NBER WORKING PAPER SERIES

THE PROMISE AND PITFALLS OF CONFLICT PREDICTION: EVIDENCE FROM COLOMBIA AND INDONESIA

Samuel Bazzi
Robert A. Blair
Christopher Blattman
Oeindrila Dube
Matthew Gudgeon
Richard Merton Peck

Working Paper 25980
http://www.nber.org/papers/w25980

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2019

We thank seminar participants at ESOC, MWIEDC, ISF, NBER Economics of National Security, and NEUDC for helpful feedback. Miguel Morales-Mosquera provided excellent research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia
Samuel Bazzi, Robert A. Blair, Christopher Blattman, Oeindrila Dube, Matthew Gudgeon, and Richard Merton Peck
NBER Working Paper No. 25980
June 2019
JEL No. C52,C53,D74

ABSTRACT

Policymakers can take actions to prevent local conflict before it begins, if such violence can be accurately predicted. We examine the two countries with the richest available sub-national data: Colombia and Indonesia. We assemble two decades of fine-grained violence data by type, alongside hundreds of annual risk factors. We predict violence one year ahead with a range of machine learning techniques. Models reliably identify persistent, high-violence hot spots. Violence is not simply autoregressive, as detailed histories of disaggregated violence perform best. Rich socio-economic data also substitute well for these histories. Even with such unusually rich data, however, the models poorly predict new outbreaks or escalations of violence. "Best case" scenarios with panel data fall short of workable early-warning systems.

Samuel Bazzi
Department of Economics
Boston University
270 Bay State Road
Boston, MA 02215
and CEPR
and also NBER
sbazzi@bu.edu

Robert A. Blair
Brown University
Department of Political Science
Watson Institute for International and Public Affairs
Providence, RI 02912
robert_blair@brown.edu

Christopher Blattman
Harris School of Public Policy
The University of Chicago
1307 E 60th St
Chicago, IL 60637
and NBER
blattman@uchicago.edu

Oeindrila Dube
University of Chicago
Harris School of Public Policy
1307 E 60th St
Chicago, IL 60637
and NBER
odube@uchicago.edu

Matthew Gudgeon
OEMA & Department of Social Sciences
607 Cullum Road
West Point, NY 10966
matthew.gudgeon@westpoint.edu

Richard Merton Peck
Northwestern University
2211 Campus Dr
Evanston, IL 60208
RichardPeck2021@u.northwestern.edu

1 Introduction

Advances in data and computing techniques have kindled hopes that civil society, police, or peacekeepers could predict costly violence ahead of time. Such early-warning systems could be used to target scarce security personnel and resources, and prevent violence from occurring or escalating.

Until recently, prediction focused on large-scale, country-level events, including coups, civil wars, and terror attacks.[1] These macro-level efforts have informed policy, the science of prediction, and our understanding of violence. But such high-level predictions are not easy to act on. Scholars such as Cederman and Weidmann (2017) argue that country-level conflict predictions are unlikely to improve much in the future: there is simply too much complexity and randomness, they argue, to develop reliable forecasts over such wide time and space.

Sub-national data or higher-frequency predictions could prove more fruitful. The past decade has seen the study of conflict push down to the micro-level causes, processes, and consequences, and we can avail ourselves of these data to investigate prediction. If governments, police, or peacekeepers can reliably predict what places will see escalations of violence, for instance, they may be able to act to prevent it. Policy options to prevent an ethnic riot or local unrest are likely better than policy options to prevent a civil war. The feasibility of these early warning systems are unknown, however. Now is a good moment to take stock of what existing methods and the richest available micro data can deliver.

This paper takes advantage of high-quality and extensive data in two countries, Colombia and Indonesia. Both countries have been ravaged by violence for decades—a situation that typically does not bode well for data availability. But both countries are also wealthy enough (and have strong enough states and research communities) to produce some of the highest-

[1]The Political Instability Task Force's prediction efforts are likely the most well known (Goldstone et al., 2010). For other examples of cross-national prediction studies, see Beck et al. (2000); Brandt et al. (2011); Celiku and Kraay (2017); Gleditsch and Ward (2013); Gurr and Lichbach (1986); Harff (2003); Hegre et al. (2013, 2016); Perry (2013); Ward et al. (2013). For an early exception see Schrodt (2006) who studies violence in the Balkans.

quality local data in the developing world. This includes a trove of information on local socioeconomic conditions and other characteristics, plus at least a decade of subdistrict- or municipal-level data on local violence.

We chose these two cases because they are among the current "best case" scenarios in terms of both micro-level data on violent events, as well as a wide range of predictors of violence, in panel form. If conflict prediction proves fruitful in these two cases, they could be models for other prediction efforts. If not, then we must ask where or with what additional data we can expect early warning systems to bear fruit. Finally, both countries have suffered recurring episodes of local violence during transitions to national peace. Anticipating and preventing these episodes is of both substantive and practical importance.

