Making a Case for (Imperfect) Statistical Modeling as the Basis for Genocide Early Warning

Jay Ulfelder, Ph.D.

October 5, 2011

For today's seminar, I was asked to talk about the promise and potential challenges of using statistical models to provide early warning of genocide and mass killing. I can sum up my views on this topic as follows:

- Statistical modeling should be able to give us early warnings of genocide and mass killing that are more accurate than expert judgments.

- Even statistical forecasts will never be as sharp as we would like, however, to allow us to pinpoint preventive action in a world of limited resources.

- It's still worth trying.

- There are a few low-cost ways we might improve on previous warning tools.

I'll start at the top of that list. For me, the promise of statistical modeling is clear. Thanks to Philip Tetlock's very careful work, we now know that, when trying to anticipate trends in international politics, expert judgment is not very reliable. Even simple statistical models will usually provide more accurate forecasts than subject-matter experts can. From personal experience, I also know that well-chosen statistical methods will usually lead to models that can outperform the simple extrapolatory ones Dr. Tetlock used as his benchmark. So, if our goal is to provide warnings that are as accurate as possible, we have good reason to think that statistical models are the right tool for the job.

That said, statistical models are not crystal balls, either. As producers of early warning, we want it all. We want to avoid being surprised by important events without ever “crying wolf.” To approach that ideal, our forecasts have to do an excellent job distinguishing the situations where relevant events will occur from the ones where they will not. A single “surprise” can be very costly, but lots of false alarms make it difficult to distribute preventive resources in an effective way. In technical terms, what we want is a forecasting system that produces no false negatives while also minimizing the ratio of true positives to false positives.

Unfortunately, that kind of precision is pretty much impossible to achieve, and that near-impossibility is not just a function of the complexity of human social behavior—as if that weren't enough. As Kaiser Fung nicely shows in his book *Numbers Rule Your World*, the same difficulties bedevil detection systems in many fields, from testing for illegal drug use or rare diseases to the polygraph tests used to screen people for security clearances. Imprecision is an unavoidable consequence of the fact that major political crises are rare, and it is virtually impossible to predict rare events of any kind as precisely as we would like.

An example helps to show why. Imagine that there are 150 countries worldwide, and that five of those countries suffer onsets of civil war each year. Those assumptions give us an annual incidence (or onset rate) of 3 percent (5/150 = 0.03)—slightly higher than the average rate of civil-war onset in the real world over the past several decades, but close, and in round numbers that make the ensuing calculations easier to follow.

When an event is rare, it's easy to make predictions that are quite accurate by simply saying that the event will never happen. In our civil-war-onset example, that blind forecast would give us an impressive accuracy rate of 97 percent. Every year, we would make 145 correct predictions and only miss five. Of course, those predictions would not be very useful, because they wouldn't help us at all with our goals of avoiding surprises and targeting preventive action.

To do better, we have to build a system that uses information about those countries to try to distinguish the ones that are likely to slide into civil war from the ones that are likely to remain peaceful. Now let's imagine that we have done just that with a statistical model that estimates the probability of a civil-war onset in every country every year. To convert those estimated probabilities into sharp yes/no predictions, we need to set a threshold, where values above the threshold are interpreted as predictions that the event will occur and values below it are interpreted as predictions that it will not. To think about how useful those predictions are, statisticians sometimes identify the threshold that produces equivalent error rates in both groups (events and non-events, or positives and negatives) and then use the accuracy rate that results from that "balancing" threshold as a summary of the model's predictive power.

Now, back to our example. Let's imagine that we've developed a model of civil wars that sorts countries into high-risk (predicted onset) and low-risk (predicted non-onset) groups with 80-percent accuracy when that balancing threshold is employed, on par with the current state of the art in forecasting political instability. Using that model, the high-risk group would include 33 countries each year: 80 percent of the five onsets (four countries), and 20 percent of the 145 non-onsets (29 countries). Of those 33 countries identified as high risk, only four (12 percent) would actually experience a civil-war onset; the other 29 would be false alarms, or what statisticians call "false positives." Meanwhile, one of the five civil-war onsets would occur in the set of 117 countries identified as low risk.

If you're a decision-maker looking at that list of 33 high-risk countries and trying to decide how to allocate resources in an effort to prevent those wars or mitigate their effects, you are probably going to find the length of that high-risk list frustrating. The large number of high-risk cases means you have to spread your preventive actions thinly across a large group, and the one conflict your warnings miss could prove very costly as well. Your odds of hitting impending crises with your preventive efforts are much better than they would be if you picked targets at random, but they're still not nearly as focused as you'd like them to be.

Now let's imagine that some breakthrough—an improvement in statistical methods, an improvement in our data, or an improvement in our understanding of the origins of civil wars—allows us to develop a new model that is 95-percent accurate at that balancing threshold. In light of the complexity of the processes generating those events and the (often poor) quality of the data we use to try to forecast them, that's an achievement I don't expect to see it in my lifetime, but let's consider it anyway for the sake of argument. At 95-percent accuracy, we would largely have solved the problem of false negatives; only once in a great while would a war break out in a country we had identified as low risk. Meanwhile, though, our high-risk group would still include 12 countries each year, and only five of those 12 countries (42 percent) would actually suffer war onsets. In other words, we would still have more false positives than true positives in our high-risk group, and we would still have no way of knowing ahead of time which five of those dozen countries were going to be the unlucky ones. The imprecision is greatly reduced, but it's hardly eliminated.

