accumulation of random errors—is that short-term prediction is easier than long-term prediction. I am increasingly convinced that this is wrong: finely-grained short-term prediction is nearly impossible, but longer-term trends can be modeled. We can forecast tides but not individual waves.

I take this position in part based on experience. Schrodt (2000) discusses the development of hidden Markov models for forecasting levels of conflict in the former Yugoslavia, Central Asia, and West Africa at time horizons of one, three, and six months. I had expected the accuracy of the models to drop as I extended the time horizon. They do, but only in aggregate and only very slightly. This result was surprising and I spent a great deal of time checking to be certain that I was not accidentally incorporating more recent information into the long-horizon models. As far as I know the result is solid.

On further reflection, perhaps it should be expected as well. Short-term predictions often involve the *plays* of a game, which may be intrinsically random (e.g. mixed strategies) or, in the real world, subject to extensive attempts at deception.[1] Longer-term predictions, in contrast, involve the *pay-offs* of the game (and therefore the strategies) and consequently are more regular. We can forecast the expected value of a mixed-strategy game even if we cannot forecast the individual plays.

This fits nicely into Pasteur's quadrant. Contrary to many popular perceptions, in my experience the policy community is actually more interested in long-term predictions than in short term. Six months is the modal value for what is considered a "useful" time horizon because it is enough time to do something: food can be shipped, troops mobilized, diplomatic initiatives undertaken. With a horizon of one day—or even one month—all you can do is stand by and watch the train wreck, perhaps with a bit of extra time to figure out how to explain that it wasn't your fault. Knowing the tides is sufficient for most of the policy community.

The policy community is (rightly) quite skeptical of any claims of accurate predictions in the very short term, and it would be helpful for quantitative forecasters to make it clear what we *can't* do. In particular, we could not have predicted that terrorist attacks would occur on 11 September 2001. We are not in the business of creating that type of crystal ball.[2] But with

---

[1] Kuran (1989, 1995) provides a formal analysis of this problem in the context of predicting political revolution.

[2] Speaking of crystal balls, it is interesting to note that in J. K. Rowling's *Harry Potter* series, the only form of magic that is treated with skepticism is divination. More generally, the principle driving all of Rowling's plots is that wizards and witches have nearly total control over their physical world, but lack extraordinary access to *information* critical to their lives. Harry can fly, transform his body, and become invisible, but must rely on the same techniques as Sherlock Holmes in determining who is friend and who is foe.

Rowling's *opus*, by the way, should be required reading for any graduate student aspiring to a career as a political methodologist. But methodologists only get to spend summers at Hogwarts; the academic year is spent at the Dursley's.

suitable methods and attention to the issue, we probably *could* have predicted that there would be some sort of terrorist act directed against the U.S. in the fall of 2001 if we had been looking for it. Such predictions— presumably derived from qualitative models—were in fact quite common in the weeks before 9-11. Al-Qaida had been regularly staging attacks on the U.S. through the 1990s at a frequency of about one every 18 months and the September events were consistent with that general pattern.[1]

Which formal models are best for forecasting political behavior?

The one clear methodological advantage to prediction over explanation is that one can assess whether one model works better than another. Predictive models—even unconditional prediction—can be determined to be better or worse, useful or useless when their performance is tracked over a long period, either in real time or simulated through out-of-sample tests using only temporally prior information for the parameter estimates. Explanatory models—that is, models whose objective is solely the estimation of parameters using in-sample techniques—cannot so assessed in this manner. Once one has eliminated all of the obvious statistical problems, any other changes simply reapportion the explained variance based on information, such as Bayesian priors or assumptions about the underlying structural, temporal or spatial error structure, supplied *a priori* by the analyst. However elaborate the computational pyrotechnics, what you get out is determined by what you put in, and alternative explanations are indistinguishable.[2]

This emphasis on in-sample explanation instead of out-of-sample prediction has in turn lulled the profession into dependence on a single method: least-squares estimation of linear models. Almost all of our statistical findings over the past twenty years depend on either regression analysis (albeit sometimes elaborate variants on OLS that deal with complex error structures) or logit/probit, which has almost all of the same strengths and weakness. Like the farmers of the Great Plains who once planted tracts the size of small countries with a single variety of wheat, only to watch the entire crop succumb to disease, we have created a methodological monoculture.

