Expert Judgment versus Statistical Models: Explanation versus Prediction

Gregory A. Caldeira

One of my favorite political scientists (and mentors in graduate school) defined a great idea as one he would have liked to have had himself. The Supreme Court Forecasting Project is a great idea, one I wish I had hatched. It is such a good idea that it is a little surprising no one had attempted it earlier.[1] I commend this team and urge them, as I shall say in more detail later, to continue, expand, and perhaps redirect this undertaking. That this team combines legal academics and political scientists is a happy sign of the more general recognition in both disciplines of the great things we have to offer each other.

Andrew Martin and colleagues seem, at least to me, to be a little confused about the purpose(s) of their project. On the first page of their contribution to this symposium, they claim that the best test of an explanatory theory is its ability to predict future events or outcomes; therefore, political scientists and legal academics should attend to future and not just past cases. Yet, as they recognize, this seldom happens. This retrospective orientation puzzles them because, they say, in other areas of political science, scholars do attempt to make forecasts. Later in their essay, however, the team notes that "these variables [in the statistical model] were not chosen for explicitly theoretical reasons. . . . This forecasting exercise was not one of theory testing, but rather one of determining whether systematic patterns can be uncovered using these variables."[2] This statement comes closer to the mark, but this team has much more on its hands than mere prediction for its own sake or comparison of expert opinion to a statistical model.

First I want to focus on their opening assertions about the role and prominence of prediction in political science; for the moment I leave the lawyers to our law professors. Contrary to their claim, political scientists spend little time or effort on forecasting. Those who do live in the world of consulting or applied policy research; from what I know, they are not interested in theory and theory testing, even though they rely on theories or pieces of them to improve the accuracy of prediction. The two examples the team cites are not really apt. Predicting the outcome of presidential elections is not a central concern of electoral studies; indeed, some regard it an academic parlor trick rather than a serious scholarly undertaking. It does not rely on well-specified theory or explanation and is often openly ad hoc in its specifications. It purports to predict outcomes, not to explain them; but, as the 2000 election demonstrated, its value as a forecasting model is unclear. Keith Poole and Howard Rosenthal's work repeatedly demonstrates that nearly all of the roll-call behavior of members of Congress, except for short historical periods, can be explained by a single left-right dimension.[3] It is thus explicitly retrospective in its orientation. To be sure, from time to time, Poole has generated forecasts of roll calls, as in the cases of the Robert Bork's nomination to the Supreme Court in 1987 and President Clinton's impeachment in 1998 and 1999. These are neat by-products of what is otherwise a theoretical undertaking.

Most political scientists are theoretically-oriented social scientists. We strive to develop persuasive accounts of or stories about political phenomena we regard as important. One nonreactive indicator of our priorities is that one of today's most frequently cited works on the strategy of research, *Designing Social Inquiry,* does not discuss prediction.[4] Applied political scientists are not in the mainstream of our disciplinary conversation, as the table of contents of any major journal readily demonstrates. For all of the palaver about the "public role" of political scientists, most of us are content to go about the work of explaining things without concern for whether we can use our explanations to forecast the future. Practicing lawyers, of course, value dearly, as do their clients, the ability to predict the outcomes of cases. This is in essence what they do in advising clients and in formulating legal strategy. As Justice Holmes famously wrote, "[T]he law is a prediction of what judges in fact will do." Perforce, practicing lawyers

Gregory A. Caldeira is Distinguished University Professor in the Department of Political Science at Ohio State University and was Jack N. Pritzker Distinguished Visiting Professor of Law at Northwestern University Law School when he wrote this commentary (caldeira.1@osu.edu).

https://doi.org/10.1017/S1537592704040526 Published online by Cambridge University Press

are forecasters. And to the extent legal academics engage in consulting, they, too, are forecasters.

Prediction, or goodness-of-fit, is only one criterion for judging the quality of explanation. Just as important—perhaps more important—are an explanation's elegance, parsimony, verisimilitude, fit with the data, or persuasiveness. Our research often involves trade-offs among these criteria. To gain predictive power, we may have to sacrifice parsimony, elegance, and persuasiveness. Indeed, Martin and associates make precisely this sort of trade-off in formulating their statistical model. Whether this tradeoff is beneficial depends on one's goals.

The team's statistical model and specification, as they explain, is a function of convenience and availability. It is inductive rather than deductive, and it does not provide an explanation of what the Court or individuals do. To be sure, one can examine the independent variables in the model and think of theoretical rationales for their presence, but doing so is an ad hoc exercise. One of the difficulties in their statistical model is that the variables specified are at a high level of generality—circuit of origin, issue area, type of petitioner, type of respondent, ideological direction of decision in the lower court—and therefore cannot discriminate among different cases with similar values on the independent variables, cases that may be very different in factual and legal circumstances not specified in their statistical model. No doubt, the variables in their model provide a fair amount of discrimination, or spread, among cases, and it might be sufficient for their purposes; whether it is for forecasting, I do not know. This problem, of course, works against accurate predictions.

