Predicting Local Violence[*]

Robert Blair Christopher Blattman Alexandra Hartman[†]

December 16, 2014

Abstract

This paper tests the feasibility of local-level violence forecasting. We apply standard prediction models to new data from 242 Liberian communities to investigate whether it is to possible to predict outbreaks of local violence with sensitivity and accuracy, even with limited data. We first trained our models to predict communal, extrajudicial, and criminal violence in 2010 using 2008 risk factors. We then made forecasts of violence in 2012, before collecting data. Our model predicts up to 88% of actual 2012 violence. This comes at the cost of many false positives, for overall accuracy of 33 to 50%. Policy-wise, states and peacekeepers could use such predictions to prevent and respond to violence. The models also generate new stylized facts for theory to explain. In this case, ethnic cleavages and power-sharing predict violence, while economic variables typically do not. We illustrate how forecasting can be widely more applied to micro-level conflict data.

[*] *Acknowledgements*: This research was supported via grants from the United Nations Peacebuilding Fund in Liberia (via UNHCR), Humanity United, the MacMillan Center for International Affairs at Yale University, and the World Bank's Italian Children and Youth Trust Fund. For comments we thank Peter Aronow, Kate Baldwin, Neal Beck, Grant Gordon, Guy Grossman, Stathis Kalyvas, Elisabeth Wood, and seminar participants at Yale University, the 2011 Northeast Methodology Program meeting, a 2014 APSA Annual Meeting panel, and a 2014 BREAD conference. Bryan Plummer, Mathilde Emeriau, Tricia Gonwa, Rebecca Littman, Benjamin Morse, Grant Gordon, Daniel Masterson, Johnny Ndebe, Richard Peck, Gwendolyn Taylor, Prince Williams, and John Zayzay provided superb research assistance.

[†] Blair, Yale University, Political Science, 77 Prospect St, New Haven CT (robert.blair@yale.edu); Blattman: Columbia University, SIPA and Political Science, 420 W 118th St, New York, NY 10027 (chrisblattman@columbia.edu); Hartman: Yale University, Political Science, 77 Prospect St, New Haven CT (alexandra.hartman@yale.edu).

1 Introduction

Prediction of town- and village-level violence can be a valuable tool for both policymaking and theory-building. Police, peacekeepers, and governments aim to stop local violence before it escalates, but resources for prevention are scarce. Any information that helps to better allocate them should have high returns. Prediction can also aid the study of conflict. Good theories generalize, and prediction can be used to test, validate and refine theoretically-motivated empirical models (Beck et al., 2000; King and Zeng, 2001; Ward et al., 2013). In fields like conflict studies, where the explanatory power of existing models is often small, prediction is also an important inductive tool, as it can reveal patterns, puzzles and stylized facts that may, in turn, suggest areas for new research (Schrodt et al., 2013).

The amount of data available for predicting conflict is large and growing. This includes "Big Data" generated through automated text-scraping algorithms, as well as surveys, censuses, and administrative databases that measure crime, political violence, and associated risk factors. Combined with advances in computing power and data mining techniques, this information is a tremendous resource for understanding violence.

Yet studies of violence have focused on description and causal inference rather than prediction. The few existing forecasting models mainly predict national-level events, especially internal conflict (Hegre et al., 2013; Ulfelder, 2013; Ward et al., 2013), state failure (Goldstone et al., 2010; King and Zeng, 2001), and inter-state wars (Beck et al., 2000; Gleditsch and Ward, 2013).[1]

The most active frontier of conflict research in recent years has addressed local-level incidents, including riots, crime, and mob justice. Local violence is important

[1] For reviews and discussion see Ward et al. (2010, 2013); Schrodt et al. (2013).

to understand not only for its local impacts on development and well-being, but also because it can shape, and be shaped by, national-level conflict processes.[2] Much of this micro-level work has been descriptive and explanatory, however. Some scholars have used military reports, news archives, and social media to generate detailed data on local violence. But in the absence of micro-level data on risk factors, local violence forecasting has been relatively rare and mainly limited to univariate time series analysis.[3]

We have little sense of whether local forecasts are more or less reliable than national ones. Sub-national and cross-national forecasting models focus on different kinds and levels of violence, and the results from the cross-national models cannot be easily applied to the local violence. If feasible, local-level forecasts may prove more actionable than national-level ones. Preventing a civil war requires marshaling tremendous resources and coordinating among many disparate actors and institutions, and good predictions might not influence actors on the ground. Predictions could be more actionable in preventing ethnic clashes or crime requires, if only because there are fewer actors to coordinate.

This paper uses original data from Liberia to investigate the feasibility of predicting local-level violence. We harness existing data mining and machine learning techniques to tackle two questions. First, how well can we predict future outbreaks of local violence (such as ethnic riots) from low-frequency data on risk factors? Second, do the correlations we observe between risk factors and local violence suggest important theoretical avenues for future research to explore?

While we are interested in such substantive findings, our goal is not necessarily

[2] This argument has been made for rioting (Brass, 1997; Horowitz, 2003), post-war violence (Autesserre, 2010), and property disputes (Blattman et al., 2014).

[3] e.g. Schrodt and Gerner (1997); Yonamine (2014). There is also a large literature on crime forecasting in Western cities, but most of these studies similarly rely on the autoregressive properties of crime (Perry et al., 2013).

to generalize from a handful of tests on one dataset. As with any micro-level or case analysis, we cannot be sure how our results will fare when applied to other contexts. Rather, a broader purpose of this exercise is to illustrate how social scientists can use forecasting tools on the growing number of micro-level datasets on violence and its correlates around the world. In general, the innovations in data mining and machine learning of the past decades are poorly understood and underused in political science, especially at the sub-national level. Applying these innovations to the study of violence is a promising frontier, and the sum of the resulting substantive findings should provide new insights to motivate further causal research.

Our data consist of three cross-sectional surveys covering 242 rural towns and villages in Liberia in 2008, 2010, and 2012. We use surveys of both residents and local leaders to measure seven main forms of violence (including riots, murders, and extrajudicial killings) as well as dozens of geographic, social, and economic predictors.

We first used 2008 data to train models for predicting local violence in 2010, simulating forecasts using cross-validation. Our models were based on four common techniques from the forecasting and machine learning literature—logit, lasso, random forests, and neural networks. This process involved running over 30 million specifications. We then collected data for a true forecast of violence in 2012, and compared our predictions to actual incidents in that year.

We highlight four findings. First, the models perform significantly better than chance and are robust to changes in specification. This is especially true of one of the simplest methods, lasso, a form of constrained logit. If we are interested in maximizing correct predictions of violence (true positives) and minimizing false negatives, lasso predicts up to 88% of all violence in 2012 while holding overall accuracy at 33% or greater. This comes at a cost of roughly 4.5 false positives for every true positive. The area under the Receiver Operating Characteristic (ROC) curve is 0.65 for this lasso forecast, comparable to the performance of early cross-national models. Given that local violence prediction remains largely unexplored terrain, we view these results as a promising first step.

Second, some of the best-performing models are surprisingly parsimonious. The optimal lasso model includes just 5 of the 56 available risk factors, suggesting that it may be possible to generate reasonably reliable forecasts at relatively low cost in terms of data collection. Predictive performance, moreover, is not driven by serial correlation in violence.

Third, the models identify patterns that might not have been detected through hypothesis testing alone. For example, economic conditions such as poverty and adverse economic shocks are correlated with violence, but their predictive power is poor. This is consistent with many accounts of intergroup violence, which see little role for poverty in inciting attacks (e.g. Horowitz, 2003), but contrasts with studies that identify a causal causal relationship between economic shocks and crime (Freeman, 1999), extrajudicial killings (Miguel, 2005), and inter-group violence (Chassang and Padro-i Miquel, 2009).

Also, in the best-performing model, the single most robust predictor of violence is whether or not minority ethnic groups are included in local government (i.e. power-sharing). Counterintuitively, the prevalence of violence is *higher*, not lower, in power-sharing communities. This relationship is not evidence of causality, and we caution against generalizing, especially given our limited number of time periods and communities. Nonetheless, this result suggests that the interplay of local institutions and ethnic cleavages plays a role in violence, and is consistent with the argument that power-sharing can backfire.

Fourth, violence appears to be more a function of systematic than idiosyncratic risk factors, and seemingly disparate types of violence may have common roots. Some scholars argue that few episodes of violence are alike, in large part because their causes are too heterogeneous to capture in regression models or early warning systems (e.g. Cramer, 2007). Our results provide reason for optimism that local violence may be less idiosyncratic than these accounts suggest.

One limitation of our study is the small number of time periods and communities. While we could probably improve forecasts with "bigger" data, our dataset does have some important advantages. In addition to the high level of granularity, our dataset contains a large number of precisely-measured covariates, or risk factors. These are often lacking in sub-national datasets, and are crucial for the theory-building and model-testing aspects of our work. Our data on violence also have the advantage of accuracy. We triangulate reports of violence from multiple survey respondents to minimize the risk of reporting bias, then validate these reports through case-by-case qualitative interviews. "Big data" are often collated from news stories, administrative reports (such as crime statistics), and other sources. These are usually incomplete, noisy, and have unknown but potentially serious forms of bias. Thus one risks modeling the bias rather than the underlying patterns.

We also draw several substantive conclusions from this pilot. To start, local violence forecasting shows empirical promise, and could be an important tool for allocating resources and security forces to the places they are most needed. Prediction may also be useful for building theory. Generating stylized facts can point research in new directions or force us to think differently about violence and its correlates. For example, further research is required to confirm and understand any correlation between violence and power-sharing institutions designed to manage inter-ethnic relations and violence. Is this relationship causal? Or, perhaps, is power-sharing caused by past conflict, which in turn causes present conflict? We leave these questions for future research to explore.

Local-level forecasts are needed in more settings, ideally at higher frequency in larger and longer panels. Relevant administrative data are available for a growing number of countries, especially middle-income ones. Cheap, real-time data collection would also help for these purposes, especially in the poorest and most fragile states where administrative data are less common, and we discuss strategies that police and peacekeepers could employ to collect such data even. At the national level, datasets like the Integrated Crisis Early Warning System (ICEWS) have begun to provide fairly accurate monthly forecasts of insurgency, ethnic violence, and other forms of instability in about 29 countries (Ward et al., 2012). Nothing like this exists at the sub-national level, but developing higher-frequency models of this sort is a fruitful avenue for future conflict research.

2 Setting

Liberia is a West African nation of nearly four million people that suffered two civil wars from 1989 to 2003. Five main features of the setting are relevant to this study.

First, in many respects Liberia stabilized over the study period, 2008-12. It held a pair of free, competitive elections in 2005 and 2011 that brought to power a largely legitimate, professional, and open regime. The state has been supported by ample aid and a large United Nations (UN) peacekeeping operation.

Second, social cleavages remain deep. There are fifteen major ethnic groups (or "tribes") in addition to an "Americo-Liberian" elite that has historically dominated power. There are tensions not only between neighboring groups, but also within communities—for instance, between largely Christian/animist indigenous groups and Muslim traders who are regarded as lesser citizens or even "strangers" in spite of generations of cohabitation. These cleavages are widely thought to be one of the main sources of local violence.

Third, local violence is endemic. Among our study communities, in 2010 alone, 10% reported an act of collective violence, 15% reported a murder or rape, and 9% a lynching or trial by ordeal.

Fourth, these episodes of local violence can escalate, and a number of high-profile local incidents have threatened national peace. For example, in 2010, the murder of a girl in one of our study villages escalated into countywide ethnic riots. Curbing such escalation is a government and peacekeeping priority.

Finally, state capacity is weak, and the ability to prevent or respond to incidents of local violence is limited. Liberia's wars devastated the security and justice sectors. Courts are largely inaccessible, and unable to cope with the volume of cases. The army and police are underfunded and under-equipped, but their role and funding is growing as UN peacekeepers draw down. Liberia is in these respects an ideal test case for local violence forecasting, and there is widespread interest in developing early warning and response systems to inform day-to-day law enforcement, prevent the escalation of local conflicts, and allocate scarce government resources, especially as UNMIL withdraws.