When the resources that might be used to respond to those warnings are scarce, those false positives are a serious concern. If you can only afford to muster resources for serious preventive actions in a handful of countries each year, then trying to choose which of those 33 countries—or, in an unlikely world, that dozen—ought to be the targets of those actions is going to be a daunting task.

Unfortunately, that uncertainty turns out to be an unavoidable product of the rarity of the events involved. We can push our prediction threshold higher to shrink the list of high-risk cases, but doing that will just cause us to miss more of the actual onsets. The problem is inherent in the rarity of the event, and there is no magic fix. As Kaiser Fung puts it (p. 97), "Any detection system can be calibrated, but different settings merely redistribute errors between false positives and false negatives; it is impossible to simultaneously reduce both."

If this is the best we can do, then what's the point? Well, consider the alternatives. For starters, we might decide to skip statistical forecasting altogether and just target our interventions at cases identified by expert judgment as likely onsets. Unfortunately, those expert judgments are probably going to be even less accurate than our statistical forecasts, so this “solution” only exacerbates our problem.

Or, we could take no preventive action at all and just respond to events as they occur. If the net costs of responding to crises as they happen are roughly equivalent to the net costs of prevention, then this is a reasonable choice. Maybe responding to crises isn't really all that costly; maybe preventive action isn't effective; or maybe preventive action is potentially effective but also extremely expensive. Under these circumstances, early warning is not going to be as useful as we would like.

If, however, any of those last statements are false—if waiting until crises are already underway is very costly, or if preventive action is (relatively) cheap and sometimes effective—then we have an incentive to use forecasts to help guide that action, in spite of the lingering uncertainty about exactly where and when those crises will occur.

Even in situations where preventive action isn't feasible or desirable, reasonably accurate forecasts can still be useful if they spur interested observers to plan for contingencies they otherwise might not have considered. For example, policy-makers might be concerned about the risk of mass killing in another country but still fail to plan for that event because they don't expect it to happen any time soon. A proven forecasting model which identifies that country as being at high or increasing risk of mass atrocities might encourage those policy-makers to reconsider their expectations and, in so doing, lead them to prepare better for that event.

The bottom line is this: even though forecasts of rare political events are never going to be as precise as we'd like, they can still be accurate enough to be helpful, as long as the events they warn about are ones for which prevention or preparation stand a good chance of making a (positive) difference.

To my mind, genocide and mass killing are just that type of event. These aren't natural disasters, in which the dollar costs of infrastructure repairs are being weighed against the dollar costs of preventive engineering. These are mass murders, in which no amount of money can repair the damage done. The resources for, and effectiveness of, preventive action will inevitably be limited, but the immeasurable costs of mass killing mean that it behooves us to do whatever we can to make the most of what we've got. Better forecasts are one thing we can do.

How might we improve on available statistics-based early-warning efforts? I can think of a few ways to try to advance the state of the art in this area.

- *Aim wider.* The most prominent effort to develop a statistical model to predict genocides—a Political Instability Task Force (PITF) project led by Barbara Harff— defined the event of interest in such a way that it had only occurred a few dozen times in the past half-century. When the phenomenon about which we're trying to warn is that rare, statistical analysis will usually produce less reliable results. A more recent PITF-funded project led by Ben Valentino broadened the event of interest from genocide/politicide to mass killing, giving us a larger set of cases on which to train our models. In addition to its statistical advantages, a broader definition also avoids political hassles that sometimes arise from the use of the term "genocide."

- *Aim lower.* For similar reasons, it might also behoove us to focus our early-warning efforts on smaller-scale atrocities that might or might not represent the start of a larger campaign of killings. I don't think it's possible or even useful to try to forecast specific incidents, and I know from personal experience that it's also extremely hard to predict variations in the scale of violence against civilians over time. What my experience tells me could work well is forecasting the occurrence of significant atrocities—say, more than 25 noncombatants killed in a calendar quarter, or more than 100 killed in a year—in cases where such killings were not already occurring. A threshold in that neighborhood would give us a larger set of cases on which to train. At the same time, it would also shift the conceptual emphasis from mass killing to what we might call incipient mass killing, a shift that could leave more time for efforts to avert wider atrocities.

- *Use better statistical methods.* To my knowledge, most of the studies that have tried to develop statistical models to help predict onsets of mass killing have used logistic regression models. From structured comparisons, we know that other statistical techniques can probably lead us to more accurate forecasts. Bayesian model averaging is my current favorite, but there are other strong candidates, too, such as Bayesian ensemble forecasting and random forests. There are no major barriers to entry here; all of these techniques can be implemented in open-source software.

- *Focus on forecasting.* To my knowledge, all of the published statistical models of mass killing have emphasized theory testing as much if not more than predictive accuracy. The problem is that the research design appropriate for theory testing is not the best design for developing a forecasting tool. Theory testers are primarily interested in the marginal effects of specific variables. Forecasters are primarily interested in the predictive power of specific models (or combinations of models). To get a forecasting tool that works as well as possible, it's important to check and compare models as you go according to their out-of-sample predictive power. That kind of cross-validation usually leads to more accurate forecasts, and it has the added bonus of giving you some information up front about the real-world predictive power of the tool you're using.