Martin and his colleagues express surprise at the superior performance of the statistical model in predicting outcomes. Or perhaps they are surprised that the experts do better than the model in predicting individual votes. If their surprise is about the former, then I am a little surprised. Scholarship on "judgment and decision" has predicted such a result as early as 1954, in Paul Meehl's well-known study of clinicians.[5] "Experts" and "novices" are, in varying degrees and under varying circumstances, victims of the same sorts of errors in judgments we see in all human decision making. Richard Nisbett and Lee Ross devote an entire chapter to "prediction" and detailed discussion of various sorts of errors.[6] On the question of the predictive capacities of experts and models, they state: "Not only clinicians, but human beings in general, make predictions inferior to those of actuarial formulas. This result has been obtained under almost all conditions studied in which the information employed by the human judges and statistical formulas is at all diagnostic of the outcomes."[7] And they make two other points of great interest in this context: "Human judges are not merely worse than optimal regression equations; they are worse than almost any regression equation. . . . The judges typically are not amateurs, but professionals, making predictions in domains of which they have special knowledge and often years of training, and in which they commonly have had repeated opportunities for feedback on the accuracy of their predictions."[8] For example, experts and novices alike are prone not to take into account base-rate data, probabilities, and information about the central tendencies in the matter predicted; and to fail to recognize "regression effects" in the outcome. In many instances, experts do worse—much worse—than novices, precisely because they have so much information; they have commitments to ideas and conclusions and confidence in their judgments, both of which make them resistant to discordant information. From the team's description of their experts, it seems very likely that this group contained many people who had taken positions in court and in print on issues in their specialty and who may have had strong ideological commitments to these positions. One hypothesis, then, is that the greater the expertise among the experts, the more likely we are to observe errors.

That the team's expert judges did not do nearly as well as the statistical model in predicting case outcomes is therefore entirely in keeping with nearly fifty years of experiments and other studies on judgment and decision making. In light of the literature, that the experts did slightly better than the model in predicting individual votes seems all the more impressive. The clear expectation from the literature, together with the performances of the model and experts, raises the question of just how accurate a forecast must be to be considered a good forecast. The statistical model was accurate in 75 percent of the cases; the experts, 59 percent. Ought we to expect better of the model? Perhaps. I would like to see them discuss the standards by which we ought to judge the quality of forecasts. This leads back to the high level of generality of the model's independent variables; many of the errors surely result from the model's failure to include more nuances of particular cases.

Martin and associates speculate on and investigate several sources of errors. They pinpoint cases on technical matters of (1) jurisdiction and procedure and (2) federalism. The former are hard to predict on ideological grounds, the basis of the model, because decisions on jurisdiction and procedure can cut in both directions and do not inevitably hurt or help liberals or conservatives; the latter are difficult because they usually involve two or more dimensions, as in state criminal procedure. Presumably, if we looked further into the team's cases and predictions, we would see many other issues involving multiple dimensions or ambiguous ideological consequences. *Breuer v. Jim's Concrete,* affirmed 9–0 by the Court—predicted accurately both in votes and outcomes by the experts, and wrongly both on outcome and votes by the model—provides an excellent example of the drawbacks of the statistical model. This case, on my reading of the opinion, looks like one in which the Court simply wanted to resolve an intercircuit conflict: the Eleventh Circuit was in agreement with the Court and the conflicting cases in the other direction, favoring the Fair Labor

https://doi.org/10.1017/S1537592704040526 Published online by Cambridge University Press

Standards Act claimant, came from other circuits. All in all, it has the look of an "easy" case, one on which liberals and conservatives agree because the statutory situation is fairly clear. Oddly, then, on this easy issue, the model was completely befuddled. Is this because the conservative ideological direction of the lower court and the overwhelming tendency of the Court to grant cases and reverse them together pushed the model in the wrong direction?

One other source of errors comes immediately to mind. The experts had at their disposal the lower court opinion and any other materials they wished to examine, including the petitions for certiorari; briefs on the merits; and briefs of amici curiae, if present; and the joint appendix, if they were really energetic. Cert petitions typically present a number of issues; some present a long list of issues. Even seasoned practitioners, despite the Court's clear signals, will go beyond the central issue in the case just in case the Court might have an interest in one or another. The Solicitor General's Office is presumably the most careful in restraining the tendency to throw everything into a petition; but even a cursory perusal of last term's petitions for the United States shows the Solicitor General regularly offering up a variety of issues. Yet seldom does the Court tell the parties which issues it has targeted. Since the Court ultimately decides cases on one or a limited set of issues, it regularly engages in what has been called "issue suppression."[9] Expert judges do not know precisely which issues the decision will turn on and therefore must predict the issue(s) on which the Court will base its decision before they can attempt to predict the outcome.