We identified, collected, and merged dozens of subnational datasets in each country. This gives us an unusually rich array of hundreds of covariates per locality, including covariates that the empirical and theoretical literatures commonly associate with conflict onset and escalation (Blattman and Miguel, 2010). This data gathering also gives us multiple measures of the outcome we are trying to predict. For instance, using data from 1998 to 2014 in Indonesia, we are able to study conflicts related to interethnic or religious tensions, as well as electoral and resource disputes. In Colombia, our data span 1988 to 2005. We predict clashes between state, guerrilla, and paramilitary forces during a period of protracted civil conflict.

We then deploy several machine learning methods to generate predictions of local violent incidents at the annual level. In our main year-ahead predictions, we train the algorithms on six to fourteen years of data, and forecast local conflict during the following year. We also examine predictive power across space as well as time, and with new outbreaks of violence as well as escalations.

Our results illustrate both the promise and pitfalls of local violence forecasting. An ensemble of machine learning models effectively identifies locations at risk of having a violent incident. We are particularly effective at identifying "hot spots" with high concentrations

of violence, defined as five or more incidents in a single year. Indeed, our ensemble model, which leverages the best new methods, performs better than previous sub-national attempts (Blair et al., 2017; Colaresi et al., 2016; Weidmann and Ward, 2010; Witmer et al., 2017). We view these results as especially important given that such local hot spots can pose an especially serious risk of regional or national escalation.

We find that local violence is not merely autoregressive, as a model consisting of the lagged dependent variable alone performs consistently poorly. Rather, our algorithms' strong performance is mainly driven by forecasts of where, but not when, violence is likely to occur. While a simple lagged dependent variable model yields disappointing results, more nuanced histories of violence tend to be the best predictors of hot spots in particular. In other words, to predict future violence, it is not enough to know where violence occurred in the past. But detailed and disaggregated histories of violence—including the severity of particular incidents (e.g., number of deaths, property damage) and the identity of the actors involved—perform very well.

Even without such detailed and disaggregated histories, however, the available covariates also predict hot spots almost equally well. This suggests that much of the information contained in these violence histories is representative of observable characteristics of the units in our two samples. The most predictive risk factors tend to be slow-moving or time-invariant. In Colombia, for example, one of the most reliable predictors is terrain ruggedness. In Indonesia, robust predictors include religious and ethnic diversity as well as sectoral shares of the local economy.

Importantly, predictive accuracy improves little when we add time-varying factors, including natural disasters, elections, and fluctuations in rainfall, temperature, commodity prices and drug production. This lies in contrast to a large causal literature on conflict, where an array of findings associate economic and political shocks to intensified violence (Miguel et al., 2004; Blattman and Miguel, 2010; Dube and Vargas, 2013; Berman and Couttenier, 2013; Bazzi and Blattman, 2014; Burke et al., 2015; Wright and Signoret, 2016).

Our models also perform poorly when we attempt to predict annual deviations from average levels of violence over the study period. That is, even rich histories of violence, rich covariates, and the most common economic and political shocks do not help us identify what hot spots are likely to get hotter in the coming year.

In contrast, the models perform well when we forecast conflict across space, using training data from all available years in one set of locations to predict conflict in another set of locations. In these cross-location predictions, time-varying shocks typically improve performance. The models leverage the longest time series currently available at the local level. This leads us to believe that a lack of common support in the training and testing periods may explain the limited predictive performance of time-varying shocks when we attempt to forecast conflict over time. Thus, early warning performance should improve over time, but this (by definition) means that better one-year-ahead predictions are a long ways off.

Taken together, our results are both encouraging and disappointing. On the one hand, we are able to predict hot spots for local violence remarkably accurately, and much more accurately than previous exercises of this sort. Anticipating where violence is most likely to occur is potentially highly valuable to resource-constrained governments in conflict-affected states. On the other hand, early warning systems would ideally be able to predict not just the location but also the timing of new outbreaks of violence. Our inability to do so, with some of the richest and most systematic subnational panel data in the developing world, is important but disheartening news for conflict forecasters.

One interpretation of our results is that local conflict prediction is less fruitful than hoped. This pessimistic conclusion would resonate with warnings that big data and machine learning may not deliver the precision that policymakers long for (Jasny and Stone, 2017, p. 469). It also aligns with a view that conflict breaks out for largely idiosyncratic reasons, as reflected in the warning by Gartzke (1999) that "war is in the error term".

Another interpretation is that early warning systems may yet be feasible, but require longer, more frequent data on additional or different risk factors. Longer training samples

could give algorithms more variation to train on, and in principle this could help them identify more complex relationships with time-varying predictors. If true, this implies that there will be large gains to collecting longer conflict time series, for example, by diving into past historical archives. There could also be gains from using higher-frequency data, or more data on new risk factors, including data from mobile phones, social media, or other forms of media.[2] These data innovations, and our increasing ability to collect newer and wider forms of big data, may enhance our capacity to forecast conflict over time.