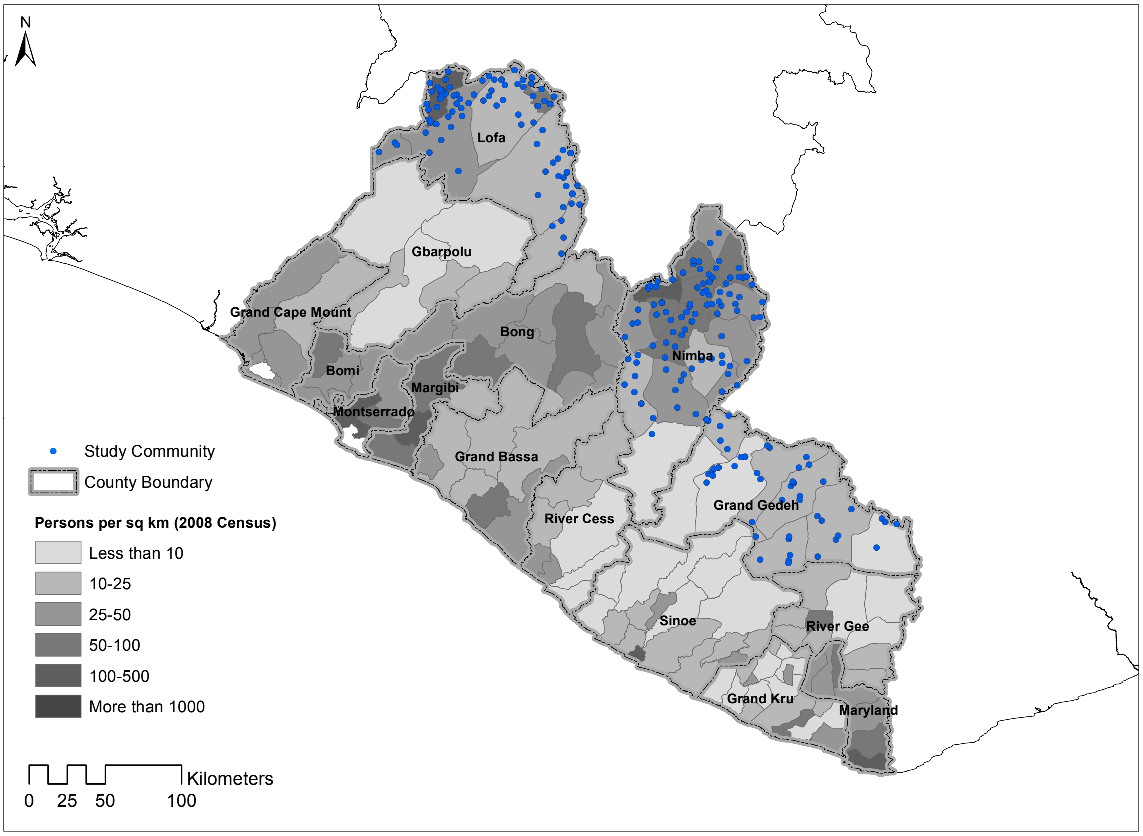
3 Data and measurement

We collected survey data from residents and local leaders in 242 villages and towns in three of Liberia's 15 counties: Lofa, Nimba, and Grand Gedeh.[4] Figure 1 displays a map of Liberia and the communities in our sample.

In each community we individually surveyed a representative sample of roughly 20 villagers plus four purposively selected leaders—typically a town chief plus women's,

[4] See Appendix A for further details, including sampling of villages and potential selection issues.

Figure 1: Communities in the study sample, with population density by district



youth, and minority leaders. We collected data in three rounds: March to April 2009, November 2010 to January 2011, and February to April 2013. In the first and second rounds we selected a new cross section of villagers each time. In the third round we interviewed leaders only for the dependent variable.

We use these surveys to construct a panel dataset, aggregating residents' responses into community-level means. We also use the leaders' responses to code communal events or characteristics, taking the modal leader response for binary variables (e.g. any collective violence) and the mean response for continuous variables (e.g. distance to the nearest road).

Note, however, that the data were collected in the context of a randomized evaluation of a government-sponsored alternative dispute resolution program, and the villages are not a representative sample of all villages in the three counties. Rather, county officials nominated these communities because they were thought to be more dispute-prone than others. We cannot say how this affects predictions, but it implies that we should interpret our results as conditional on a minimum pre-existing level of risk. This is similar in spirit to some cross-national forecasting models, which distinguish between countries with high and low ex ante probabilities of war before generating forecasts (Beck et al., 2000).

Dependent variable: Incidents of violence

Each round of the survey asked leaders about seven types of violence in the previous 12 months, which we classify into three categories: collective violence (violent strikes and protests, or violent confrontations between tribes); interpersonal violence (murders, rapes, or fights/assaults with weapons); and extrajudicial violence (beatings or killings

Table 1: Number of towns with any incident of crime or violence, 2008–12 (n=242)

Dependent variable	2008	2010	2012
Any major incident of crime or violence	90	42	40
Any collective violence	25	9	7
Any violent strike or protest	8	5	4
Any violent confrontation between tribes	17	6	3
Any interpersonal violence	64	32	21
Any murder	9	14	5
Any rape	33	17	11
Any serious fight with weapons	43	9	6
Any extrajudicial violence	23	8	16
Any trial by ordeal	23	7	16
Any witch killing or beating	1	1	0

of suspected witches, or trials by ordeal).[5]

We construct a separate indicator for each category of violence, as well as an aggregate indicator, equal to one if any of the seven types of violence occurred in the past year.

Table 1 reports the prevalence of violence in each round. In 2008, 37.2% of communities reported at least one major incident of crime or violence. By 2010, that proportion roughly halved, to 17.4%.[6] By 2012, however, the rate of decline had slowed, and the proportion of communities experiencing at least one major incident remained high (16.5%).

[5] Both men and women can be accused of using occult powers to inflict harm on others; in some cases, mobs respond by beating or killing the accused (Miguel, 2005). Trial by ordeal is another manifestation of occult beliefs, and requires suspects to withstand a heated machete on their skin, or to survive eating the bark from the poisonous sassywood tree.

[6] At baseline the "fights with weapons" question was less specific, asking only about "serious fights". This likely accounts for a significant proportion of the decline in fights from 2008 to 2010, but the decline in other incidents is similar or greater. If we omit all fights (with and without weapons) from the aggregate indicator, prevalence rates in 2008, 2010 and 2012 are 29%, 16% and 15%, respectively. Thus the fall from 2008 to 2010 is still precipitous.

Aggregation In this paper we focus on the aggregate indicator. There are two arguments in favor of aggregation. First, because cross-validation requires splitting the data into subsets, the aggregate indicator increases empirical tractability and statistical power by making these rare events less rare. Cross-national models take similar approaches, for instance by aggregating different crises into an indicator for "political instability" (Goldstone et al., 2010).

Second, our qualitative fieldwork (discussed below and in Appendix B) suggested that many incidents of violence were ambiguous, and that leaders rarely abided by our three categories in their understanding of events. A murder may incite a protest, which may pit members of different tribes against one another in episodes of tit-for-tat violence. Thus the categorization in Table 1 is not so clear cut. Similarly, in his ethnographic studies of violence in India, Brass (1997) finds that different acts of violence are interrelated and may be difficult to sort into unambiguous conceptual categories. He also argues that the ex post interpretation of violence is just that—an interpretation—and may be far removed from the actual event. We test the validity of the aggregation below.

Independent variables

We construct variables for 56 potential risk factors. Table 2 reports summary statistics in 2008 and 2010. In some cases we have data from both residents and leaders and so have two measures of the same predictor.

Table 2: Risk Factors

Covariate	2008	2010
Town population	2,032	3,117
# of households	238	337
% male	56%	48%

Table 2: Risk Factors

Covariate	2008	2010
% under 30 yrs. old	20%	27%
% non-native/strangers (leader)	2%	3%
% non-native/strangers (residents)	13%	29%
% ex-combatants (leader)	2%	2%
% ex-combatants (residents)	9%	8%
% returned from internal displacement	58%	36%
Mean educational attainment (years)	5.17	5.54
% with no education	45%	42%
% receiving any "peace training"	28%	34%
Group participation (0–9)	3.74	3.69
Collective public goods index (0–11)	1.79	1.61
% who contribute to public facilities	86%	89%
% saying town is safe at night	52%	43%
% saying neighbors helpful	70%	50%
% rely on NGOs for public goods	53%	53%
% rely on gov't for public goods	14%	17%
% describing police/courts as corrupt	33%	44%
Perceived equity in institutions (0–3)	2.60	2.37
# of tribes in town	2.63	2.66
% in largest tribe	82%	87%
% Muslim (leader)	9%	5%
% Muslim (residents)	12%	12%
Indicator for mosque in town	45%	21%
% accepting inter-religious marriage	66%	73%
% say Muslims shouldn't be leaders	27%	55%
% believing other tribes violent	30%	65%
% believing other tribes dirty	15%	56%
Minority tribe in town leadership	59%	84%
% reporting burglary or robbery	13%	19%
% reporting assault	19%	8%
% reporting any land conflict	25%	21%
Any major destabilizing event	37%	17%
% of town landless (leader)	1%	0%
% of town landless (residents)	17%	12%
% of town farmers	18%	55%
Unemployment rate	4%	7%
Wealth index	-0.02	-0.02
S.D. of wealth index in town	0.69	0.76
Exposure to war violence (0–13)	4.28	5.19

Table 2: Risk Factors

Covariate	2008	2010
Participation in war violence (0–3)	0.31	0.45
% reporting loss of land during war	10%	9%
% displaced or refugee during war	80%	84%
Social services in town (0–14)	5.61	6.81
1 if police or magistrate in town	19%	18%
Freq. of police/NGO visits (0–2)	1.26	1.23
Town >1 hour from nearest road	14%	3%
1 if mobile coverage in town	58%	74%
1 if <2 radio stations in town	87%	4%
# natural resources in 2 hours (0–5)	1.44	1.77
Basic commodities price index (0–4)	0.60	0.69
% affected by human disease	4%	19%
% affected by livestock disease	16%	30%
% affected by crop failure	26%	29%

Qualitative data

We conducted in-depth, semi-structured qualitative interviews with all leaders who reported incidents of violence, described in more detail in Appendix B. We did this to validate our dependent variable, assess measurement error, and evaluate our approach to categorization and aggregation. By validating all events, we minimize the risk of under- or over-reporting of violence in the survey data.

4 Forecasting methods

Statisticians have developed a vast range of tools for forecasting, and have made a number of attempts to assess their relative merits.[7] One important lesson from this literature is that no single model, or even class of models, dominates all others across

[7] Probably the most comprehensive and best known of these attempts was the StatLog project of the early 1990s (King et al., 1995). For a more recent example, see Caruna and Niculescu-Mizil (2006).

all applications and metrics. We selected four models prior to collecting the 2012 data.[8] Rather than attempt an exhaustive comparison, we chose a small selection of techniques (logit, lasso and random forests) with long and proven track records. These methods minimize the need for ad hoc decisions, and are transparent to and replicable by the average applied empirical social scientist. We also chose one more complex technique (neural networks) that can accommodate a large number of collinear regressors and interactive relationships. While limited, our selection captures some of the most important variation across classes of forecasting models: variable selection and coefficient shrinkage techniques (lasso), ensemble and tree-based methods (random forests), and non-linear adaptive weighting models (neural networks). We also see promise for newer techniques such as Bayesian model averaging or support vector machines for future research, but restrict our analysis to the preanalysis plan. We describe our models briefly here, with more technical details in Appendix C.

4.1 Models

Logit Logit has the virtue of simplicity and familiarity, but has potential drawbacks. Risk factors with limited predictive power may introduce more noise than signal, reducing accuracy. Also, when violence is rare but the number of potential risk factors is high, logit is overdetermined. It is common to prune variables manually or using some mechanical criterion in order to avoid these problems, but this selection process is not systematic or transparent.

Constrained linear regression: The "lasso" method The Least Absolute Shrinkage and Selection Operator is among the most widely-used variable selection tech-

[8] The models, risk factors, and predictions were published as a conference working paper prior to 2012 data collection (Blair et al., 2012), which functions as a preanalysis plan. A small number of minor changes were made following these predictions, and effects are documented in Appendix D.1.

niques in statistics (Tibshirani, 1996). In effect, it is a constrained logit regression. Lasso penalizes model complexity, shrinking all coefficients and reducing some to zero. It thus prunes variables automatically, using a transparent, replicable selection mechanism. Lasso models often include only a handful of predictors with non-zero partial effects. This parsimony is advantageous when deciding which subset of risk factors to track over time.[9]

Random forests Random forests are collections of decision trees (Breiman, 2001). A decision tree sorts observations into subgroups (or "leaves") and makes the same prediction for observations on the same leaf. The algorithm begins by identifying the single predictor that most efficiently distinguishes positives (in our case, incidents of violence) from negatives (no violence). This first split partitions the data into two subsamples that minimize the sum of squared deviations from the mean in each sample. At each node, the model then identifies secondary and tertiary predictors to further reduce mean squared error. Each observation is passed down the tree until it reaches a terminal node, at which point a prediction is made based on the average outcome for the training data observations at that node.[10] Random forests are comprised of many trees over many random subsamples of the training data and many random subsets of predictors. The prediction of the random forest is the average prediction for each tree in the forest, increasing stability.[11]

[9] We use the `glmnet` package in R (Friedman et al., 2010). There are also Stata packages, such as `lars` (Mander, 2014).

[10] In our case, each tree is fit to 24 randomly selected observations (roughly 10% of the sample) using 7 randomly selected risk factors and a maximum of five terminal nodes. Each random forest is comprised of 1,000 trees constructed in this manner. For a given observation, the algorithm generates a "final" prediction by taking the average of these 1,000 predictions.

[11] We use the `randomForest` package in R (Liaw and Wiener, 2002). We are not aware of a Stata package.

Neural networks There are many non-linear interactive techniques to choose from. We opt to use neural networks, which are especially well-established in the machine learning literature, and have been applied in political science as well (e.g. Beck et al., 2000). Neural networks make no parametric assumptions, instead using an iterative algorithm to approximate the underlying structure of the data. The algorithm begins by constructing several different weighted sums of the available risk factors (each called a node). It takes that layer of weighted sums and weights them into another weighted sum, which maps onto the prediction space. In principle there can be many layers and nodes, but ours is more parsimonious, with one layer and 5 nodes. The weights are initially chosen at random, and then tuned iteratively to minimize mean-squared error.[12]

4.2 Model training and simulated forecasting

Before collecting 2012 data on violence, we trained our models and simulated a forecast through a conventional 5-fold cross-validation. First, we randomly partitioned our 2008–10 sample of communities into five equal sized subsets. For each subset, we trained our model on four of the subsets and then generated predictions for the fifth, thus generating a prediction for each observation based on a model that was not fit to that observation. Cross-validation has been shown to approximate out-of-sample accuracy rates and is arguably the most accurate and widely-used method for estimating prediction error without new data (Hastie et al., 2009).