To return to the question of this study's purpose, at the outset I pointed out the ambiguity about the team's goals: predictive capacity as ultimate test of an explanation; statistical forecasting not related to theory-testing; and a competition between lawyers and the machine. But they have much better justifications for doing this research. This study and what I hope will be future extensions and expansions is, in my view, more important for what it can tell us about the qualities of experts and statistical models and for the estimates of the justices' ideal-points[10] than it is for its competition between lawyers' and statistical models. Martin and his associates have in their hands a rich lode of information for assessing under what conditions and to what extent experts and the model make mistakes. Why not take features of the cases, of which Spaeth's database has legions, and use likely candidates to account for mistakes? Thus, for example, cases with multiple issues might be harder to predict. It is unfortunate that they have so few experts for each case because it makes consensus estimates unreliable, and it restricts the variance in the level of expertise. Moreover, matching up experts only with cases in their designated fields of expertise means that Martin and colleagues do not have a good way of calibrating across experts and cases. If these legal experts are like other experts, we might well see those with lesser expertise doing better than those with more. If somehow they could persuade experts to make predictions on all cases, this would give them a very rich set of data and permit them to do all sorts of interesting analyses.

Martin and associates of course have data for only one term for their experts, but they have eight terms for their statistical model. Why not start with an initial set of estimates from the first term and use them to predict the next, take the estimates from the two terms and predict the third, and so on? Doing so would tell them how quickly the statistical model improves in predictive capacity and it would as a by-product tell them about how many cases the model needs to produce good estimates of ideal-points.

Two other issues merit brief discussion. First, we see that the team excluded eight cases, most for good reason, but I wonder about their having excluded the three with coding difficulties.[11] These cases ought, it seems to me, to cause great difficulty for the statistical model but less trouble for the experts; I question the equity of putting these cases aside. Second, as they explain, the team used "circuit of origin" as one of its predictive variables and coded the state cases based on the circuit in which they are located. This may well have caused difficulties because states and the state high courts within circuits can differ considerably, as in the case of the Ninth Circuit, which ranges from Alaska and Arizona on the political right to Hawaii on the left. Yet, today, among the circuits, the Ninth would be the furthest to the left on balance. Therefore, coding cases from the Arizona Supreme Court as originating from the Ninth Circuit probably does not help the statistical model.

Finally, statistical models, expert judges, or the combination thereof, are not the only ways of forecasting future events of the behavior of individuals or institutions. One of the most successful and intriguing methods uses the device of a stock market to aggregate the information from individual investors and thereby predict some outcome. Its rationale derives from the notion of efficient markets: if people invest money, they will have an incentive to gather accurate information and, when aggregated, an accurate prediction will result. The Iowa Political Stock Market, run for more than a decade from the University of Iowa's College of Business, is a brilliant example of a futures market at work.[12] It has done well in predicting the outcomes and margins of the three presidential elections and relatively well on congressional elections. Investors, both at the University of Iowa and around the world, purchase stock in a candidate; then online trading in shares begins. Share prices respond to political events, the economy, and so on. By tracing the relative value of candidates' share values, which the market transforms into a sum of 100 percent, one can see how well candidates are doing. Once the presidential race hits the summer and early fall before the election, the share prices are amazingly stable and close to the final outcome, in contrast to polls, which bounce all around, by day. This is a very powerful instrument and one might apply it to the Supreme Court. Parenthetically, last year the Department

https://doi.org/10.1017/S1537592704040526 Published online by Cambridge University Press

of Defense, in response to political pressure, gave up on a stock market it had in the works to predict terrorist events. It would have worked very much like the Iowa Market.[13]

Lest my critical comments mislead, I want to reemphasize my great excitement about this project and its future extensions. They will have much to tell us about expert judgments, statistical models for forecasting, estimation of ideal points, and about the Supreme Court itself.

Notes

A complete reference list for the entire symposium appears on pp. 791–93, below.

1 Attempting to predict Supreme Court decisions has a fine precedent and pedigree in Harold J. Spaeth's syndicated columns of the 1960s, wherein he would use his statistical model, together, no doubt, with his immense knowledge of the Court, to venture a prediction. He did quite well, as I recall.
2 Martin et al. 2004, 762.
3 Poole and Rosenthal 1997.
4 King, Keohane, and Verba 1994.
5 Meehl 1954.
6 Nisbett and Ross 1980.
7 Ibid., 140.
8 Ibid., 141.
9 McGuire and Palmer 1995.
10 Martin and Quinn 2002.
11 Martin et al. 2004, note 10.
12 See Forsythe et al. 1992; Berg et al. 2000.
13 Meirowitz and Tucker 2004.

https://doi.org/10.1017/S1537592704040526 Published online by Cambridge University Press