Our paper establishes several results that future work should leverage to explore the promise of machine learning methods in conflict. Overall, conflict can be forecasted well across space, but not over time. While violence is not simply autoregressive, detailed conflict histories can substitute for a broader array of covariates, which are potentially more expensive to collect. But if such detailed histories of violence are unavailable (as is the case in most countries), a more limited set of common or easy-to-measure covariates can also predict hot spots remarkably well, at least in our two cases. Basic hot spot prediction systems are probably feasible in a wide range of countries, even if an escalation early warning system may not be.

We see these patterns consistently across two different country cases, with very different forms of violence. These findings suggest that the combination of newer and longer conflict time series, combined with machine learning methodologies, may be particularly instructive for improving conflict prediction over time.

[2] A case in point is work by Mueller and Rauh (2017) which successfully uses topics analyses from newspapers to forecast conflict within countries. Likewise, Berger et al. (2014) use cell phone call patterns to predict temporal variation in conflict.

2 Settings

2.1 Indonesia

Following the 1998 collapse of Suharto's authoritarian regime, Indonesia experienced large-scale collective violence.[3] Separatist movements in Aceh, East Timor (as it parted from Indonesia), and Papua resulted in over 10,000 deaths. At the same time, religious and ethnic conflict reached new highs.

Collective violence abated by 2003, and the separatist conflict in Aceh ended in 2005. Post-2004, there were far fewer fatalities, and the composition of violence shifted as electoral and resource-related violence rose. The violence also had different consequences: after 2004 it was more likely to lead to injuries and property damage than to deaths. Deadly violence nevertheless remained prevalent across the archipelago, primarily concentrated in regions with histories of large-scale violence.

Scholars debate the drivers of today's conflict in Indonesia, highlighting fixed factors like ethnicity and religion, as well as local resources like forests, minerals, and plantation crops. Regional variation in violence has been linked to political and economic shocks associated with decentralization and electoral reforms (Bazzi and Gudgeon, 2017; Pierskalla and Sacks, 2017), and with natural disasters, weather shocks, and commodity price fluctuations (Barron et al., 2009; Wright and Signoret, 2016). While the literature has identified a variety of proximate causes, it is not clear which of these factors are the best predictors of violence.

2.2 Colombia

Colombia's long-running civil war has resulted in 220,000 deaths and 25,000 disappearances, and forcibly displaced over five million civilians (Historical Memory Group, 2013).

During our analysis period, 1988–2005, the conflict mainly involved left-wing guerrilla groups, the government military, and right-wing paramilitary groups. The insurgency was

[3]For detailed accounts see Barron et al. (2014, 2016); Tadjoeddin (2014).

launched by communist guerrillas in the 1960s. Paramilitaries arose in the 1980s when landowners organized in response to extortion and violence perpetrated by the guerrillas. Paramilitaries and the government colluded extensively, though their relationship varied over time and space.

Low violence levels prevailed through the 1980s, but escalated in the 1990s when paramilitaries expanded and centralized authority. Intensity remained high until the paramilitaries demobilized, a process that began in 2003 and continued until 2006, when the main paramilitary organization officially disbanded. The conflict subsided further as the death of a number of guerrilla leaders weakened their respective groups. It drew to an official end in 2016, when the largest guerrilla group signed a peace deal with the government.

Scholars have linked regional variation in conflict to a host of political and economic factors, including shocks to drug production (Angrist and Kugler, 2008), fluctuations in commodity prices (Dube and Vargas, 2013), revenue decentralization (Chacon, 2014), collusion between paramilitaries and politicians (Acemoglu et al., 2013), American military aid (Dube and Naidu, 2015), and military incentives in the targeting of civilians (Acemoglu et al., 2016).

3 Data

An important contribution of this study are the two datasets we assembled. In each country we collected and stitched together dozens of local-level data sets, most of which had not been consolidated before. The result is a uniquely rich trove of data that we can draw on for purposes of prediction.

3.1 Indonesia

Our units of analysis are Indonesia's third-tier administrative units, known as subdistricts or *kecamatan*. The country had 7,094 subdistricts in 514 districts in 2014. These subdistricts

had a median population of around 22,000.[4] While districts are the key autonomous administrative units, responsible for providing major public goods, subdistricts are also important sites of political organization.

Subdistrict-Level Violence Data Our main measures of violence come from the Indonesian National Violence Monitoring System (known by its Indonesian acronym, SNPK). Coverage begins in 1998 for nine conflict-prone provinces and increases to 15 provinces plus parts of 3 provinces in greater Jakarta beginning in 2005. The data is not formally representative of Indonesia, but by 2005 it spans all major island groups and covers a majority of the population.

The SNPK is built from local media reports of violence. SNPK researchers collected all available print archives of 120 local newspapers, recording over 2 million images. Coders then used a standardized template to code each incident based on the underlying trigger, beginning with broad groupings: domestic violence, violent crime, violence during law enforcement, and conflict. Within conflict, the coders further sorted into identity, elections/appointments, governance, resource violence, popular justice, separatist, and other (could not be classified). Appendix C.1 (Appendix page 10) defines each of these.