Because a single cross-validation can yield idiosyncratic results, we repeated the steps above across 200 trials for each model. Within each trial, we identified the model parameters that allowed us to maximize sensitivity while maintaining an overall accuracy rate above 50%. As discussed below, given the high costs of violence, we

[12] We use the `nnet` package in R (Venables and Ripley, 2002).

argue that this constitutes a reasonable balance between false positives and false negatives (though we calculate the area under the ROC curve for all models as well). We then calculated the average of those parameters, applied them across another 200 trials, and calculated average accuracy, "sensitivity" (true positive), and "specificity" (true negative) rates.

In training and testing our models, we standardize all independent variables to have zero mean and unit standard deviation.

4.3 Forecasts

For each of our models, we applied the optimal parameters from the 2008–10 cross-validations to predictors measured in 2010. We then generated predicted probabilities of violence for each community in our sample prior to collecting new data in 2012.

This test sets a high bar for our models to pass. Liberia is a country in flux.Between 2010 and 2012, for instance, Liberians living in Cote d'Ivoire lost their refugee status, precipitating an influx of once-displaced persons into the country; dozens of riot police officers were sent from Monrovia to semi-permanent deployments in the periphery; UNMIL withdrew several troop contingents; a presidential election was held. Our goal was to determine whether our models would retain their accuracy despite these changes. Because the performance of any given model can be somewhat idiosyncratic over time, both the simulated and true forecast results are of interest.

4.4 Evaluating model performance

In any prediction exercise there is a trade-off between sensitivity (true positive rate) and specificity (true negative rate). For continuous predictions, this trade-off is regulated by the threshold above which we predict an outcome will occur. The higher the threshold, the lower the true positive rate. A ROC curve plots this trade-off over the entire range of possible thresholds. The area under the curve (AUC) is a common performance metric, capturing the model's overall predictive power relative to chance.

While AUC is a useful benchmark, it can be misleading because it weights false positives and false negatives equally. This weighting may be sensible in some applications, but less so in others. For example, local violence may be sufficiently costly that a policymaker would be willing to tolerate many more false positives for the sake of fewer false negatives. To address this possibility we also provide performance metrics at the point on the ROC curve that maximizes true positives while maintaining an overall accuracy rate of at least 50%. This ensures that we correctly classify as many incidents of violence as possible while sustaining an overall accuracy rate that is at least as good as chance.[13]

5 Results

5.1 Simulated forecasts in 2010

We use cross-validation to simulate forecasts of 2010 aggregate violence using 56 predictors measured in 2008. Figure 2 plots the average ROC curve for each simulated model.[14] The solid circles indicate the thresholds that maximize sensitivity while maintaining accuracy at or above 50%.

Lasso outperforms the other models at almost all thresholds.[15] It achieves the largest area under the curve, 0.58, compared to 0.53 or less for the other models (see

[13] By comparison way of comparison, Goldstone et al. (2010) opt for the threshold that equalizes sensitivity and specificity, weighting false positives and false negatives equally.

[14] We estimated the ROC curve 200 times for each model. The lines displayed represent the average.

[15] The only exception is random forests at low discrimination thresholds, which may be of less interest because of the high number of false negatives this would imply.

Figure 2: ROC curves for simulated forecast in 2010

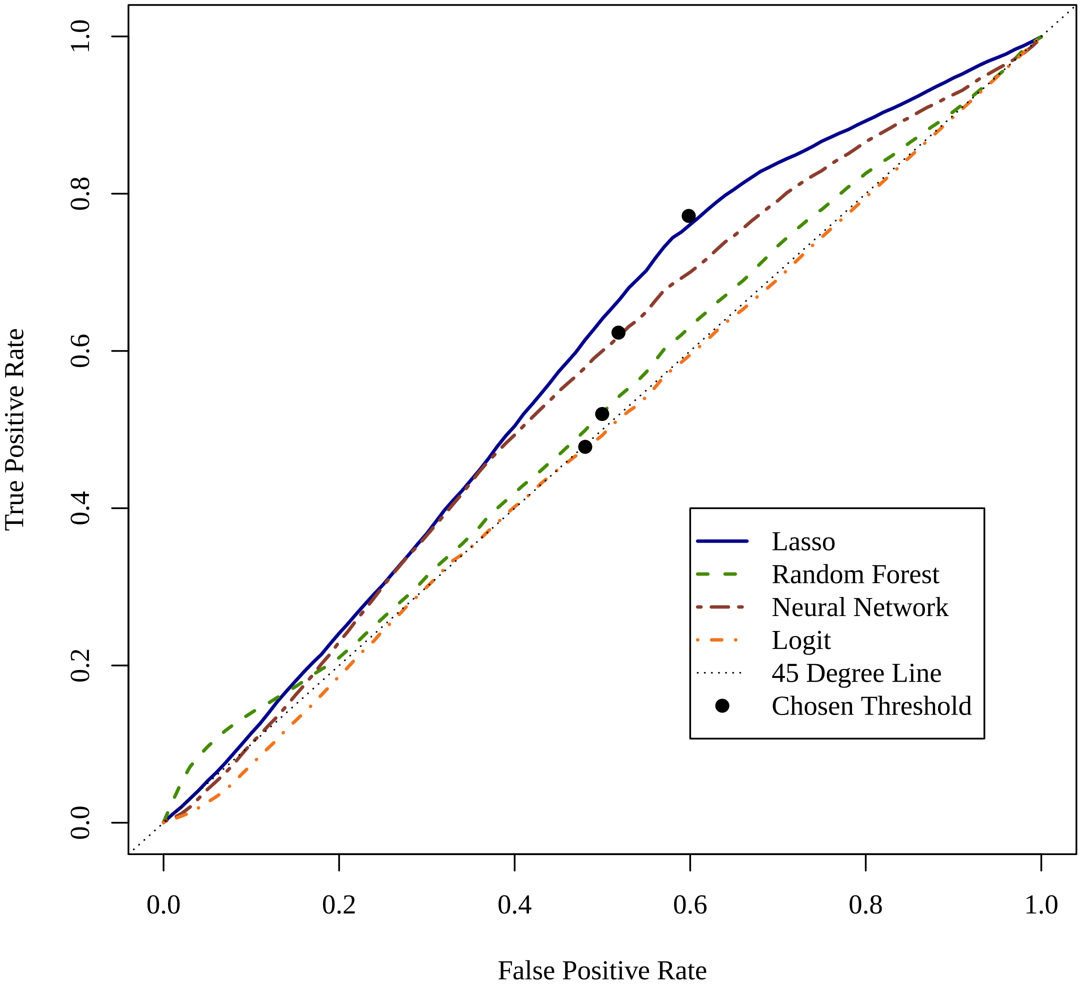


Table 3). Logit performs no better and perhaps even slightly worse than chance, perhaps because of the limitations noted in Section 4.1 above.

Moving left on the ROC from each solid circle reduces the number of villages in which we predict violence, lowering sensitivity but increasing specificity and (in some cases) accuracy. This would be appropriate, from a policy perspective, if it were difficult or costly to expend resources on prevention. While our preferred threshold was in some respects arbitrary, it was also judicious, as the solid circles correspond to the points of maximum distance from the diagonal for both lasso and neural networks. This implies that further improvements in sensitivity beyond our preferred threshold would have required increasingly large penalties in terms of specificity and accuracy.

Table 3 reports eight measures of performance for each model: AUC plus sensitivity, specificity, accuracy and other metrics at the thresholds indicated by the solid circles in Figure 2. As a benchmark, we also report the results of simply predicting that 2008 violence will recur in 2010.

Lasso again outperforms the alternatives, with a true positive rate of 77% and an accuracy rate of 47%.[16] For every true positive, the lasso generates almost four false positives but few false negatives. This level of sensitivity is attained without sacrificing much in terms of accuracy compared to the other methods (though lasso's specificity rate is considerably lower—a result of over-predicting violence). The standard deviations on these predictions are small, suggesting that the results are relatively stable over cross-validated trials. The results are also robust to most modeling choices.[17]

The sensitivity of these models is not driven by serial correlation in violence. As

[16] Since we identified the optimal parameters within each of 200 cross-validated trials, then applied the average of those parameters across another 200 trials, some models fail to achieve accuracy rates at or above 50%.

[17] Appendix D.2 reports detailed robustness checks.

Table 3: Simulated forecasts of 2010 aggregate violence using 2008 risk factors

	Dependent variable: Aggregate violence				
Performance metric	Logit	Lasso	Random Forests	Neural Networks	Simple recurrence
AUC	0.49 (0.04)	0.58 (0.03)	0.52 (0.02)	0.53 (0.04)	
True positives (sensitivity)	48% (0.08)	77% (0.05)	52% (0.05)	62% (0.06)	48%
True negatives (specificity)	52% (0.04)	40% (0.02)	50% (0.02)	48% (0.03)	65%
Overall accuracy	51% (0.03)	47% (0.02)	50% (0.02)	51% (0.03)	62%
Ratio of false + to true +	4.88 (0.73)	3.71 (0.26)	4.62 (0.46)	3.99 (0.44)	3.5
Ratio of false - to true +	1.15 (0.36)	0.30 (0.09)	0.94 (0.19)	0.62 (0.16)	1.1
Violence predicted (% villages)	48% (0.04)	63% (0.02)	50% (0.02)	54% (0.03)	37%
False negatives (% villages)	9% (0.01)	4% (0.01)	8% (0.01)	7% (0.01)	9%

we see from Column 5 of Table 3, 2008 violence predicts less than half of 2010 violence, with a true positive rate of just 38%.[18] Further, as we will see, the lagged dependent variable does not appear among the top 10 strongest predictors of violence for any model.

Lasso's relatively strong performance does not imply that it is superior to logit, random forests, or neural networks in any universal sense. Further optimization may have improved the performance of these other models. For example, dropping a subset of poorly-performing risk factors before running the simulated forecasts increases the sensitivity of the random forests and neural networks models to levels similar to

[18] Naturally it is possible that further lags of violence would improve overall model performance, and that lagged violence could displace other covariates in predictive performance. Our data cannot say.

Table 4: Performance of 2008–10 models in predicting 2012 aggregate violence

	Dependent variable: Aggregate violence				
Performance metric	Logit	Lasso	Random Forests	Neural Networks	Simple Recurrence
AUC	0.67	0.65	0.60	0.60	
True positives (sensitivity)	93%	88%	65%	63%	38%
True negatives (specificity)	35%	22%	41%	52%	87%
Overall accuracy	45%	33%	45%	54%	79%
Ratio of false + to true +	3.54	4.49	4.62	3.84	1.8
Ratio of false - to true +	0.08	0.14	0.54	0.60	1.7
Violence predicted (% villages)	69%	79%	60%	50%	17%
False negatives (% villages)	1%	2%	6%	6%	10%

lasso.[19] Given the relatively small number of villages and time periods and the large number of potential risk factors, lasso's built-in process for variable selection may have given it an edge over the alternatives.

5.2 Forecasts

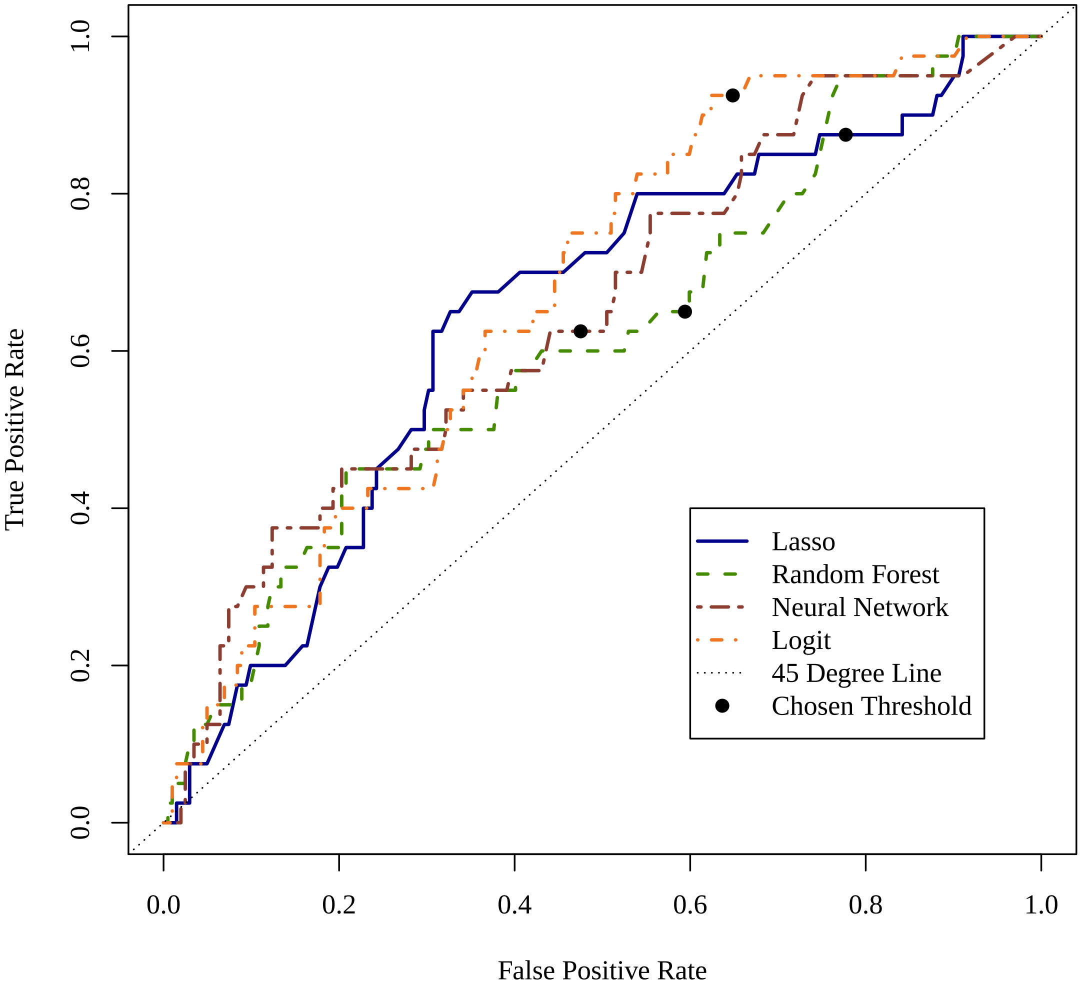
Figure 3 plots the ROC curves for our true forecasts of 2012 violence using 2010 risk factors and optimal model weights identified through cross-validation. Table 4 compares actual aggregate violence in 2012 to our predictions when we maximize sensitivity while holding accuracy in the training data at or above 50% (the points on each curve in Figure 3).