OLS linear regression has a lot going for it: it is robust, well-understood, computationally efficient, and—under unrealistic assumptions—has all of the nice properties we would like in an estimator. We should also not under-rate the importance of our ability teach regression to graduate students—few of whom choose political science out of a love of mathematics—in one or two semesters to a point where they can employ the technique at a level that only creates difficulties. This, in turn, has gradually led to an environment where one can present a regression

---

[1] Such a prediction, however, may be in Edison's quadrant of applied research rather than Pasteur's quadrant: it is not particularly interesting from a theoretical perspective. This is in contrast to the earlier example of how to deploy U.S. peacekeepers in the Middle East, which raises a number of theoretically interesting questions.

[2] An extreme example of this phenomenon is the massive literature attempting to differentiate between the democratic peace ("Democracies don't fight each other") and liberal peace ("Developed states with trade linkages don't fight each other." Or, in Thomas Friedman's formulation, "States with McDonald's restaurants don't fight each other"). After about fifteen years and dozens if not hundreds of refereed statistical articles using a wide variety of [linear] methods, the issue remains open. Why?—*because the two hypotheses cannot be differentiated given the available historical evidence.* The various approaches simply re-arrange the variance explained.

In fifty years we may be able to differentiate between these two hypotheses. Today we cannot. No improvements in estimation will change this fact. There is no there there. Give it a break!

model in a policy setting and there is a good possibility that at least some of the individuals listening will understand the presentation.[1] I am not suggesting we discard regression altogether.

But regression also has at least three substantial weaknesses.[2] It has no systematic means of dealing with missing values (particularly if these occur non-randomly); it is highly sensitive to collinearity; and in the presence of sub-populations (or outliers; same thing), it will confidently (that is, with low standard errors) estimate coefficients having the opposite sign of the true effects that are found within the sub-populations. All three of these problems are very common in political science data, particularly in data sets found in comparative politics and international relations. In addition to these fundamental problems, the assumption of linearity usually has been accepted with little or no experimentation with alternatives.

A focus on prediction would not solve the problems with regression, but it would highlight situations where regression models do not work very well and this, in turn, could lead to greater exploration of alternative models. And there are alternatives! For example, neural network and ID3 models are substantially more robust in the presence of missing values than regression. Clustering algorithms such as K-Means and correspondence analysis are more robust with respect to sub-populations. (Cluster analysis is also likely to be more effective than regression analysis in situations where analogical reasoning appears to be important.) Systematic selection of cases (or, where possible, experimentation) can address the issue of collinearity, although of the three problems I've mentioned, collinearity is the most difficult. There is more to computational political analysis than the linear model.[3]

---

[1] Similarly, one will now occasionally see elite publications such as the *Economist* report the correlation coefficient on the lines that are blithely drawn through scattergrams. On this issue we are, however, at least a decade or more behind the public opinion survey community, which has been able to convince journalists to report 5% confidence intervals and sample sizes as a matter of routine.

[2] Goertzel (2002) provides a semi-popular review of some examples of regression as "junk science" in sociology; Freedman (1991) a more extended critique with specific attention to the problem of causality.

[3] Two years ago I attended the international meetings of the "RC33"—quantitative methods—section of the International Sociological Association in Cologne, Germany. Despite its affiliation with the discipline of sociology, this organization is effectively the international equivalent of APSA Organized Section on Political Methodology. Not a post-modernist in sight, presentations featured slide after slide of dense, nearly incomprehensible linear algebra. And no one wore ties.

Yet with a very few exceptions, the mathematics were not those of the linear econometric models that dominate U.S. "political methodology." Instead, they were the mathematics of correspondence analysis, a clustering technique (Greenacre 1984).

One witnessed a deep methodological divide across the Atlantic. Not over the issue of the utility of sophisticated quantitative methods—the Europeans were every bit as sophisticated in this regard as the Americans—but in the form. Americans saw the world in linear models; the Europeans in clusters.

One possible explanation for this is institutional: econometrics developed in the U.S. whereas correspondence analysis was developed in France and popularized by a South African now teaching in Spain who has spent sabbaticals in the UK and Norway. Generations of U.S. graduate students learned econometrics because that is what their instructors were researching; generations of Europeans learned correspondence analysis for the same reason.

But there may also be a cultural issue here. The U.S. contains 280-million people in a relatively uniform social, political and economic environment, and so the homogeneity of the regression approach extracts little penalty. Europe contains about the same number of people scattered across a diverse set of legal systems, languages, and national histories, and therefore a clustering approach seems essential.