We also draw on additional measures of violence from a triennial administrative census of villages known by its acronym, *Podes*. Podes asks local government officials about a host of village characteristics, including recent violent events.[5]

Outcome Measurement Our main outcome is an indicator for any "social conflict." This groups all of the various forms of violence, except domestic violence and crime, into one category. It guards against miscoding of conflict triggers. Predictive performance is similar when retaining domestic violence and crime.[6]

[4]To deal with Indonesia's pervasive administrative unit proliferation we harmonize all observations to boundaries in 2000.

[5]To the extent that local leaders face strategic incentives to misreport violence, *Podes* measures may be more biased than those from external media reporting (for discussion, see Barron et al., 2014).

[6]Results available upon request.

In addition to indicators of any social conflict (which occur in around half of the subdistricts each year) we also predict an indicator for at least five social conflict incidents in a given year. This is meant to capture higher intensity episodes. These episodes occur in around 10% of subdistricts each year. We predict indicators rather than counts in order to simplify the interpretation of performance: the models either correctly predict the incident or they do not. We predict counts in Online Appendix A.3 (Appendix page 4); this exercise does not meaningfully change our conclusions.

Of course, levels of violence tend to be persistent, and we are often interested in predicting the onset and escalation of violence after a period of peace. There is no natural definition of "onset" here, as in the civil war literature, since there are no discontinuities in subnational event-level data. Instead, we construct an indicator for a standard deviation increase in violence since the previous year, and seek to predict this escalation. A standard deviation increase is around 4.7 acts of violence in a year, and we observe an increase of this size in 3.3% of subdistrict–years.

Covariates In addition to lagged violence measures from SNPK and *Podes*, we assemble a set of over 400 subdistrict-level covariates from multiple data sources, several of them new. Details on sources and variable construction can be found in Online Appendix C.1 (Appendix page 10). Predictors include, among others: (a) population and population density; (b) age, religious, and ethnic composition; (c) topographical traits and transportation infrastructure; (d) local commodity production, mine locations, and relevant commodity prices; (e) sectoral shares of output; (e) economic output measured by unemployment, light intensity, and district-level GDP; (f) rainfall and temperature histories and fluctuations; (g) local and national election outcomes; (h) subdistrict and district public revenues and expenditures, and (i) local security personnel and posts. Unless otherwise specified, these measures are available at the subdistrict level or finer, and are aggregated to their subdistrict boundaries in 2000.

3.2 Colombia

Our unit of analysis for Colombia is the municipality. There are 1,023 in our study period, averaging 37,000 residents (slightly larger than the typical Indonesian subdistrict). Municipalities play a key role in the allocation and contestation of public resources in Colombia.

Municipality-Level Violence Data The Conflict Analysis Resource Center (CERAC) provides data on armed confrontations from 1988 to 2005, collected from 25 major newspapers and supplemented by reports from a network of Catholic priests. The priests are seen as neutral actors, often serving as negotiators between the two sides. Their accounts are crucial for remote areas. CERAC cross-checks the events in this database against National Police, Human Rights Watch, and other sources (see Restrepo et al., 2004).

CERAC codes events as bilateral clashes between sides or unilateral attacks by any one side against another. Clashes occur between all three, though government versus paramilitary clashes are rare.

We only use CERAC data after 1992 for the training sample, since a consistent set of covariates is unavailable before then.

Outcome Measurement We construct indicators of any attack or clash, analogous to the indicators for Indonesia. This grouping combines attacks initiated by the government with attacks initiated by other armed actors. Results are similar when we remove government-initiated violence.[7]

In Colombia, our indicator of any conflict occurs in about one-third of municipalities each year. "Hotspots" with five or more incidents occur about 8% of the time. A standard deviation change is 3.4 events, and we observe such a change in 4.4% of municipality-years.

[7]Results available upon request.

Covariates As in Indonesia, we assemble a broad set of predictors from multiple sources: over 350 in all (including lags of time-varying predictors).[8] These include: (a) population and population density; (b) topographical traits and road accessibility; (c) military presence and U.S. military spending (Dube and Naidu, 2015); (d) local commodity production and relevant commodity prices (as in Dube and Vargas, 2013); (e) illicit drug production and prices; (f) measures of local poverty and inequality; (g) rainfall and temperature histories and fluctuations; (h) colonial population and infrastructure (Acemoglu et al., 2015); (i) annual municipality revenues and spending; and (j) electoral outcomes in local and federal elections.

4 Prediction Methods

4.1 Training and Testing

For each year t we forecast violence in year $t+1$. While the events are coded with specific dates, we aggregate to the annual level because few predictors are measured at sub-annual frequency, and because disaggregation would serve to exacerbate a class imbalance problem (i.e. the fact that there are far more non-events than events). Our procedure is as follows:

1. For each model, we take predictors measured from t_0 to $t-1$ as our training set. Correspondingly, violence measures up to and including period t are there training outcomes.