All four forecasts yield higher AUCs and sensitivity rates than in the simulations.[20] Unlike in the simulations, lasso no longer dominates the other models. The lasso AUC still exceeds that of random forests and neural networks, but the random forests and neural networks ROCs cross the lasso ROC at the lower and upper ends of the

[19] See Appendix D.2.

[20] An exception is simple violence recurrence: 2010 violence is a poorer predictor of 2012 violence than in 2008–10.

Figure 3: ROC curves for 2012 forecasts



distribution. Most surprisingly, the 2012 logit AUC exceeds that of the lasso, in spite of the fact that the 2010 logit model performed no better than chance. (Again, our results do not indicate that any one model is universally superior to the others, as relative performance can vary over time). In general, each model's AUC is robust to modeling choices, but the optimal threshold (and the tradeoff between specificity and sensitivity) varies a good deal, especially for random forests and neural networks.[21]

5.3 Is this predictive performance "high"?

Our forecasts achieve an AUC of 0.65 using lasso, 0.67 using logit and 0.60 using random forests or neural networks. How good is this performance? 0.65 is roughly the predictive power of the variables used in the classic Fearon and Laitin (2003) cross-national analysis of insurgency, where a model trained on pre-2000 data is used to predict post-2000 outbreaks of civil war (Ward et al., 2013). More recent cross-national models have reached AUCs exceeding 0.8 (Hegre et al., 2013; Ulfelder, 2013).

Our pilot performs about as well as some of the earlier cross-national models. This is unsurprising. The cross-national literature has generated many more datasets covering many more time periods. Moreover, lessons learned over more than a decade of research have iteratively improved performance. Similar improvements in the prediction of local violence may be possible with further replication. Some of the cross-national improvements have come from new risk factors, but some of the biggest gains are from accounting for correlations over time and space. This suggests an important direction for micro-forecasting research.

Finally, many cross-national models achieve high AUCs by under-predicting violence, and thus generating a large number of potentially costly false negatives. In contrast, our local-level lasso performs best by over-predicting violence. This is partly

[21] See Appendix D.2.

intentional, as we selected thresholds and model parameters to guard against false negatives. This comes at the cost of many false positives (4.5 for every true positive).[22][23] This may be a price that policymakers are willing to pay, especially if the costs of prevention are relatively low and the costs of violence relatively high.[24]

5.4 Model averaging

Rather than arbitrate between models, one alternative would be to average across them. A simple approach, "majority voting," generates a single prediction according to what the majority of the models predict. Majority voting achieves a sensitivity of 72% in the simulations and 90% in the true forecasts. Another approach, where we average predicted probability across models and classify towns according to a discrimination threshold chosen by cross validation, obtains sensitivity of 59% in the simulations and 83% in the true forecasts, with accuracy rates of 50% and 46% respectively.[25] Overall performance is similar to our best performing models. Appendix D.5 discusses detailed performance and alternatives.

5.5 What variables predict violence in these models?

Table 5 ranks our 56 risk factors by the magnitude (in absolute value) of their coefficients in the lasso model and by their importance scores in the random forests model (where "importance" is calculated as the average decrease in mean squared er-

[22] For a visualization of lasso accuracy, see Appendix D.3.

[23] Random forests and neural networks, meanwhile, correctly classify fewer cases of violence, and thus have higher specificity and overall accuracy. They do so, however, with two to three times the proportion of false negatives to true positives as lasso.

[24] Note that the surest way to increase accuracy would be to predict peace in every community. This approach would achieve sensitivity of 0%, specificity of 100%, and overall accuracy of 83% for both 2010 and 2012.

[25] Some of the more complex "ensemble" methods, such as Bayesian model averaging, are difficult or impossible to implement on our data (e.g. Montgomery et al., 2012).

Table 5: Rankings of risk factors by model

	Lasso		Random Forests	
Risk factor	Rank	Coeff.	Rank	Importance
Minority Tribe in Town Leadership	1	0.30	9	0.0003
Town Population	2	0.15	1	0.0021
Percent Believing other Tribes are Violent	3	0.07	11	0.0003
Percent in Dominant Group	4	-0.05	3	0.0006
Percent Who Contribute to Public Facilities	5	0.01	24	0.0001
Mean Educational Attainment			2	0.0008
Percent reporting Loss of Land During War			4	0.0006
S.D. of Wealth Index In Town			5	0.0006
Number of Households			6	0.0005
Number of Tribes			7	0.0004

ror achieved by the addition of each variable to the model).[26] Neural network weights cannot be meaningfully ranked in this way and are omitted.

The optimal lasso model is parsimonious, achieving high sensitivity and consistency across time periods and model specifications using just five variables. In some sense, random forests is also parsimonious, in that the importance scores decline after the first five to ten (and even the first) variables in the ranking.[27] The most important predictors do, however, vary somewhat across models. Just three of the five lasso risk factors are in the top 10 for random forests. Even within models, changes in specification lead to (modest) changes in risk factors or their ordering (Appendix D.2). One possible interpretation for this result is that different risk factors capture similar underlying characteristics of the communities in our sample. For example, "minority tribe in town relationship," "percent believing other tribes are violent," "percent

[26] A comparison to logit, and a list of all 56 risk factors and their corresponding coefficients, is in Appendix D.6.

[27] Logit is not as parsimonious and, as a result of multiple correlation, assigns opposite signs to related risk factors. This makes the relative importance of related predictors difficult to assess—a limitation of logit relative to lasso.

in dominant group" and "number of tribes" are all related to ethnic heterogeneity and polarization.[28] Nonetheless, model dependence suggests that we should exercise caution in interpreting the importance of specific risk factors.

With this caveat in mind, consider the risk factors themselves. Some are intuitive, even mechanical. For example, given that our dependent variable is binary, the correlation between violence and town population is unsurprising: more people implies more potential disputants and more potential violence.

Other risk factors are consistent with existing theories. For example, the larger the degree of ethnic heterogeneity, measured as the proportion of residents in the majority tribe, the higher the risk of violence in these communities.

Other results are more surprising. One is the absence of economic risk factors in the lasso model. Wealth levels are not significant predictors, nor are wealth shocks such as droughts, floods and pest infestations. Economic variables appear in the random forests model, but these capture inequality and loss of land during the war—sources of economic grievance rather than income shocks.

Perhaps most striking is the relationship in the lasso model between violence and power-sharing (an indicator for whether the minority tribe is represented in village leadership). The coefficient is large and positive, meaning that power-sharing predicts an *increase* rather than a *decrease* in the probability that violence will occur. Power-sharing appears among the top ten random forests risk factors as well, though is lower ranked. We return to this result below.

[28] Another possible explanation is that lasso tends to favor uncorrelated regressors, while random forests tends to favor correlated ones.

5.6 Disaggregated violence

While our decision to aggregate different categories of violence into a single indicator was partly necessary for statistical power, it obviously implied some conceptual slippage. As we show through qualitative data in Appendix B, different categories of violence often proved difficult to distinguish from one another. As it turns out, the predictors of different types of violence are also reasonably similar.

Cross-category forecast performance If our three categories of violence are indeed interrelated, then they should share some of the same correlates, and it should be possible to predict one category reasonably accurately using the same risk factors that predict the others. Table 6 tests this possibility. The table displays AUC only, with other performance metrics in Appendix D.8.

With a few exceptions, the models continue to perform better than chance. Most notably, models trained on interpersonal and extrajudicial violence predict collective violence with AUCs ranging from 0.52 (neural networks) to 0.6 (lasso) in the simulations, and from 0.58 (neural networks) to 0.7 (lasso) in the 2012 forecast. These results are comparable to the full model performance in Table 4. The lasso generally (though not universally) continues to outperform the alternatives, producing AUCs above 0.5 in six of the nine models, and an AUC below 0.5 in only one. Most important, the models generally perform about as well predicting the excluded categories (the second and third rows in Table 4) as they do predicting the included ones (the first row).

Within-category forecast performance Finally, Table 7 reproduces our simulated and true forecasts with the dependent variable disaggregated into its three

Table 6: How well do different categories of violence predict one another?

	Area under the curve (AUC)								
Categories used for prediction (No. of positives in 2008):	Interpersonal + extrajudicial (38)			Interpersonal + collective (37)			Collective + extrajudicial (16)		
Omitted category (No. of positives in 2010, 2012):	Collective violence (9, 7)			Extrajudicial (8, 16)			Interpersonal (32, 21)		
Model:	Lasso	RF	NN	Lasso	RF	NN	Lasso	RF	NN
Training the 2008-10 model on pairs of categories (Simulated forecast)	0.57	0.54	0.58	0.57	0.53	0.58	0.50	0.50	0.46
Success at predicting the excluded category of violence in 2010 (Current forecast)	0.60	0.54	0.52	0.44	0.48	0.44	0.50	0.54	0.50
Success at predicting the excluded category of violence in 2012 (True forecast)	0.70	0.62	0.58	0.58	0.50	0.52	0.58	0.46	0.47

Table 7: Simulated and true forecasts for disaggregated violence

	Area under the curve								
	Collective violence			Extrajudicial violence			Interpersonal violence		
Performance metric	Lasso	RF	NN	Lasso	RF	NN	Lasso	RF	NN
Simulated forecast, 2010	0.48	0.65	0.58	0.42	0.47	0.38	0.65	0.57	0.67
True forecast, 2012	0.48	0.52	0.66	0.71	0.52	0.61	0.61	0.60	0.71

component parts.[29]

Performance varies but overall is poorer with the disaggregated indicators, largely because we are predicting rarer events. In the 2012 forecasts, lasso predicts extrajudicial and interpersonal violence as well as it predicts aggregate violence, but its performance no longer dominates. Lasso also performs poorly in predicting extrajudicial violence in the simulations, and in predicting collective violence in both the true and simulated forecasts. This is likely because instances of extrajudicial and collective violence are rarer in the dataset. In general, the rarer the event, the worse the models perform. This reaffirms our statistical power justification for aggregation.

6 Discussion and conclusions

Prediction has a long and successful track record in election forecasting and cross-national violence. Our results suggest that the prospects for local violence prediction and data-driven early warning are promising as well. We illustrate tools and techniques that the social scientist familiar with regression should be able to understand and replicate, in many cases on existing data. In this Liberia pilot, a relatively simple, parsimonious, and transparent variable selection model (the lasso) outperforms most alternatives, and a simple majority vote aggregation of four models performs

[29] Full performance metrics are in Appendix D.8.

similarly well.

We draw several initial conclusions, to be tested in other contexts, ideally with longer term event data. First, local violence seems to be driven at least in part by systematic, quantifiable risk factors. Second, the number of risk factors needed to predict local violence may be relatively few and possible to measure—e.g., population and ethnic heterogeneity. This suggests that future early warning exercises may achieve similar results at lower cost.

Third, model performance can likely improve through larger, longer-term datasets on which to train and test prediction models. Spatial models may have promise as well (though it is not clear the types of violence we study are prone to geographic spillovers). In the meantime, however, even simple models can serve as useful probabilistic tools, forecasting violence with relatively few risk factors. Governments, peacekeepers and civil society organizations may be able to leverage these models to allocate scarce resource and address the conditions conducive to violence before it occurs.

There are at least two practical ways to move ahead. One is to harness large, subnational administrative datasets on violence and potential risk factors—data which are more commonly available in middle-income countries but not the least developed countries. Municipal or district-level crime and conflict data, along with a large number of covariates, are available for a host of mainly middle-income countries with conflicts, including at least two of the five largest countries in the world India and Indonesia.[30] The growing amount of web-scraped news data on violent events, meanwhile, could be geolocated and merged with administrative data on covariates to produce similar datasets for other large countries.