A second promising, if pedagogically more complex, development is the slow but on-going shift from frequentist to Bayesian approaches in social science statistics. Most policy-makers approach information as intuitive Bayesians. They have some idea about the value of a parameter relevant to policy—for example, will Israel be more likely to change its policies in response to the "stick" of diplomatic threats or the "carrot" of economic aid; same question for the Palestinian Authority—and they adjust the parameter value (and their confidence about the accuracy of the value) on the basis of new data—for example when Israel eases curfews in the West Bank after criticism by the European Union. The frequentist concept of parameter estimates as the idealized outcome of a infinite repeated sampling under a oftentimes implausible null hypothesis is, in contrast, completely alien to the policy community, an observation that will certainly not surprise anyone who has taught the standard canon of frequentist statistical methods to students interested in policy.

Bayesian methods are also likely to deal with one of the greatest problems with frequentist regression, the more or less random allocation of explanatory power among collinear independent variables. With an informed and possibly quite restricted prior—applied with the appropriate "garbage in, garbage out" caveats—one can often circumvent this problem. Yes, such priors open the possibility that the analysis will simply confirm the decision-maker's prejudices, but one still will be checking the model for predictive accuracy (we'll do out-of-sample assessment of our models now, right?). And since the frequentist alternative is "collinearity in, garbage out," at least we've done no harm.

So why aren't we exclusively using Bayesian methods? Largely because we've spent the last fifty years teaching frequentist methods as the norm and Bayesian methods as the extension, an extension requiring a fairly serious understanding of probability theory that can otherwise be finessed. In addition, the computational power (and software) required to effectively implement Bayesian methods has only become available in the last couple of decades. There may also be a subtle psychological resistance to Bayesian methods because they require additional assumptions to be stated explicitly in the priors, whereas frequentist estimation hides many assumptions in the null hypothesis.[1]

Which alternatives are best? This is an empirical question. As Achen (2002) has forcefully pointed out, the best model will not necessarily be the most complicated model or the model requiring the largest number of machine cycles to estimate. For example, I am continually puzzled by researchers who employ an elaborate logit analysis where a difference-of-means test would make the same point. Analysis-of-variance, one of the oldest, most straightforward, and most robust of the linear methods, is rarely even taught to graduate students in political science, much less utilized in published research. I am fairly confident, however, that for most questions, the trendiest current econometric technique—say some Bayesian randomized coefficient

[1] An example is the typical mistake that students (and not a few published articles) make in interpreting regression coefficients estimated on large samples. The t-test is above 1.96 so the coefficient is significantly different from zero. Therefore it is important! Well, no—all the t-test has said is that the coefficient is not *zero*, and in a world where everything is related to everything else, the true coefficient probably is not zero, and you've just managed to collect a sufficiently large data set to make this apparent. This does not mean that the coefficient is sufficiently large to make any theoretical or predictive difference. This interpretative problem occurs all the time in frequentist analyses because zero is the computationally-convenient null hypothesis, whereas a [good] Bayesian would only impose a prior of zero if there is a theoretical reason to do so.

multinomial tobit model, co-integrated, Gibbs-sampled, Monte-Carlo-Markov-chained, and super-sized with a Cauchy-distributed prior[1]—will not magically be the best possible approach.[2]

How should cases be chosen in small-N situations?

Or to phrase this more commonly, why are some cases more interesting than others? This issue is not uniquely linked to the problem of policy-relevant forecasting, but tends to be associated with it for two reasons. First, such forecasting tends to be directed at specific cases rather than a random or universal sample: for example policy-makers have little interest in the fact that 95% of all states have no plans for developing weapons of mass destruction; they are very interested in the remaining 5% who might be. Second, the information available for forecasting is often limited due to a situation changing quickly over time (for example assessing the likely policies of a regime that has recently taken power during a civil war), or the number of cases considered comparable is small (for example states governed by self-identified Islamic fundamentalists).

In 1994, King, Keohane and Verba made a strong argument that the same norms for case selection should apply in both large-N and small-N studies. Their assertion has subsequently been widely contested by researchers who use small-N approaches (see McKeown 1999, Ragin 2000, and *APSR* 1995). Small-N researchers, almost without exception, consider some cases to be more interesting than others.

But why? My assessment of this literature—particularly after trying to teach it—is that we have a lot of intriguing arguments, but *at the present time*, we do not have a theory for small-N case selection comparable in power to the sampling theories of large-N studies. Indicative of this problem is the fact that the closest thing we have to a canonical justification are appeals to John Stuart Mills' mid-19th-century formulation of the small-N "method of agreement" and "method of differences" (for example Van Evera 1997, Gerring 2001) that, ironically, Mills explicitly said should *not* be used in the study of social behavior (Ragin 2000: 204).