2. We use 5-fold cross validation to choose optimal "tuning parameters" specific to each machine learning algorithm (see Section 4.2). We choose tuning parameters to maximize out-of-sample area under the receiver operating characteristic, or ROC, curve (AUC). 5-fold cross validation simulates out-of-sample prediction. First, the data is randomly partitioned into 5 equal sized subsamples. A model is then fit to four subsamples and used to predict results for the fifth. This is repeated for each of the five

[8] See Online Appendix C.2 on Appendix page 16 for full details.

subsamples, so that there is an "out-of-sample" prediction of each observation. We replicate this exercise for each tuning parameter value in the parameter search space. The best performing parameter, in terms of AUC, is chosen.

3. We repeat step 2 ten times with different random partitions, in order to generate 10 "optimal" tuning parameters, and then we take the average over these 10 trials.

4. Using the selected tuning parameter values, we fit the model to the entire training set.

5. With this fitted model, we use the predictors measured up to year t and the estimated parameters to forecast violence in year $t + 1$.

We generate out-of-sample predictions starting in 2008 (and ending in 2014) for Indonesia and in 1998 for Colombia (ending in 2005). We use all data up to the test year to generate out-of-sample forecasts. So, to predict violence in year $t + 1$, we train each model on data through year t; to predict violence in $t + 2$, we train each model on data through $t + 1$; etc. For Indonesia, this procedure generates seven predictions per algorithm. The first uses six years of data to forecast conflict in 2008, and the last uses twelve years of data to forecast conflict in 2014. For Colombia, we generate eight predictions per algorithm. The first uses six years of data to forecast conflict in 1998, and the last uses fourteen years of data to forecast conflict in 2005.

4.2 Machine Learning Algorithms

We apply several machine learning methods to the above procedure. Since each has its own strengths and drawbacks discussed below, we also take a weighted average of the four using an **Ensemble Bayesian Model Average**.(Beger et al., 2016) Starting with the prior that each algorithm is equally appropriate, we use cross-validation to update our weights based on the accuracy of each model (Montgomery et al., 2012). Bayesian model averaging is especially important for our auxiliary analyses in which we explore different subsets of predictors and alternative prediction tasks. Since some procedures may be more or less

suited to particular tasks, the model average allows us to consider the potential of the suite of algorithms as a whole.

1. **LASSO** (Tibshirani, 1994) is a logistic regression model that penalizes large coefficients and forces all but the most important to zero. This algorithm is the simplest of the five that we test, and the least susceptible to overfitting. It is less suited to identifying complex relationships between covariates and outcomes. It is also most familiar to social scientists.

2. **Random Forests** comprise many independent decision trees. Each tree is a sequence of rules that splits the sample into subsets, called leaves, based on variable cutoffs. The prediction for each leaf is the mean outcome for the observations on that leaf, and trees are fit so as to minimize mean squared error. Each tree is constructed by sampling a random subset of the training data and a random subset of the predictors. Each of these trees generates a prediction, and the overall prediction of the Random Forest is the average of the predictions from each tree. Random forests are very flexible–able to model complicated interactions between variables. Random forests are also relatively straightforward in terms of choosing tuning parameters.

3. **Gradient Boosted Machines** are a variant of Random Forests. Trees are fit neither randomly nor independently. Instead, each tree is fit sequentially to the full dataset, but observations are weighted by the error rates of previous trees in the forest, such that later trees are fit with a larger weight on observations that previous trees found difficult to predict. In this way, each new tree slightly improves the model (Freund and Schapire, 1999). Gradient boosted machines can improve upon random forests by fitting trees in a more targeted manner, but they also require more decisions about tuning parameters and are more susceptible to overfitting.

4. **Neural Networks** consist of systems of "nodes," which are each functions of predictors. The functions input a linear combination of predictors and output a value

between zero and one. The outputs of these nodes are then further combined to produce a single output, with an organization evoking the structure of the human brain (Hastie et al., 2001). The optimization problem is to choose appropriate weights in each linear combination.[9] Neural networks are widely applied in industry and are best suited for the most complex classification tasks such as image and speech recognition. However, neural networks require relatively large datasets and computing resources to achieve high performance.

Online Appendix B (Appendix page 8) reports further details about hyper-parameter choices and the mechanics of each algorithm.

While there is an enormous variety of additional algorithms we might have tested, we focus on these four because they are well established in the machine learning literature, because they have been used (albeit infrequently) for purposes of forecasting in economics and political science, and because they reflect much of the variation across the most prominent categories of machine learning models: selection and shrinkage techniques (LASSO), ensemble and tree-based techniques (Random Forests and Gradient Boosted Machines), and non-linear adaptive weighting techniques (Neural Networks). Our goal is not to be exhaustive, but rather to evaluate the predictive power of well-established models applied to uniquely rich within-country data on conflict and its correlates.