[30] An example is crime and civil war violence in India, which has mainly been used for testing causal hypotheses (Iyer and Topalova, 2014; Fetzer, 2013).

Where such data do not exist, there may be technological solutions (such as mobile phone-based surveys), news- and web-scrapeable data, or even traditional surveys that cheaply increase sample size, time periods, and frequency of measurement. For instance, Berger (2014) uses mobile phone data in Côte d'Ivoire to improve predictions of UN data on local violence. Also, Hirose et al. (2014) use survey data on civilian attitudes to improve predictions of Taliban attacks. Even single cross-sections that gather time-invariant data could improve predictive power. Cost is an impediment but not insurmountable. Peacekeeping missions like the one in Liberia expend huge resources on patrols, field-based observers, and other forms of data collection. In principle, the leader-based survey we conducted could be performed by mobile phone. To narrow the number of key risk factors to measure, peacekeepers could use an intensive survey-based approach in a small random sample of communities. The effort and expense would be small relative to the overall budgets of peacekeeping missions.

Beyond these practical implications for future research, the exercise generated substantive patterns worth investigation. The power-sharing result is one example. Several studies have found that exclusionary institutions foment conflict between dominant and marginalized groups (Cederman et al., 2010), or that power-sharing helps mitigate the risk of violence (e.g. Hartzell and Hoddie, 2003). Our results are inconsistent with these studies, and instead align with a growing literature suggesting that power-sharing may often be unstable, and may itself be a consequence of long-standing inter-group struggles (e.g. Lake and Rothchild, 1996; Gates et al., 2014). One possible interpretation is that power-sharing arrangements are responses to past conflicts, and that past conflicts continue to predict future ones. Similarly, villages where customary institutions (which tend to exclude minority groups) dominate may represent particularly strong cases of majority ethnic rule that manage to suppress violence. There is a potential parallel to the cross-national forecasting literature, which finds that ethnically factionalized semi-democracies are among the least stable regime types (Goldstone et al., 2010). While we are careful not to generalize too much from this one case, and one out-of-sample forecast, these parallels deserve further exploration.

In the end, we think this argues for a more even balance between forecasting and hypothesis-testing. Currently the balance of conflict research—even of comparative politics and international relations research in general—is almost entirely on the hypothesis-testing side, with little work on the prediction side. At this extreme, the marginal gains from more prediction exercises almost surely exceed the gains from more causal investigations. Large quantities of existing micro-level data could be harnessed for purposes of prediction, and new datasets are built every day. This unexplored frontier is one of the discipline's most intriguing and promising.

References

Autesserre, S. (2010). *The trouble with the congo: Local violence and the failure of international peacebuilding.* Cambridge: Cambridge University Press.

Beck, N., G. King, and L. Zeng (2000). Improving quantitative studies of international conflict: A conjecture. *The American Political Science Review 94*, 21–35. 1.

Berger, D. (2014). Violence and cell phone communication patterns: Evidence from cote d'ivoire. *Working paper*.

Blair, R., C. Blattman, and A. Hartman (2012, November). Predicting local level violence. Columbia University.

Blattman, C., A. Hartman, and R. Blair (2014). How to promote order and property rights under weak rule of law? an experiment in changing dispute resolution behavior through community education. *American Political Science Review 108*(1).

Brass, P. R. (1997). *Theft of an idol: Text and context in the representation of collective violence.* Princeton: Princeton Univ Press.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Caruna, R. and A. Niculescu-Mizil (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*.

Cederman, L.-E., A. Wimmer, and B. Min (2010). Why do ethnic groups rebel? new data and analysis. *World Politics 62*(01), 87–119.

Chassang, S. and G. Padro-i Miquel (2009). Economic shocks and civil war. *Quarterly Journal of Political Science 4*, 211–228. 3.

Cramer, C. (2007). *Violence in developing countries: war, memory, progress*. Bloomington: Indiana University Press.

Fearon, J. D. and D. D. Laitin (2003). Ethnicity, insurgency and civil war. *American Political Science Review 97*, 75–90. 1.

Fetzer, T. (2013). Can workfare programs moderate violence? evidence from india. *Working paper*.

Freeman, R. B. (1999). The economics of crime. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 5, pp. 3529–3572. Elsevier.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*(1), 1.

Gates, S., K. W. Strøm, B. A. Graham, Y. Lupu, and H. Strand (2014). Powersharing, protection, and peace. *Working paper*.

Gleditsch, K. S. and M. D. Ward (2013). Forecasting is difficult, especially about the future using contentious issues to forecast interstate disputes. *Journal of Peace Research 50*(1), 17–31.

Goldstone, J. A., R. H. Bates, D. L. Epstein, T. R. Gurr, M. Lustik, M. G. Marshall, J. Ulfelder, and M. Woodward (2010). A global forecasting model of political instability. *American Journal of Political Science 54*(1), 190–208. 1.

Hartzell, C. and M. Hoddie (2003). Institutionalizing peace: Power sharing and post-civil war conflict management. *American Journal of Political Science 47*(2), 318–332. 2.

Hastie, T., R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani (2009). *The elements of statistical learning*, Volume 2. Springer.

Hegre, H., J. Karlsen, H. M. Nygård, H. Strand, and H. Urdal (2013). Predicting armed conflict, 2010–2050. *International Studies Quarterly 57*(2), 250–270.

Hirose, K., K. Imai, and J. Lyall (2014). Can civilian attitudes predict civil war violence? *Working paper*.

Horowitz, D. L. (2003). *The deadly ethnic riot*. University of California Press.

Iyer, L. and P. Topalova (2014). Poverty and crime: Evidence from rainfall and trade shocks in india. *Working Paper*.

King, G. and L. Zeng (2001). Improving forecasts of state failure. *World Politics 53*(4), 623–658.

King, R. D., C. Feng, and A. Sutherland (1995). Statlog: Comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence 9*(3), 289–333.

Lake, D. A. and D. Rothchild (1996). Containing fear: The origins and management of ethnic conflict. *International security 21*(2), 41–75.

Liaw, A. and M. Wiener (2002). Classification and regression by randomForest. *R news 2*(3), 18–22.

Mander, A. (2014). LARS: Stata module to perform least angle regression. *Statistical Software Components*.

Miguel, E. (2005). Poverty and witch killing. *Review of Economic Studies 72*, 1153–1172. 4.

Montgomery, J. M., F. M. Hollenbach, and M. D. Ward (2012). Improving predictions using ensemble bayesian model averaging. *Political Analysis 20*(3), 271–291.

Perry, W. L., B. McInnis, C. C. Price, S. C. Smith, and J. S. Hollywood (2013). *Predictive policing: The role of crime forecasting in law enforcement operations.* Rand Corporation.

Schrodt, P. A. and D. J. Gerner (1997). Empirical indicators of crisis phase in the middle east, 1979-1995. *Journal of Conflict Resolution 41*(4), 529–552.

Schrodt, P. A., J. Yonamine, and B. E. Bagozzi (2013). Data-based computational approaches to forecasting political violence. In *Handbook of Computational Approaches to Counterterrorism*, pp. 129–162. Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Ulfelder, J. (2013). A multimodel ensemble for forecasting onsets of state-sponsored mass killing. *Working paper*.

Venables, W. N. and B. D. Ripley (2002). *Modern applied statistics with S.* Springer.

Ward, M. D., B. D. Greenhill, and K. M. Bakke (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research 47*(4), 363–375.

Ward, M. D., N. W. Metternich, C. Carrington, C. Dorff, M. Gallop, F. Hollenbach, A. Schultz, and S. Weschle (2012). Geographical models of crises: Evidence from ICEWS. In *Walker, PB, Angelo, A., & Davidson, IN (2012).“Network Discovery: Measuring Cause and Effect Behind Event and Social Networks,” Paper presented at 2nd International Conference on Cross-Cultural Decision Making: Focus*, pp. 21–25.

Ward, M. D., N. W. Metternich, C. L. Dorff, M. Gallop, F. M. Hollenbach, A. Schultz, and S. Weschle (2013). Learning from the past and stepping into the future: Toward a new generation of conflict prediction. *International Studies Review 15*(4), 473–490.

Yonamine, J. (2014). Predicting future levels of violence in afghanistan districts using GDELT. *Working paper*.

Appendix for online publication

A Survey sites

There were originally 246 communities in the sample. We have survey data on 242 of the 246 originally selected communities, as surveyors could not reach two extremely remote villages for any survey round, one tiny village disbanded before the first follow-up, and one village is missing data.

The data were collected in the context of a randomized evaluation of a government-sponsored alternative dispute resolution training intervention, and the villages are not a representative sample of villages in these counties. Rather, county officials nominated these communities because they were thought to be more dispute-prone than other villages. It is difficult to say how this sampling approach affects our predictions and predictive power. It could improve or reduce performance. From a policy perspective, one can imagine that understanding predictive power in the subset of places perceived to be risky is a useful policy variable, and reasonable under budget constraints. Nonetheless, one would like to be able to test these assumptions using a representative sample of villages, or stratified random sample. Unfortunately this is not available in this case, and is recommended for future research where possible.

No census frame existed, so for the representative sample, a team walked each community and divided it into blocks, chose a random pathway, counted all houses in that path, and randomly chose a set number. Household members were selected randomly. Non-response was typically less than 5 to 10% per community.

We validate our events qualitatively, as described in the following section. It is not possible to validate or compare these events to conflict events measured by ACLED or other sources. First, we measure events that are inherently smaller scale. Second, the majority of major conflict events occurred during not after the conflict, so the datasets do not overlap much in time. Third, news-based databases such as ACLED are inherently incomplete and biased due to the bias in news coverage, especially in countries where the majority of news coverage is radio-based and news organizations are relatively unprofessional and disorganized, with no presence outside the capital.

B Qualitative investigation of event classification

We collected three main forms of qualitative data:

1. Between the first and second phases of data collection (2008–10), we and three Liberian research assistants conducted 104 formal interviews with respondents in 20 purposefully selected communities. We selected communities with high and low levels of conflict, as well as those showing variation along potentially important correlates of conflict (exposure to wartime violence, remoteness and size).

2. Following the second wave of data collection (2010–11), we sent our Liberian research assistants to investigate and verify all incidents of collective violence reported in the survey through interviews and written notes. While we did not have the resources to back-check other types of violence, these interviews helped us validate the survey data and explore the interconnections between apparently disparate violent events.

3. Finally, during the third wave of data collection, enumerators digital interview in depth any leader who reported any incident of violence (from the seven categories described above). This exercise served two main purposes. First, it helped us to further validate the survey data, building our confidence that the

dependent variable is measured with as little reporting bias as possible. Second, along with the earlier interviews, it informed our decision to aggregate different forms of violence into a single category.

In general, respondents who reported incidents in the survey continued to do so during qualitative follow-up. Our interviews also suggested, however, that grouping incidents into non-overlapping categories would be challenging. Respondents described the same incidents in strikingly different ways. Much of this ambiguity resulted from the dynamics of conflict escalation. The police and courts in Liberia are notoriously inept, inaccessible and corrupt. Because victims cannot rely on these institutions to resolve disputes, violence easily mutates from one form to another (e.g., a murder turns into a riot or mob justice). Riots, lynchings and trials by ordeal often serve as extrajudicial mechanisms for adjudicating other types of crime (e.g., rape and murder); it's a hardly a surprise that they often go hand in hand. There are numerous specific examples from our fieldwork, and we outline a handful here for illustration.

- In our largest study town—Voinjama, Lofa County—the mysterious disappearance and killing of a girl provoked a peaceful protest which quickly turned violent. The (Lorma and Christian) mother of the missing girl accused the town's Mandingo (Muslim) population of abducting and murdering the girl in a ritual killing. Traditional leaders ("Zoes") were called to attempt to divine the identity of the perpetrator, fomenting allegations of witchcraft. Riots ensued, killing four. In our interviews, respondents varied dramatically in how they categorized the sequence of events—as a riot, a murder, a violent confrontation between tribes, a fight between men, or a lynching of suspected witches.

- In one of the larger towns, a hit-and-run accident provoked a violent protest by the motorbike union. In their descriptions of the incident, some respondents focused on the hit-and-run, others on the violent protest.

- In several villages, respondents described how trials by ordeal had been used to identify suspected murderers, typically in cases involving an unusual or mysterious death. As one local leader explained: "A little girl. . . passed away within this community and everybody was surprised of that particular death, so the parents of that little girl decided to go for sassy wood [trial by ordeal]. . . The sassy wood man came and he...used hot cutlass—they put the cutlass on the fire and...if you ain't part of it will just be like water on your skin. . . They started to do it going around all the people in the neighborhood.... They started touching them with the hot cutlass...and the cutlass was able to grab one person. . . because that particular person was the doer of the act."

- In a small village, we directly observed a seemingly intoxicated woman attack a man. Shortly thereafter another woman—a female relative or friend of the man—attacked the first woman. As the two women grappled, male family members and friends gathered and began to exchange insults. A physical fight between two of the men ensued. The crowd continued to grow, and several bystanders began agitating to join the fray. While the incident was eventually diffused, its interpersonal and collective dimensions remained difficult to disentangle.

C Forecasting models

This section describes in more detail the estimation methods for each prediction model, summarizing the generic method and highlighting the specific modeling choices we made in each case.