In Bohr's quadrant, the domain of pure research, we could simply avoid doing small-N work, and arguably there is a bias in the quantitative literature towards large-N problems. This is not an option in Pasteur's quadrant. Studies with the states of the European Union as the unit of analysis will remain small-N for the foreseeable future, as will studies of the deployment of U.S. peacekeeping troops and (hopefully) studies of states developing weapons of mass destruction. In situations of relatively rapid political and economic change—much of the Third World—the period over which it is reasonable to expect parameters (or even processes) to remain sufficiently constant to justify parameter estimation will involve only a small-N time-series. Furthermore most qualitative and intuitive forecasting is based on small-N situations such as historical

[1] hold the anchovies...

[2] This is not to say that one should not experiment with new methods. It is merely to say that methods relevant to economics are not automatically relevant to political science. For example, a great deal of effort was expended in the 1990s on applying co-integration models—originally developed in economics to deal with the random walk hypothesis posited by the theory of efficient markets—to time series such as Presidential popularity and event data that were bounded (by definition for popularity; by source limitations for event data) and therefore could not possibly have unit roots. Likewise many computationally-intensive estimation methods make only minor—and seemingly random—differences in the estimates of coefficients and their standard errors but nonetheless have been pursued far beyond the point where any new knowledge about the underlying data could be obtained.

analogies. In applied settings, answering the question of what makes a good case (or cases) is clearly critical.

The good news is that I think that progress is being made on this issue, in large part as a response to King, Keohane, and Verba's challenge. The Consortium for Qualitative Research Methods (http://www.asu.edu/clas/polisci/cqrm/; accessed 1 August 2002) appears to be developing norms and advanced training for small-N qualitative research in a manner similar to that provided earlier in the large-N quantitative realm by the summer courses of the ICPSR and the summer conference of the APSA Organized Section on Political Methodology. The gradual movement towards Bayesian approaches—which unlike frequentist approaches are conditioned on the data—should also place greater emphasis of the impact of individual cases. We do not have a systematic theory of small-N research at the moment, but could well in the near future.

How do we best impute or otherwise deal with missing information?

The final issue I will address is that of missing values. These complicate our analysis in at least three ways: First, they are missing. Second, they are missing non-randomly. Third, they are frequently missing non-randomly for the cases we would most like to study.

In Bohr's quadrant, we have the option of simply ignoring—with due documentation— the missing cases. Here's the model, it is estimated only with data from developed states and less-developed states that happen to have English-language newspapers;[1] you got a problem with that? In Pasteur's quadrant we do not have this luxury. Yes, we've got pretty good estimates of Osama bin Laden's popularity in New York, London, and Paris. But we really need to know his popularity in Kabal, Karachi, and Khartoum, or else work around the absence of that information.

While missing value problems are prevalent in any social science research, they are particularly problematic in prediction problems for two reasons. First, real-time prediction does not allow for slow and meticulous back-checking to fill in missing values, and even if the required information might *someday* be available (e.g. through declassified archival sources), this does not solve the problem in the present. Second, intuitive forecasters use a variety of heuristics—for example prior values or stereotypes—to get around the missing value problem. These may or may not be effective—as noted earlier, this is an empirical question—but it does raise yet another gap separating the day-to-day environment of the policy analyst and the idealized demands of the quantitative forecaster.

There are at least two different ways of dealing with this problem. First, we can experiment with models that are robust against—or can actually utilize the information contained in—missing values. OLS regression can do neither. Neural networks are a good example of a method known to be insensitive to missing values, although possibly at the price of having a diffuse parameter structure that is nearly impossible to interpret. ID3 is one of the best-known of the pattern recognition algorithms that can treat "missing" as simply another value, a value that possibly contains valuable information.

[1] This is a fairly accurate, if uncharitable, description of the sampling design of the CREON event data set (Hermann et al 1973).

A more sophisticated approach would involve imputation, which in recent years has been used with great effectiveness (accompanied by considerable political controversy) in demographic studies. Due to problems of inhomogeneity, it seems unlikely that imputation will be as effective in international and comparative politics as it has been in census-based studies in the U.S. and Europe, but it could still provide a substantial improvement over the existing approaches of throwing the case away, throwing the variable away, or (less frequently) replacing it with a mean value.