4.3 Performance Metrics

To evaluate our models, we focus on the area under the ROC curve, although other performance metrics such as the mean squared error and accuracy are reported in Online Appendix A.1 (Appendix page 1). ROC curves plot the tradeoff between true and false positives for a given model. The area under the curve, or AUC, captures the probability that a randomly

[9]Because there is a separate set of weights for each node, the number of free parameters can grow very quickly. Since we do not have many observations relative to our predictor dimensionality, we must first cut down the number of predictors by taking principal components of the covariates. We use 30 principal components in Indonesia and 20 in Colombia.

chosen pair of observations is correctly ordered in terms of predicted risk of violence. A model that performs no better than chance would have an AUC of 0.5; a perfect model would have an AUC of 1.

An advantage of the AUC is that it does not require that we specify a probability threshold above which we predict violence will occur. Selecting a specific threshold requires making a tradeoff between accuracy, sensitivity (the proportion of incidents correctly predicted), and specificity (the proportion of non-incidents correctly predicted). The threshold one chooses depends on one's relative tolerance for false positives and false negatives.

For example, a policymaker with ample resources might choose a low threshold, increasing sensitivity at the cost of specificity and accuracy, while a policymaker with scarce resources might choose a high threshold, increasing specificity at the cost of sensitivity and accuracy. We are more interested in overall performance than in performance at any given threshold, and so opt to focus on the AUC. But we recognize that the AUC has some limitations as well, especially in the presence of class-imbalanced data, and report alternative performance metrics in the appendix.[10]

5 Results

5.1 Next Year's Violence is Predictable

Table 1 shows that all machine learning methods have strong predictive performance. For the ensemble average (EBMA), the AUC is above 0.82 for predicting $\geq$1 event, above 0.91 for $\geq$5 events, and above 0.79 for escalations of $\geq$1 standard deviation. In general, AUCs of 0.8 and above are considered very good, and AUCs of 0.90 and above are considered excellent.

To fix ideas, given a random pair of Indonesian subdistricts in which one location experi-

[10] An alternative metric that also does not require selecting a threshold is the mean squared error (MSE). Online Appendix A.1 (Appendix page 1) reports the MSE as well as accuracy, sensitivity, and specificity at two different thresholds (one that maximizes accuracy and one that maximizes sensitivity while keeping accuracy above 50%). Our results are qualitatively unchanged when we use the MSE to compare models and predictors rather than the AUC.

ences 5 or more incidents and the other does not, there is a 0.935 probability that the more violent subdistrict would have a higher predicted probability of violence. The lower AUC for escalations implies that changes are inherently more difficult to predict, and that increasing the number of true positives comes at the cost of more false positives.

By way of comparison, Blair et al. (2017) report a maximum out-of-sample AUC of 0.74 in a sample of 250 Liberian towns over three years, Weidmann and Ward (2010) achieve a maximum out-of-sample AUC of 0.78 in Bosnia 1992–95, and Witmer et al. (2017) find a maximum *in*-sample AUC of 0.85 across sub-Saharan Africa using 1 degree gridded monthly data, 1980–2012.[11] Gains in the range of 0.05 or 0.10 represent 10–20% of the difference between the worst and best possible prediction.

The models perform similarly well in Indonesia and Colombia, and performance is similar across algorithms. LASSO performs roughly as well as the more sophisticated algorithms, notable given its relative simplicity.

Table 1: Out-of-Sample (One Year Ahead) Performance of Prediction Models, Area Under the Curve (AUC)

	LASSO (1)	Random Forest (2)	Adaptive Boosting (3)	Neural Network (4)	EBMA (5)
	Indonesia (social conflict)				
Indicator of any violent event	0.806	0.81	0.814	0.78	0.814
Indicator of ≥ 5 violent events	0.922	0.927	0.932	0.91	0.935
≥ 1 s.d. increase in violent events	0.851	0.797	0.82	0.803	0.841
	Colombia (attacks and clashes)				
Indicator of any violent event	0.845	0.846	0.848	0.826	0.851
Indicator of ≥ 5 violent events	0.914	0.909	0.91	0.88	0.916
≥ 1 s.d. increase in violent events	0.803	0.787	0.792	0.747	0.798

Notes: Each model is trained on all available data preceding the out-of-sample prediction year. Training data starts in 1991 in Colombia and 2002 in Indonesia. Out-of-sample prediction begins in 1998 in Colombia and 2008 in Indonesia. The AUC is the area under the ROC curve, a measure of the trade-off between the true positive rate and false positive rate at different thresholds. We report average performance over the out-of-sample years.

[11] *In-sample* performance refers to models that are trained and tested on the same data. *Out-of-sample* refers to models that are trained on one subset of data and tested on another.

5.2 We Predict Time-Invariant Risk that Varies Over Space rather than Time

In this section, we run various analyses to try to understand the nature and sources of predictability, and conclude that our predictions mainly capture fixed risks of violence. Table 2 reports results. For purposes of comparison, Column (1) reproduces the results from Table 1 above.