C.1 Lasso method

Given some dataset $(\mathbf{x_i}, y_i)$ where $\mathbf{x_i}$ denotes a set of j standardized predictor variables and y_i a vector of responses, the lasso coefficients β are given by:

$$l\left(\beta\right) = \text{-}\sum_{i=1}^{N} \ln\left(1 + \exp\left(-\beta^{\mathbf{T}}\mathbf{x_i}y_i\right)\right) + \lambda\sum_{j=1}^{k}|\beta_j|$$

where the first expression on the right hand side is a standard maximum likelihood estimator, and the second is the penalty function specific to lasso. $\lambda \geq 0$ is a tuning parameter that controls the degree of coefficient shrinkage; the coefficients on poor performers are forced to 0 and thus dropped from the model. A ridge regression looks similar, except that the penalty is $\lambda \sum_{j=1}^{k} \beta_j{}^2$. The key difference between lasso and ridge regression is that the latter assigns non-zero coefficients to all predictors, though the coefficients on poorly performing indicators can be very small. Ridge regression thus performs coefficient shrinkage only, while lasso performs both coefficient shrinkage and variable selection, and so generally produces more parsimonious models. We opt for an elastic net, which uses a penalty that is a weighted sum of the lasso and ridge penalty:

$$l\left(\beta\right) = \text{-}\sum_{i=1}^{N} \ln\left(1 + \exp\left(-\beta^{\mathbf{T}}\mathbf{x_i}y_i\right)\right) + \lambda\left(\alpha\sum_{j=1}^{k}|\beta_j| + (1-\alpha)\sum_{j=1}^{k}\beta_j^2\right)$$

We use a variation known as “elastic net optimization.” that involves a scalar α, which regulates the weight given to lasso ($\alpha = 1$) versus ridge ($\alpha = 0$) optimization. In our preferred model we set $\alpha = .95$, thus weighting the lasso penalty much more strongly than the ridge. We use a modification of lasso analogous to logit in order to accommodate our binary dependent variable. Thus, for a given observation, our model generates a predicted probability of violence between 0 and 1. We then classify each observation as 0 or 1 (violence or no violence) according to a discrimination threshold that is chosen by cross-validation to maximize sensitivity, keeping accuracy above 50%.

Our pruning procedure is as follows. First, we split the sample into five subsets, or folds. We then train a lasso model on four of the five folds. This is the initial training set. The lasso is fit over a sequence of 80 lambdas in the training data, in effect producing 80 lasso models, each with a different lambda (and thus a different vector of coefficients). The lambda that maximizes sensitivity while maintaining accuracy above 50% in the training set is then applied to the test set. We iterate this process over the five possible combinations of folds into training and test sets. This is one cross-validation. We then repeat this process 200 times and calculate the average optimal lambda across these 200 cross-validations—80,000 regressions in total. Finally, we repeat the cross-validation procedure 200 additional times, this time applying the average optimal lambda to every model. We calculate performance metrics within each of these 200 trials, then report the average of each metric.

C.2 Random forests

Given some dataset $(\mathbf{x_i}, y_i)$, a regression tree sorts observations into leaves and makes a prediction, $\hat{y}$, for each leaf. Trees are constructed stepwise. Initially, all observations are on the same leaf. The observations are then divided into two leaves based on values of one of the k predictors, so that the sum of squared deviations from the mean in each leaf is minimized. More formally, we minimize:

$$MSE = \sum_{l=1}^{2} \left(\sum_{i=1}^{N_l} (y_i - \bar{y}_j)^2 \right)$$

where $\bar{y}_j$ is the average outcome in leaf j and N_j is the number of observations in leaf j. In the next step, each of the these leaves is split again based on the predictor that most reduces the sum of squared deviations from the mean in the leaf (this could be the same predictor that was chosen in the first step). In principle, this process could continue until all leaves contain only observations of the same value (and $MSE = 0$). However, researchers typically employ some sort of stopping criteria before that happens. In our case, we set the maximum number of nodes to be 5. Because we use regression trees rather than classification trees, each observation is assigned a predicted probability rather than a binary (0/1) prediction.

Random forests are comprised of many trees fit to random subsets of the data with random subsets of predictors available for splitting. Each tree generates a distinct predicted probability for each observation, and the prediction for the entire random forest model is just the average of the predictions of each tree in the forest.

For our random forests model, we grow 1,000 trees with a maximum of 5 terminal nodes each and $\sqrt{56}$ variables sampled (without replacement) at each node.

C.3 Neural networks

Neural networks are layered systems of weighted sums of predictor variables with a final weighted sum mapped into the prediction space. In order to control model complexity, practitioners specify the number of layers and the number of weighted sums (called nodes) that comprise each layer. Our model has one layer and 5 nodes, and a weight decay of 0.1 with randomly selected near-zero starting values. Using five different sets of weights, the 56 predictors plus a constant are mapped onto each of the 5 nodes. Then, these five nodes plus another constant are mapped, by some linear combination of weights, to a scalar. Finally, this scalar is mapped by a logistic function to the interval $[0, 1]$, our prediction space. Hence, our network is defined by

287 weights $(56 \times 5 + 5 + 2)$. they are initially chosen at random, and then tuned iteratively to minimize the mean-squared error in the prediction space. The net is trained via back-propagation.

More specifically, a neural network is a two-stage regression or classification model, typically represented as a "network diagram" with K units at the top; the kth unit models the probability of class k. In our classification model $k = 1$ and the response $Y_{k=1}$ is simply a binary variable.

Neural networks capture interactivity by generating "derived features," denoted Z_m, from linear combinations of the predictors, then modeling the response as a function of linear combinations of the derived features:

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), m = 1, ..., M$$

$$T_k = \beta_{0k} + \beta_k^T Z, k = 1, ..., K$$

$$f_k(X) = g_k(T), k = 1, ..., K$$

where $Z = \{Z_1, ..., Z_m\}, T = \{T_1, ..., T_k\}, \sigma(\nu)$ is an initial non-linear transformation of the predictors, and $g_k(T)$ is a final transformation of the output vector T. For the special case where $\sigma = 1$, the network collapses to a linear model. For a more thorough explanation of neural networks and their analogies to maximum likelihood, see Beck et al. (2000).

D Supplemental tables

D.1 Original forecasting models and effects of subsequent changes

The original Lasso model results, calculated before the 2012 data collection, are presented in Column 1 of Table C.1. A small number of relatively minor technical changes were made after the 2012 data were collected, not with an eye to improving the specific performance of the models on these new data, but to correct small errors or adhere to expert recommendations. Table C.1 details each change in turn, including the cumulative effect on predictions and risk factors.

First, we estimated simulated forecast errors by applying a single set of optimal parameters across 200 cross-validated trials, rather than estimating the error using a varied set of optimal parameters identified within each trial (Column 2). Finally, we standardized dummy variables, which we had not been doing in previous models. These changes had little material effect on model performance or risk factor rankings in the simulated forecasts.

D.2 Robustness of prediction model performance

Lasso

Table C.2 reports various robustness checks for the simulated forecasts (Panel A), true forecasts (Panel B), and corresponding risk factors and rankings (Panel C), limiting to the latter to the top five factors only. The columns are as follows:

1. *Main specification:* From Tables 3 and 4.
2. *New Seed:* We specify an alternate randomization seed for the selection of folds which are used to select parameters and estimate error rates using 2010

Table C.1: Reconciliation of original to current Lasso predictions

(a) Simulated forecast (2010)

	Original Model	Change in cross validation	Standardized indicators
Performance metric	(1)	(2)	(3)
AUC		0.56	0.58
True positives (sensitivity)	69%	67%	77%
True negatives (specificity)	49%	44%	41%
Overall accuracy	52%	48%	47%
Ratio of false + to true +	3.98	4.00	3.68
Ratio of false - to true +	0.52	0.50	0.31

	Original Model	Change in cross validation	Standardized indicators
Performance metric	(1)	(2)	(3)
AUC		0.66	0.65
True positives (sensitivity)		85%	88%
True negatives (specificity)		35%	23%
Overall accuracy		43%	33%
Ratio of false + to true +		3.88	4.46
Ratio of false - to true +		0.18	0.14

(b) 2012 Forecasts

	Original Model	Change in cross validation	Standardized indicators
Performance metric	(1)	(2)	(3)
Minority Tribe in Leadership	1	1	1
Town Population	2	9	2
Proportion in Largest Tribe	3	6	4
Percent Muslim	4	10	
Percent Reporting Armed Robbery or Burglary	5		
Percent Contributing to Public Facilities	6	4	5
Number of Tribes	7		
Percent Reporting Loss of Land During War	8		
Number of Resources Available	9	8	
Community Wealth Index	10	14	
Percent Farmers		2	
Percent Believing Other Tribes Are Violent		3	3
Participation in War Violence		5	
Frequency of Police Visits		7	

(c) Risk Factor Rankings

outcomes.

3. *Dummies Not Standardized:* We keep binary predictors on a (0,1) scale rather than standardizing them to have mean 0 and standard deviation 1.

4. *10-fold Cross Validation:* We identify optimal parameters and estimate forecast error rates using 10-fold cross validation rather than 5-fold cross validation.

5. $\alpha = 1$: α is the weight placed on the Lasso penalty (sum of coefficient magnitudes) relative to the Ridge penalty (sum of squared coefficients). Our usual value for this is .95. When $\alpha = 1$, we have a pure Lasso penalty.

6. $\alpha = .5$: sets the penalty to be half-way between a Lasso and Ridge penalty.

7. *Subset (30) from OLS:* We first fit an OLS model to the 2008/2010 data to determine the 30 coefficients of greatest magnitude. We then use only those 30 predictors for the model.

Random forests

Table C.3 reports various robustness checks for the simulated forecasts (Panel A), true forecasts (Panel B), and corresponding risk factors and rankings (Panel C), limiting to the latter to the top five factors only. The columns are as follows:

1. *Main specification:* From Tables 3 and 4.

2. *New Seed:* We specify an alternate randomization seed for the selection of folds which are used to select parameters and estimate forecast error rates using 2010 outcomes.

Table C.2: Lasso model robustness checks

(a) Simulated forecast (2010)

	Main		Dummies Not	10-Fold Cross-			Subset (30)
Performance metric	Specification	New Seed	Standardized	Validation	$\alpha = 1$	$\alpha = .5$	from OLS
AUC	0.58	0.58	0.56	0.58	0.58	0.57	0.61
True positives (sensitivity)	77%	75%	68%	80%	76%	74%	83%
True negatives (specificity)	40%	40%	43%	42%	40%	38%	30%
Overall accuracy	47%	46%	47%	49%	46%	45%	39%
Ratio of false + to true +	3.71	3.84	4.06	3.45	3.82	3.96	4.02
Ratio of false - to true +	0.30	0.34	0.49	0.25	0.33	0.35	0.20
Violence predicted (% villages)	63%	63%	59%	62%	63%	64%	72%
False negatives (% villages)	4%	4%	6%	3%	4%	4%	3%

	Main		Dummies Not	10-Fold Cross-			Subset (30)
Performance metric	Specification	New Seed	Standardized	Validation	$\alpha = 1$	$\alpha = .5$	from OLS
AUC	0.65	0.66	0.66	0.61	0.66	0.58	0.68
True positives (sensitivity)	88%	88%	88%	88%	88%	88%	88%
True negatives (specificity)	22%	25%	33%	17%	25%	16%	24%
Overall accuracy	33%	36%	42%	29%	35%	28%	35%
Ratio of false + to true +	4.49	4.31	3.86	4.77	4.34	4.83	4.37
Ratio of false - to true +	0.14	0.14	0.14	0.14	0.14	0.14	0.14
Violence predicted (% villages)	79%	77%	70%	83%	77%	84%	78%
False negatives (% villages)	2%	2%	2%	2%	2%	2%	2%

(b) 2012 Forecasts

	Main		Dummies Not	10-Fold Cross-			Subset (30)
Performance metric	Specification	New Seed	Standardized	Validation	$\alpha = 1$	$\alpha = .5$	from OLS
Minority tribe in town leadership	1	1	1	1	1	1	1
Town Population	2	2	10	2	2	2	5
Percent believing other tribes are violent	3	3	3	3	3	3	2
Percent in dominant group	4	4	8	4	4	5	6
Percent who contribute to public facilities	5	5	4		5	4	3
Percent of town farmers			2			9	4
Participation in war violence			5				9

(c) Risk factor rankings

3. *Classification:* Observations on a given leaf are classified by majority vote (with a weight given to positive votes that is chosen by cross-validation). For all other specifications, observations on a given leaf are assigned a predicted probability of violence and are then classified based on a discrimination threshold.

4. *10-fold Cross Validation:* We identify optimal parameters and estimate forecast error rates using 10-fold cross validation rather than 5-fold cross validation.

5. *10 nodes:* We limit each tree in the forest to have no more than 10 nodes. For all other models, we limit trees to 5 nodes.

6. *Trees Fit to Larger Sample:* We fit each tree to a random sample of 36 observations rather than 24, the sample size for all other models.

7. *10,000 Tree Forests:* We compose the Random Forest from 10,000 trees rather than 1,000 trees as we do for all other models.

8. *Subset (30) from OLS:* We first fit an OLS model to the 2008/2010 data to determine the 30 coefficients of greatest magnitude. We then use only those 30 predictors for the model.

9. *Subset of 5 Lasso variables:* We use only the 5 risk factors selected by the main Lasso model, listed in Table 5.