Conclusion

In comments on a draft of this paper, Michael Ward noted[1]

> By continuing to use only canonical tools and canonical data we just attempt to transform all our problems into ones that the boy genius, Karl Gauss, figured out two centuries ago. If we don't forecast and engage with real data that the various policy communities are using, we don't get any real feedback on how we are doing, i.e., on whether this actually makes sense. This is debilitating to the max, intellectually as well as organizationally. It isn't just that we could do this, as a discipline we can't afford not to do it.

Working in Pasteur's quadrant is not always easy because involves frequently dealing with two quite distinct professional cultures, at least in the foreign policy arena. For the international relations community, the two big problems policy makers have with academics are that academics don't care about policy—being content instead to look at "nice" (read: familiar and easily published) problems—and, quite ironically for political scientists, don't appreciate the complexity of the policy process. The biggest problems that academics have with the policy community are its obsession with secrecy, and the inflexibility in contracting for social science research.

As I noted in the introduction, there are plenty of really interesting, forward-looking questions out there in the contemporary world, and one does not have to sacrifice any academic or theoretical integrity to study them. There's more to the study of international politics than trying to second-guess—with the benefit of hindsight—Wilhelm II.[2]

This is not to suggest that a model, even one that has a solid theoretical justification, small standard errors and splendid out-of-sample predictions, will be immediately embraced by as a guide to policy community. We are dealing with politics, and in my experience the bureaucratic environment is far more complex and dysfunctional than even our most pessimistic models suggest. In the short term, the best one can hope for is a seat at the table—much as quantitative

[1] Email, 5 August 2002; used with permission.

[2] I wish this were a joke. A recent APSA meeting featured a roundtable devoted to "problems of strategy." It was held in a large room at a prime time; the participants were from distinguished universities, and quite a few State Department, Pentagon, and—presumably—CIA analysts eagerly attended. The discussion, alas, degenerated into a 45-minute debate over what some Prussian count may or may not have said on a sultry afternoon in July 1914, a question that (attractively, from an academic perspective) has no definitive answer and thus can be debated forever. The policy folks were frankly disgusted with what they viewed as a waste of time, and this roundtable was cited repeatedly the following day, in panel featuring policy-makers, as an example of why the policy community pays no attention to the academy.

economists, demographers, and epidemiologists have achieved—but not the director's chair. Look for a voice, not a veto.

On the policy side one finds a different set of barriers. Let me start with something is *not* a problem: Contrary to the belief held by many political methodologists, the policy community as a whole is not inherently hostile to quantitative methods.[1] Bureaucrats have allocated hundreds of millions of dollars for the collection and analysis of quantitative economic and demographic data; social policies dealing with education, public health, and welfare reform are now routinely assessed using statistical methods; and politicians spend tens of millions of dollars (and euros) on statistical opinion polling. The policy community is, however, hostile to models that don't work, and they've encountered quite a few such models in the past. In Leamer's immortal phrasing, there has been a lot of con in econometrics, and quantitative political forecasting is tainted by association if not practice.

More problematic to greater participation by academics is the bureaucratic cult of secrecy and, in the United States at least, inflexible contracting rules. Control of information is a factor in all bureaucracies, but this reaches ludicrous and self-defeating heights in the realm of foreign policy, as Moynihan (1998) and Jeffreys-Jones (2002) have documented in considerable detail. As a result, academics are reluctant to become involved (or prohibited from becoming involved) in policy work that might involve classified information, while models that do not use such information may be taken less seriously by the policy community.

A recent article in *Atlantic Monthly* notes that secrecy, far from being an asset to security, is a liability. "Kerckhoffs's principle" in cryptology states that "the system should not depend on secrecy, and it should be able to fall into the enemy's hands without disadvantage."[2] Mere journalists and tourists in Gorbachev's Soviet Union could readily observe the bread-lines, failing infrastructure, and drunken, demoralized military recruits. It took a secret analysis—one severely affected by ideological blinders, conducted beyond the auspices of even the CIA itself, and yet classified until 1992—to conclude that the USSR was an awesome superpower requiring hundreds of billions of dollars of additional U.S. defense spending. (Cahn and Prados 1993). In fact, for the long-term (e.g. 6-month) projections—as distinct from knowing where an aircraft carrier will be located tomorrow—the likelihood that secret information will reverse the conclusions of a well-informed model are close to zero.[3]

---

[1] Some *individuals* are hostile—they have no desire to be second-guessed by the geeks they used to beat up in junior high school—but one can route around such people.