5.2.1 Violence Histories Alone are a Good Predictor of Future Conflict

First we examine the predictive power of violence histories alone. Importantly, these histories are not simply lagged dependent variables. For each country we have data on the actors involved in different incidents of violence, as well as on the severity of those incidents (number of deaths, destruction of property, etc.). In Indonesia, we can distinguish between different motives. Our goal is to assess whether additional covariates improve performance over these rich conflict histories.

Column (2) and (3) consider models that uses all available information about past violence. Column (3) adds measures of population and population density to reflect the fact that more populous places mechanically have more people who can engage in conflict with one another.

The results are striking. The AUCs with violence histories alone perform almost as well as, and occasionally better than, our full model in Column (1). The addition of population in Column (3) improves the performance of violence histories very little.

In Column (4), we take this exercise to the extreme by dispensing with much of the rich violence data and predicting solely based on whether there was any conflict in the previous period. Performance is not terrible, as violence is indeed persistent, but it is markedly worse than when we use a full set of violence measures. Online Appendix A.2 (Appendix page 3) compares our full model to other autoregressive and OLS fixed effects models, while Online Appendix A.5 (Appendix page 7) further explores the returns to more detailed violence data.

5.2.2 Time-Invariant Predictors are Most Effective in Our Models

If detailed, past violence predicts future violence, do we need other predictors at all? In what follows, we develop a number of tests for parsing the sources of predictability.

Table 2 Column (5) shows a model that only uses predictors that do not directly measure past violence. These include the hundreds of socioeconomic and demographic measures discussed above. Performance is comparable to the full model in Column (1) and the violence-only model in Column (2). This suggests that these socioeconomic and demographic variables contain more or less the same information as the detailed histories of violence, but add little value over them.

Of course, our models contain hundreds of variables, and it is possible that some contribute much less than others. In particular, our models include a number of predictors that change slowly or not at all. Some, such as topographical traits or colonial history, do not vary by definition. Others, such as ethnic and religious traits in Indonesia, do not vary over our sample because they are measured only once. Variables that do not change over our sample cannot, by their nature, predict the timing of violent conflict.

To examine the relative performance of time-varying and time-invariant predictors, we compare models composed entirely of one or the other. Column (6) uses only time-invariant traits to predict violence, and performance roughly matches or outperforms the model in Column (5). Column (7) uses only time-varying predictors, and performance diminishes.[12] Thus, most of our model's performance can be achieved by successfully predicting time-invariant (or at least highly persistent) violence risk.

In Figure 1, we go one step further and examine the predictive performance of clusters of related predictors. We start with a baseline model that uses only population (level, growth rate, and density) to generate predictions. We then add subgroups of predictors to that baseline model and estimate the change in predictive performance. This approach estimates

[12]This is true even for the ≥ 1 SD increase. While such increases naturally have a temporal element to them, it appears that our performance comes from leveraging *which* places are most at risk of experiencing escalations as opposed to *when* these escalations occur.

Table 2: Out-of-Sample (One Year Ahead) Performance of the Ensemble (EBMA) Method, Varying Predictor Sets

	Area under the curve (AUC) for Models with						
	Full Predictors (1)	All Past Violence Measures (2)	All Past Violence & Population (3)	Lagged Predictand (AR(1)) (4)	Full Excl. Past Violence (5)	Time-Invariant Predictors (6)	Time-Varying Predictors (7)
	Indonesia (2008–2014)						
Any Incident (AUC)	0.814	0.801	0.809	0.687	0.798	0.805	0.753
≥ 5 Incidents (AUC)	0.935	0.929	0.934	0.808	0.912	0.920	0.874
≥ 1 S.D. Increase (AUC)	0.841	0.852	0.853	0.527	0.820	0.840	0.783
	Colombia (1998–2005)						
Any Incident (AUC)	0.851	0.812	0.837	0.743	0.827	0.830	0.768
≥ 5 Incidents (AUC)	0.916	0.905	0.914	0.748	0.878	0.875	0.829
≥ 1 S.D. Increase (AUC)	0.798	0.763	0.786	0.521	0.778	0.768	0.750

Notes: Each model is trained on all available data preceding the out-of-sample prediction year. Training data starts in 1991 in Colombia and 2002 in Indonesia. Out-of-sample prediction begins in 1998 in Colombia and 2008 in Indonesia. The AUC is the area under the ROC curve, a measure of the trade-off between the true positive rate and false positive rate at different thresholds. We report average performance over the out-of-sample years above. Past violence measures include breakdowns of events by actors and outcomes such as deaths and damages. Population includes population growth rates and density.

the predictive power of sets of predictors beyond their association with population.

Figure 1 plots the change in model performance from adding the variable subgroup. We do so for our three outcomes and both countries. Each out-of-sample year is indicated with a small dot, to give a sense of the range. The first out-of-sample year is marked with a green triangle and the last is marked with a blue square. While the first out-of-sample year generally has worse performance because the training sample is smallest, it is difficult to see a clear improvement in performance over successive out-of-sample predictions. The final year of data is not necessarily the best performing. In the discussion below, we focus on the average change—reflected by the larger open circle.