Neural networks

Table C.4 reports various robustness checks for the simulated forecasts (Panel A) and true forecast results (Panel B). The columns are as follows:

1. *Main specification:* From Tables 3 and 4.

Table C.3: Random forests model robustness checks

(a) Simulated forecasts (2010)

	Main spec-	New		10-fold cross-	10	Trees fit to	10,000 tree	Subset (30)	Subset (5)
Performance metric	ification	seed	Classification	validation	nodes	larger sample	forests	from OLS	from lasso
AUC	52%	52%		52%	52%	53%	52%	51%	67%
True positives (sensitivity)	52%	54%	57%	52%	52%	53%	52%	49%	77%
True negatives (specificity)	50%	50%	48%	50%	50%	50%	50%	50%	45%
Overall accuracy	50%	50%	50%	50%	50%	50%	50%	50%	50%
Ratio of false + to true +	4.62	4.50	4.35	4.59	4.62	4.52	4.59	4.84	3.43
Ratio of false - to true +	0.94	0.88	0.77	0.93	0.94	0.89	0.93	1.05	0.31
Violence predicted (% villages)	50%	51%	53%	50%	50%	51%	50%	49%	59%
False negatives (% villages)	8%	8%	7%	8%	8%	8%	8%	9%	4%

	Main spec-	New		10-fold cross-	10	Trees fit to	10,000 tree	Subset (30)	Subset (5)
Performance metric	ification	seed	Classification	validation	nodes	larger sample	forests	from OLS	from lasso
AUC	0.60	0.62		0.61	0.62	0.59	0.62	0.61	0.63
True positives (sensitivity)	65%	75%	80%	80%	85%	83%	85%	73%	85%
True negatives (specificity)	41%	34%	37%	27%	27%	19%	29%	29%	31%
Overall accuracy	45%	40%	44%	36%	37%	30%	38%	36%	40%
Ratio of false + to true +	4.62	4.47	3.97	4.59	4.32	4.94	4.21	4.97	4.12
Ratio of false - to true +	0.54	0.33	0.25	0.25	0.18	0.21	0.18	0.38	0.18
Violence predicted (% villages)	60%	68%	66%	74%	75%	81%	73%	71%	72%
False negatives (% villages)	6%	4%	3%	3%	2%	3%	2%	5%	2%

(b) 2012 Forecasts

	Main spec-	New		10-fold cross-	10	Trees fit to	10,000 tree	Subset (30)	Subset (5)
Performance metric	ification	seed	Classification	validation	nodes	larger sample	forests	from OLS	from lasso
Town population	1	1	1	1	1	1	1		2
Mean educational attainment	2	6	6	4	8	4	6	2	
Percent in dominant group	3	5	3	2	3	2	4	1	4
Percent reporting loss of land in war	4	4	5	3	2	3	3		
Std. dev. of wealth index	5	10	13	22	16	13	11	7	
Number of households	6	3	9	14	5	7	5		
Number of tribes	7	2	2	5	4	5	2		
Wealth index	8	7	4	6	14	12	7	21	
Minority tribe in town leadership	9	8	8	10	19	6	8	5	

(c) Risk factor rankings

2. New Seed: We change the randomization seed to get different cross-validation runs and fit our models using different (randomly selected) initial weights.

3. 10-Fold Cross-Validation: We chose at threshold and estimate forecast error using 10-fold cross-validation rather than 5-fold, as in our preferred model.

4. Size = 10: We use 10 nodes in our hidden layer rather than 5, as in our preferred model.

5. Low Decay: We force our weights to decay at a rate of 0.01 rather than 0.1 as in our preferred model.

6. High Decay: We force our weights to decay at a rate of 0.5 rather than 0.1 as in our preferred model.

7. Subset (30) from OLS: We first fit an OLS model to the 2008/2010 data to determine the 30 coefficients of greatest magnitude. We then use only those 30 predictors for the model.

8. *Subset of 5 Lasso variables:* We use only the 5 risk factors selected by the main Lasso model, listed in Table 5.

D.3 Visualization of lasso accuracy

We provide another way to visualize the trade-off between true and false positives for the lasso model in the histogram in Figure C.1. Each bar represents the predicted probability of violence in one community in 2012. The discrimination threshold is the probability above which we predict violence will occur—the optimal threshold identified through our simulated forecasts above. Two features of the histogram are noteworthy. First, while the number of false positives is relatively high, the number

Table C.4: Neural networks model robustness checks

(a) Simulated forecasts (2010)

Performance metric	Main specification	New seed	10-fold cross-validation	Size = 10	Low decay (.01)	High decay (.5)	Subset (30) from OLS	Subset (5) from lasso
AUC	0.53	57%	56%	55%	55%	59%	64%	67%
True positives (sensitivity)	62%	64%	63%	60%	59%	71%	72%	79%
True negatives (specificity)	48%	48%	48%	49%	49%	46%	46%	44%
Overall accuracy	51%	51%	50%	51%	51%	50%	50%	50%
Ratio of false + to true +	3.99	3.91	3.96	4.11	4.18	3.65	3.58	3.37
Ratio of false - to true +	0.62	0.57	0.59	0.68	0.71	0.42	0.39	0.27
Violence predicted (% villages)	54%	54%	54%	53%	53%	57%	57%	60%
False negatives (% villages)	7%	6%	6%	7%	7%	5%	5%	4%

Performance metric	Main specification	New seed	10-fold cross-validation	Size = 10	Low decay (.01)	High decay (.5)	Subset (30) from OLS	Subset (5) from lasso
AUC	0.60	0.68	0.68	0.67	0.65	0.65	0.62	0.65
True positives (sensitivity)	63%	70%	70%	65%	65%	68%	63%	70%
True negatives (specificity)	52%	56%	57%	53%	61%	51%	50%	43%
Overall accuracy	54%	58%	59%	55%	62%	54%	52%	48%
Ratio of false + to true +	3.84	3.18	3.11	3.65	3.04	3.63	4.04	4.11
Ratio of false - to true +	0.60	0.43	0.43	0.54	0.54	0.48	0.60	0.43
Violence predicted (% villages)	50%	48%	48%	50%	43%	52%	52%	59%
False negatives (% villages)	6%	5%	5%	6%	6%	5%	6%	5%

(b) 2012 forecasts

of false negatives is very low. This is by design: in training the lasso model to maximize true positives, we also train it to minimize false negatives, subject only to the constraint that overall accuracy remain at or above 50%. Second, many of these false positives have relatively high predicted probabilities of violence; indeed, of the 242 communities in our sample, the two with the highest predicted probabilities are in fact false positives. This pattern does not necessarily imply that the model is inherently flawed, or that the risk of violence in these communities is in fact lower than it appears. Estimates of risk are by nature probabilistic; today's false positive may be tomorrow's true positive.

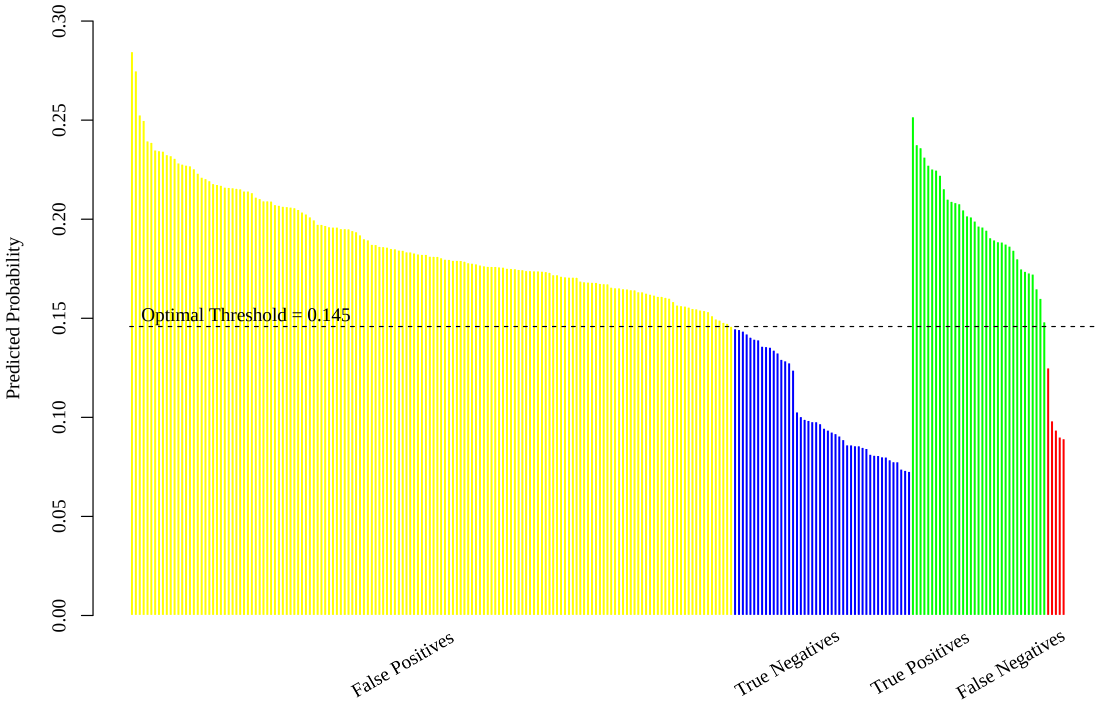
D.4 Reconciliation of lasso to logit

Figure C.2 offers a visualization of the relationship between lasso and logit. λ is the penalty placed on the sum of the magnitudes of coefficients for included variables. As we move along the x-axis and λ decreases, the model becomes more flexible, and the number of predictors included in the model (fitted with non-zero coefficients) increases. When $\lambda = 0$, our lasso objective function is just the logit objective function. The figure below shows that our optimally chosen λ is relatively restrictive. Cross-validation reveals that most of our available predictors add more noise than signal, and, consequently, we heavily penalize model complexity.

D.5 Ensemble predictions (Model averaging)

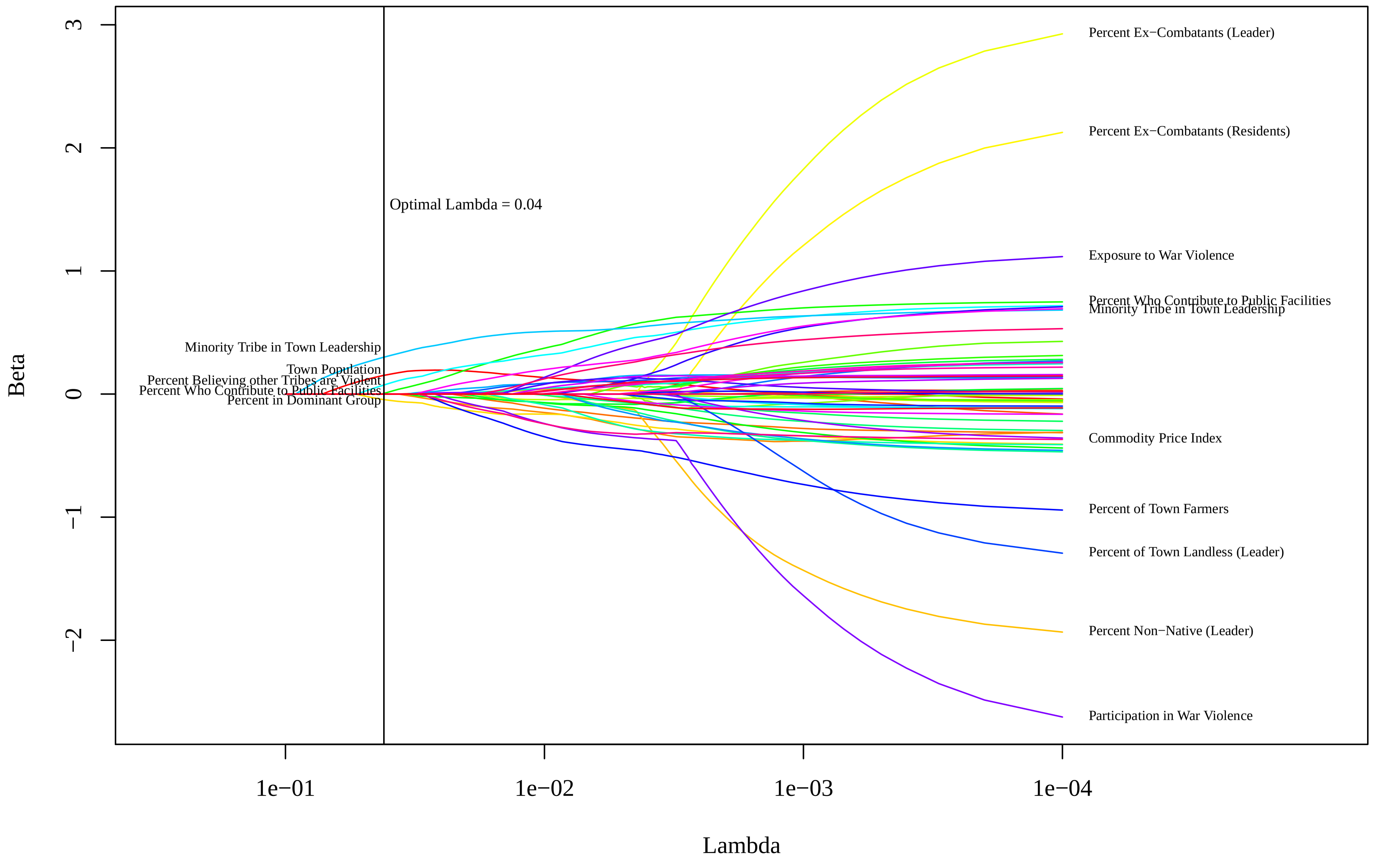
Rather than choosing the "best" model, here we consider ensemble prediction methods. One promising approach is Ensemble Bayesian Model Averaging (BMA). This method does not seem feasible in our case, as we do not have enough cross-sections of data to both train and calibrate our models. Also, it is not clear that BMA is as relevant to

Figure C.1: Predicted probabilities of 2012 violence, lasso model, by prediction accuracy



Notes: We apply the parameters from the lasso model estimated in Table 3 to 2010 data calculate the predicted probabilities of violence in 2012. The dotted line is the optimal threshold above which we predict violence, estimated in the same exercise (via 200 5-fold cross-validation trials).