[2] quoted in Charles McMann, "Homeland Insecurity" *Atlantic Monthly*, September 2002 (290:2) pg. 93.

[3] An additional—and decidedly non-trivial—problem has subsequently emerged: the supposedly high-quality secret information was thoroughly contaminated by systematically distorted data provided by U.S. agents such as Aldrich Ames and Robert Hanson who were actually working for the Soviet Union. Add to this the problem that computer technology in the FBI and CIA lags significantly behind that available in the private section—for example the FBI is unable to search their full-text data base using Boolean operators that have been available in LEXIS-NEXIS for almost two decades. And then add in the lack of field experience—how many CIA analysts, as distinct from operatives, have ever been to a refugee camp, or spent even 24 hours in a community where people live on less than $1 a day?—and the inability to speak frankly with anyone except other people in the intelligence community, and one is left with the distinct, if counter-intuitive, possibility that understanding, tools and quality of information available to analysts working *outside* the restricted world of classified information may be superior to those within that world, particularly for strategic forecasting. I have been told that even for short-term forecasting and monitoring, decision-makers are increasingly more likely to watch CNN than read a CIA briefing paper.

In the United States, government contracting for policy-oriented research is Byzantine and counter-productive. The experience of the State Failures Project, which took a very expensive route to some fairly trivial conclusions is, unfortunately, rather typical. Much of this problem stems from institutional constraints and procedures designed for the acquisition of pork bellies, not social science research,[1] though some of these constraints are due to past methodological misdeeds from social scientists in "Beltway bandit" consulting firms and academia (see Andriole and Hoople 1984; Laurence 1990).[2] On the positive side, institutions outside of the direct policy line such as the National Science Foundation and various not-for-profit research centers have considerably greater flexibility in funding, substantial experience with quantitative research in a variety of fields, and reasonably successful—if hardly flawless—quality control through the peer review system. In the near term, this may be the most promising route for funding policy-relevant work.

Would systematic forecasting be a significant improvement over the status quo? The possibility certainly cannot be ruled out. Tetlock, after all, found experts using intuitive methods fared only slightly better than chance, whereas Feder (1995) found the PoliCon expected utility models to be 90% accurate, and the best of my hidden Markov models had an accuracy of around 85% in forecasting conflict in the Balkans (Schrodt 2000). But Tetlock was asking experts to predict very difficult situations, whereas the hidden Markov models (and, I would guess, the expected utility models) gained much of their accuracy by making modal predictions in periods when little was changing. Considerably more work comparing alternative approaches needs to be done to determine what types of behavior can best be forecast, how those forecasts can be made, and how they should be assessed.

Finally, it is important to remember that the social science and policy cultures are not that far apart. In much of economic policy community—for example the U.S. Federal Reserve Board and the International Monetary Fund—one finds a much higher level of interaction between academics and the policy community than one finds in foreign policy. But even in the foreign policy area, professionals in both communities read the same newspapers, are usually working with similar concepts and theories, and have generally comparable education (in contrast to, say, the gap between the education of a political scientist and that of a surgeon). In my experience, one can usually quickly bring a conversation with a competent political analyst to a level of substantive detail comparable to that of a very good graduate seminar or professional panel. Academics clearly have the edge in discussing the latest esoteric academic theory (theories typically enjoying a life-span comparable to that of a gerbil); policy-makers have the edge on the latest nuances of the policy debate (information relevant for less than the life-span of a gerbil). But we can talk to each other.

---

[1] Although if the track record of weapons acquisition programs is any guide, the methods don't work very well for purchasing pork bellies either.

[2] A individual in the policy community who read a draft of this paper was generally sympathetic to my concerns but noted: "You make the contention that political science suffers from guilt by association to the econometric taint of 'bad modeling' (or as I like to put it: as being 'snake oil salespeople'). However, there are snake oil salespeople squarely within our ranks."

Acknowledgements

I would like to thank Robert Holt for suggesting and organizing a panel based on Pasteur's quadrant. This paper has benefited from helpful comments from Holt, John Freeman, Michael Haxton, Sara Mitchell, Naunihal Singh, and Michael Ward.

Bibliography

Abbey, Edward. 1975. *The Monkey Wrench Gang.* New York: Avon.

Achen, Christopher H. 2002. "Toward A New Political Methodology: Microfoundations and ART." *Annual Review of Political Science* 5: 423-450.

American Political Science Review. 1995. "Review Symposium: The Qualitative-Quantitative Disputation." *American Political Science Review* 89,2*:* 454-481 (June 1995)

Andriole, Stephen J. and Gerald W. Hopple, 1984, "The rise and fall of events data: From basic research applied use in the U. S. Department of Defense." *International Interactions* 10, 3-4: 293-310.

Arthur, Brian W. 1994. "Inductive reasoning and bounded rationality: the El Farol problem." *American Economic Review: American Economic Association Papers and Proceedings* 84: 406-411.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books

Axelrod, Robert. 1997. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration.* Princeton, NJ, Princeton University

Beck, Nathaniel, Gary King, and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94, 1: 21-36.

Bueno de Mesquita, Bruce. 2000. *Principles of International Politics.* Washington: CQ Press.

Burgess, Philip M. and Raymond W. Lawton. 1972. *Indicators of International Behavior: An Assessment of Events Data Research.* Beverly Hills: Sage Publications.

Cahn, Anne Hessing and John Prados. 1993. "Team B: The Trillion Dollar Experiment." *Bulletin of the Atomic Scientists* (http://www.thebulletin.org/issues/1993/a93/a93Teamb.html; accessed 5 August 2002)

Casti, John. 1989. *Alternate Realities: Mathematical Models of Nature and Man.* New York: Wiley.

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

Cockburn, Andrew. 1983. *The threat : inside the Soviet military machine*. New York, N.Y. : Random House.

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

Davies, John L. and Ted R. Gurr, eds. 1998. *Preventive Measures: Building Risk Assessment and Crisis Early Warning*. Lanham, MD: Rowman and Littlefield.

Esty, Daniel C., Jack A. Goldstone, Ted R. Gurr, Pamela Surko, and Alan N. Unger. 1995. *State Failure Task Force Report.* McLean, VA: Science Applications International Corporation.

Esty, Daniel C., Jack A. Goldstone, Ted R. Gurr, Barbara Harff, Marc Levy, Geoffrey D. Dabelko, Pamela Surko, and Alan N. Unger. 1998. *State Failure Task Force Report: Phase II Findings.* McLean, VA: Science Applications International Corporation.

Feder, Stanley. 1995. "Faction and Policon: New ways to analyze politics." In *Inside CIA's Private World: Declassified Articles from The Agency's Internal Journal, 1955-1992*. ed. H. Bradford Westerfield. New Haven:Yale University Press. Pp. 274-292 (available at http://www.diiusa.com/cia.html; accessed 14 August 2002)

Freedman, David. 1991. "Statistical models and shoe leather." *Sociological Methodology* 21: 291-313.

Gaddis, John Lewis. 1992a. "International Relations Theory and the End of the Cold War." *International Security* 17 (Winter): 5-58.

Gaddis, John Lewis. 1992b. *The United States and the end of the cold war : implications, reconsiderations, provocations.* New York: Oxford University Press.

Gerner, Deborah J. and Philip A. Schrodt. 2002. "Taking Your Academic Expertise Public: Lessons Learned from Responding to the 11 September Crisis." *International Studies Perspectives* 3,2: 221-229.

George, Alexander. 1993. *Bridging the Gap: Theory and Practice in Foreign Policy.* Washington: U.S. Institute of Peace Press.

Gerring, John. 2001. *Social Science Methodology*. Cambridge University Press.

Goertzel, Ted. 2002. "Myths of Murder and Multiple Regression." *Skeptical Inquirer* 26, 1 (January/February): 19-23.

Goldstein, Joshua S., Jon C. Pevehouse, Deborah J. Gerner, and Shibley Telhami. 2001. "Dynamics of Middle East Conflict and US Influence." *Journal of Conflict Resolution* 45, 5: 594-620.

Greenacre, Michael J. 1984. *Theory and Applications of Correspondence Analysis*. New York: Academic Press.

Gurr, Ted R., and Mark Irving Lichbach. 1986. "Forecasting Internal Conflict: A Competitive Evaluation of Empirical Theories." *Comparative Political Studies* 19 (April): 3-38.

Hermann, Charles, Maurice A. East, Margaret G. Hermann, Barbara G. Salmore, and Stephen A. Salmore. 1973. *CREON: A Foreign Events Data Set.* Beverly Hills: Sage Publications.

Hopple, Gerald W., Stephen J. Andriole, and Amos Freedy, eds. 1984. *National Security Crisis Forecasting and Management*. Boulder: Westview.