Figure C.2: Lasso coefficients as penalty for additional variables is relaxed



our problem, where we care more about sensitivity than accuracy. BMA weights are functions of log likelihoods, which are themselves functions of accuracy, not sensitivity.

A simpler "majority vote" method takes the binary predictions from each of the four models (logit, lasso, random forests, and neural networks) and generates a single prediction according to what the majority predicts. Since we have an even number of models, we code ties as a prediction of violence. This is consistent with our overall approach of erring on the side of sensitivity over specificity. This model performs similarly to our best models, without major improvement over them.

An alternative ensemble method is to stack the models using a logistic regression. takes predicted probabilities from the four models and generates a single predicted probability using a logistic regression model. Then that probability is translated into a 0,1 prediction using a discrimination threshold chosen by cross validation. It has the downside, like BMA, of giving equal weight to false negatives and false positives. Logistic stack performs surprisingly poorly.

D.6 Risk factors

Table C.6: Full Risk Factor Rankings

	Lasso		Random Forests		Logit	
Risk factor	Rank	Coeff.	Rank	Importance	Rank	Coeff.
Town Population	2	0.15	1	0.00211	47	-0.06
Number of Households			6	0.00045	50	-0.05
Number of Tribes			7	0.00039	52	0.03
Percent Muslims (Leader)			45	-0.00010	33	-0.20
Has Mosque (Resident)			40	-0.00004	23	-0.32
Percent Muslims (Resident)			35	-0.00001	27	-0.29
Percent Non-Native (Residents)			13	0.00027	45	-0.07

	Lasso		Random Forests		Logit	
Risk factor	Rank	Coeff.	Rank	Importance	Rank	Coeff.
Percent Non-Native (Leader)			26	0.00005	4	-2.01
Percent in Dominant Group	4	-0.05	3	0.00064	18	-0.42
Percent Ex-Combatants (Residents)			21	0.00012	3	2.29
Percent Ex-Combatants (Leader)			53	-0.00032	1	3.09
Percent Returned for Internal Displacement			29	0.00002	44	-0.08
Percent under 30			41	-0.00006	55	-0.01
Percent Male			48	-0.00018	49	-0.05
Mean Educational Attainment			2	0.00082	51	0.04
Percent With No Education			16	0.00019	15	0.47
Proportion Receiving Any Peace Training			55	-0.00037	46	-0.06
Group Participation (0-9)			10	0.00032	22	0.33
Percent Who Contribute to Public Facilities	5	0.01	24	0.00009	8	0.76
Percent Saying Town is Safe at Night			14	0.00025	17	-0.46
Percent Saying Neighbors Are Helpful			27	0.00005	48	0.06
Collective Public Goods			33	-0.00001	25	0.29
Percent Who Rely on NGOs			54	-0.00036	31	-0.23
Percent Who Rely on Government			22	0.00012	24	-0.31
Perceived Equity in Institutions			44	-0.00010	14	-0.48
Percent Describing Police/Courts as Corrupt			43	-0.00009	19	-0.42
Percent Accepting Inter-Racial Marriage			38	-0.00003	30	0.26
Percent Who Say Muslims Shouldn't Be Leaders			18	0.00017	41	-0.11
Percent Believing other Tribes are Violent	3	0.07	11	0.00031	10	0.73
Percent Believing Other Tribes are Dirty			12	0.00028	40	-0.12
Minority Tribe in Town Leadership	1	0.30	9	0.00034	12	0.69
Percent Reporting Burglary or Robbery			19	0.00017	39	0.13
Percent Reporting Assault			17	0.00018	16	-0.47
Percent Reporting Any Land Conflict			15	0.00020	26	0.29
Percent Reporting Any Major Destabilizing Event			25	0.00007	37	0.14
Percent of Town Landless (Leader)			32	0.00000	5	-1.38
Percent of Town Landless (Residents)			37	-0.00002	43	-0.10

	Lasso		Random Forests		Logit	
Risk factor	Rank	Coeff.	Rank	Importance	Rank	Coeff.
Percent of Town Farmers			20	0.00014	7	-0.98
Unemployment Rate			51	-0.00025	56	0.00
Wealth Index			8	0.00039	9	0.73
S.D. of Wealth Index In Town			5	0.00056	54	0.01
Exposure to War Violence			56	-0.00039	6	1.16
Participation in War Violence			30	0.00001	2	-2.80
Percent Reporting Loss of Land During War			4	0.00060	20	-0.38
Percent Displaced During War			31	0.00001	38	0.13
Social Services in Town			47	-0.00014	35	0.16
Police or Magistrate in Town			28	0.00004	34	-0.17
Frequency of Police/ NGO Visits			23	0.00011	11	0.72
Town>1 Hour from Road			46	-0.00011	28	0.27
Mobile Phone Coverage			42	-0.00007	29	0.27
Less than 2 Radio Stations			36	-0.00002	32	0.23
Natural Resources In 2 Hours			39	-0.00004	21	-0.37
Commodity Price Index			34	-0.00001	13	0.55
Percent Affected By Human Disease			49	-0.00022	53	0.02
Percent Affected By Livestock Disease			50	-0.00025	36	0.15
Percent Affected By Crop Failure			52	-0.00028	42	-0.11

D.7 Relationship between different forms of violence

Table C.7 expands Table 6 and examines the extent to which two of our categories of violence can predict the third, either in 2010 or 2012.

D.8 Disaggregated violence

Table C.8 expands Table 7 to include full performance statistics.

Table C.5: Alternate ensemble methods

(a) Simulated forecast (2010)

	Logistic stack	Model Average	Majority vote
Performance metric	(1)	(2)	(3)
AUC	0.50	0.55	
True positives (sensitivity)	45%	59%	72%
True negatives (specificity)	54%	49%	38%
Overall accuracy	53%	50%	44%
Ratio of false + to true +	4.97	4.19	4.17
Ratio of false - to true +	1.36	0.71	0.40
Violence Predicted (% of Villages)	46%	53%	64%
False Negatives (% of Villages)	9%	7%	5%

	Logistic stack	Model Average	Majority vote
Performance metric	(1)	(2)	(3)
AUC	0.56	0.65	
True positives (sensitivity)	73%	83%	90%
True negatives (specificity)	29%	39%	20%
Overall accuracy	36%	46%	32%
Ratio of false + to true +	4.93	3.76	4.47
Ratio of false - to true +	0.38	0.21	0.11
Violence Predicted (% of Villages)	71%	65%	81%
False Negatives (% of Villages)	5%	3%	2%

(b) 2012 Forecasts

Table C.7: Relationship between different types of violence

(a) Training the 2008-10 model on pairs of categories (Simulated forecast)

	Interpersonal + extrajudicial (38)			Interpersonal + collective (37)			Collective + extrajudicial (16)		
Performance metric	Lasso	RF	NN	Lasso	RF	NN	Lasso	RF	NN
AUC	0.57	0.54	0.58	0.57	0.53	0.58	0.50	0.50	0.46
True positives (sensitivity)	80%	57%	65%	78%	53%	66%	53%	50%	43%
True negatives (specificity)	33%	49%	48%	32%	50%	48%	49%	51%	51%
Overall accuracy	40%	50%	50%	39%	50%	51%	49%	51%	51%
Ratio of false + to true +	4.53	4.81	4.37	4.81	5.25	4.41	14.26	14.37	17.01
Ratio of false - to true +	0.26	0.76	0.56	0.29	0.89	0.53	0.99	1.07	1.47
Violence predicted (% villages)	69%	52%	54%	69%	51%	54%	51%	49%	48%
False negatives (% villages)	3%	7%	6%	3%	7%	5%	3%	3%	4%

(b) Success at predicting the excluded category of violence in 2010 (Current forecast)

	Interpersonal + extrajudicial (9)			Interpersonal + collective (8)			Collective + extrajudicial (32)		
Performance metric	Lasso	RF	NN	Lasso	RF	NN	Lasso	RF	NN
AUC	0.60	0.54	0.52	0.45	0.48	0.44	0.50	0.54	0.50
True positives (sensitivity)	78%	58%	59%	58%	42%	51%	51%	57%	49%
True negatives (specificity)	31%	48%	46%	30%	49%	46%	49%	52%	52%
Overall accuracy	33%	49%	46%	31%	49%	46%	49%	53%	51%
Ratio of false + to true +	23.00	23.42	24.55	35.95	38.19	33.08	6.60	5.65	6.53
Ratio of false - to true +	0.29	0.75	0.76	0.77	1.57	1.08	0.98	0.79	1.06
Violence predicted (% villages)	69%	52%	54%	69%	51%	54%	51%	49%	48%
False negatives (% villages)	1%	2%	2%	1%	2%	2%	6%	6%	7%

(c) Success at predicting the excluded category of violence in 2012 (True forecast)

	Collective Violence (7)			Extrajudicial violence (16)			Interpersonal Violence (21)		
Performance metric	Lasso	RF	NN	Lasso	RF	NN	Lasso	RF	NN
AUC	0.70	0.62	0.58	0.58	0.50	0.52	0.58	0.46	0.47
True positives (sensitivity)	86%	86%	71%	88%	63%	56%	67%	71%	38%
True negatives (specificity)	20%	32%	52%	22%	35%	46%	40%	19%	53%
Overall accuracy	22%	34%	53%	26%	36%	46%	42%	24%	52%
Ratio of false + to true +	31.17	26.50	22.40	12.57	14.80	13.67	9.50	11.87	13.00
Ratio of false - to true +	0.17	0.17	0.40	0.14	0.60	0.78	0.50	0.40	1.63
Violence predicted (% villages)	80%	68%	48%	79%	65%	55%	61%	80%	46%
False negatives (% villages)	0%	0%	1%	1%	2%	3%	3%	2%	5%

Table C.8: Simulated and true out-of-sample tests of prediction accuracy for disaggregated violence

	Collective Violence (9)			Extrajudicial Violence (8)			Interpersonal Violence (32)		
Performance metric	Lasso	RF	NN	Lasso	RF	NN	Lasso	RF	NN
AUC	0.48	0.65	0.58	0.42	0.47	0.38	0.65	0.57	0.67
	(0.07)	(0.05)	(0.07)	(0.08)	(0.06)	(0.06)	(0.03)	(0.02)	(0.03)
True positives (sensitivity)	48%	73%	60%	40%	42%	29%	85%	61%	81%
	(0.14)	(0.11)	(0.13)	(0.17)	(0.15)	(0.11)	(0.06)	(0.05)	(0.06)
True negatives (specificity)	51%	50%	50%	47%	51%	51%	39%	49%	46%
	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
Overall accuracy	51%	51%	51%	47%	51%	51%	45%	50%	51%
	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
Ratio of false + to true +	29.90	18.07	22.68	49.24	39.36	60.03	4.73	5.54	4.39
	(14.44)	(3.26)	(7.13)	(28.63)	(20.22)	(29.57)	(0.35)	(0.52)	(0.37)
Ratio of false - to true +	1.38	0.40	0.77	2.21	1.78	3.22	0.18	0.64	0.24
	(1.21)	(0.27)	(0.57)	(1.94)	(1.44)	(2.08)	(0.09)	(0.14)	(0.09)
Violence predicted (% villages)	49%	51%	50%	53%	48%	48%	65%	53%	58%
	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
False negatives (% villages)	2%	1%	1%	2%	2%	2%	2%	5%	2%
	(0.01)	(0)	(0)	(0.01)	(0)	(0)	(0.01)	(0.01)	(0.01)

(a) Simulated forecast (2010)

	Collective Violence (7)			Extrajudicial violence (21)			Interpersonal Violence (16)		
Performance metric	Lasso	RF	NN	Lasso	RF	NN	Lasso	RF	NN
AUC	0.48	0.52	0.66	0.71	0.52	0.61	0.61	0.60	0.71
True positives (sensitivity)	57%	71%	71%	88%	88%	69%	81%	71%	81%
True negatives (specificity)	46%	26%	56%	46%	19%	56%	30%	29%	47%
Overall accuracy	47%	27%	56%	49%	23%	57%	34%	33%	50%
Ratio of false + to true +	31.50	34.80	20.80	8.71	13.14	9.00	9.12	10.40	6.88
Ratio of false - to true +	0.75	0.40	0.40	0.14	0.14	0.45	0.24	0.40	0.24
Violence predicted (% villages)	54%	74%	45%	56%	82%	45%	71%	71%	55%
False negatives (% villages)	1%	1%	1%	1%	1%	2%	2%	2%	2%

(b) 2012 forecasts

Notes: RF is random forests and NN is neural networks.