Jeffreys-Jones, Rhodri. 2002. *Cloak and Dollar: A History of American Secret Intelligence.* New Haven, CT: Yale University Press.

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

King, Gary. 1986. “How Not to Lie with Statistics.” *American Journal of Political Science* 30,3: 666-687.

King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry*. Princeton: Princeton University Press.

King, Gary and Langche Zeng. 2001. “Improving Forecasts of State Failure.” *World Politics* 53, 4: 623-58.

King, Gary and Will Lowe. 2001. “An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design.” http://gking.harvard.edu/preprints.shtml, accessed 16 August 2002.

Kolodner, Janet, ed. 1988. *Proceedings of the DARPA Workshop on Case-Based Reasoning.* Palo Alto: Morgan Kaufmann.

Kolodner, Janet. 1993. *Case-Based Reasoning.* San Mateo, CA: Morgan Kaufmann

Kuran, Timur. 1989. “Sparks and prairie fires: A theory of unanticipated political revolution.” *Public Choice*, 61 (April): 41-74.

Kuran, Timur. 1995. “The inevitability of future revolutionary surprises.” *American Journal of Sociology*, 100 (May): 1528-1551.

Laurance, Edward J. 1990. “Events Data and Policy Analysis.” *Policy Studies* 23:111-132.

Leamer , Edward E. 1983. “Let’s Take the Con out of Econometrics.” *American Economic Review* 73, 1 (March): 31-43.

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

McKeown, Timothy J. 1999. “Case Studies and the Statistical Worldview: Review of King, Keohane and Verba.” *International Organization* 53,1: 161-190.

Merritt, Richard L., Robert G. Muncaster and Dina A. Zinnes, eds. 1993. *International Event Data Developments: DDIR Phase II.* Ann Arbor: University of Michigan Press.

Morgenthau, Hans J. 1973. *Politics Among Nations*. 5th edition. New York: Knopf.

Moynihan, Daniel Patrick. 1998. *Secrecy.* New Haven, CT: Yale University Press.

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

O'Neill, Barry. 1994. "Game Theory Models of Peace and War." In *Handbook of Game Theory*, vol. 2., edited by Robert J. Aumann and Sergiv Hart. New York: Elsevier.

Pevehouse, Jon C., and Joshua S. Goldstein. 1999. “Serbian Compliance or Defiance in Kosovo? Statistical Analysis and Real-Time Predictions.” *Journal of Conflict Resolution* 43, 4: 538-546.

Ragin, Charles C. 2000. *Fuzzy-Set Social Science*. Chicago: University of Chicago Press.

Ray, James Lee and Bruce Russett. 1996. “The Future as Arbiter of Theoretical Controversies: Predictions, Explanations and the End of the Cold War.” *British Journal of Political Science* 26,4 (October): 441-470.

Richardson, Lewis F. 1960a. *Statistics of Deadly Quarrels*. Chicago: Quadrangle Books.

Richardson, Lewis F. 1960b. *Arms and Insecurity*. Chicago: Quadrangle Books.

Rupesinghe, Kumar, and Michiko Kuroda, eds. 1992. *Early warning and conflict resolution.* New York: St. Martin's Press.

Schmeidl, Susanne, and Howard Adelman, eds. 1998. *Early Warning and Early Response.* New York: Columbia University Press-Columbia International Affairs Online.

Schrodt, Philip A. 2000. "A Test of Hidden Markov Models as Predictors of Conflict in the Balkans, Central Asia and West Africa." Paper presented at the American Political Science Association, Washington, September 2000. (http://www.ukans.edu/~keds/papers.html; accessed 5 August 2002)

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

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

Tetlock, Philip. 1999. "Theory-Driven Reasoning about Possible Pasts and Probable Futures in World Politics." *American Journal of Political Science* 43,2 (April): 335-366

Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science.* Ithaca: Cornell University Press.

Whelan, William. 1998. "The U.S. Agency for International Development's Famine Early Warning System." In *Preventive Measures: Building Risk Assessment and Crisis Early Warning*, ed. John L. Davies and Ted R. Gurr. Lanham, MD: Rowman and Littlefield. Pp. 194-202.

Wlezien, Christopher. 2001. "On Forecasting the Presidential Vote." *PS: Political Science and Politics* 34,1 (March). (http://www.apsanet.org/PS/march01/wlezien.cfm; accessed 1 August 2002)

Zinnes, Dina A. 1976. *Contemporary Research in International Relations.* New York: Free Press.