Consistent with the results above, time-invariant predictors appear to add the most to predictive performance. This is true even when looking at conflict escalation. The time-invariant predictors that add the most to predictive performance are also notably similar across countries and outcomes. Measures of remoteness, like distance from major cities, and

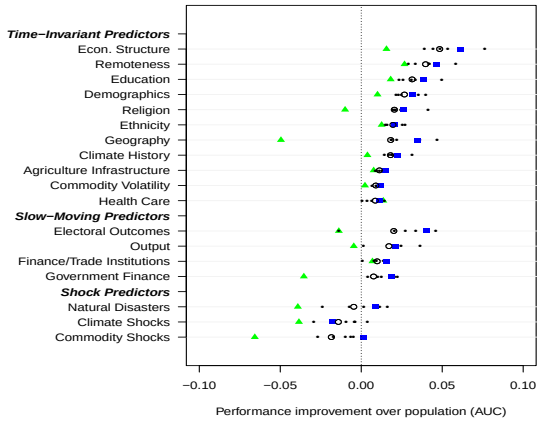
geographic traits like terrain ruggedness, are generally the best predictors. Measures of economic output and economic structure (such as sectoral shares of GDP or total employment) are also important predictors. (While these measures are not time-invariant in actuality, they are in our dataset).

In contrast, time-varying covariates seldom improve predictive power, even when predicting a ≥ 1 SD escalation. This is true of natural disasters, commodity prices and climate shocks. In some instances, adding these predictors decreases performance. However, the range of estimates is very large, which limits our ability to definitively conclude how these predictors affect model performance.[13] Elections appear to improve predictive power, though their performance added also varies. Note, moreover, that there is not a clear increase in the marginal contribution of these predictors over time. The first out-of-sample year is seldom the worst in terms of the marginal contribution of these time-varying predictors. Likewise, the final year (which uses the largest training set) is seldom the best.

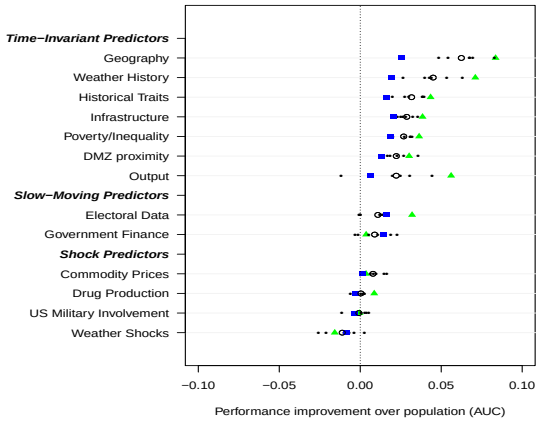
[13]The inconclusive performance of these variables in our year-ahead prediction models stands in contrast to causal studies of conflict, where variables such as commodity prices and weather fluctuations have been found to exert robust significant effects on conflict intensity and sometimes conflict onset. See, for example, Miguel et al. (2004), Bazzi and Blattman (2014), Berman and Couttenier (2013), Berman et al. (forthcoming), Burke et al. (2015) and Dube and Vargas (2013). This underscores the observation that the relationship between causation and prediction is complex (Shmueli, 2010), and the objective of prediction differs fundamentally from the objective of parameter estimation (Mullainathan and Spiess, 2017).

Figure 1: AUC Improvements from Individual Predictor Groups

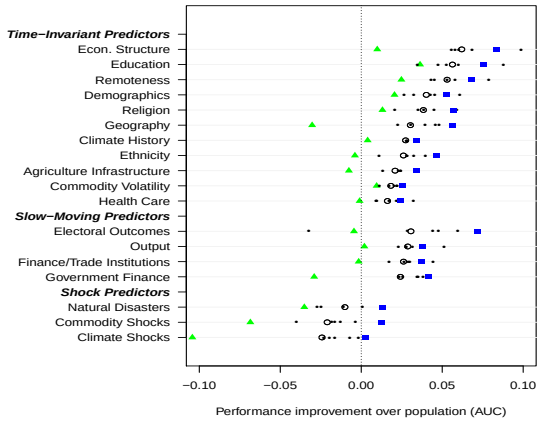
(a) Any violent event (Indonesia)



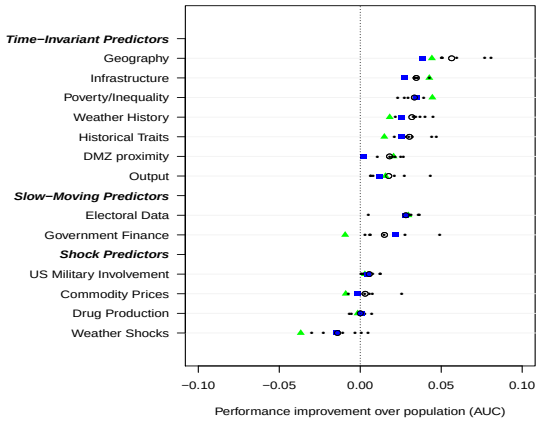
(b) Any violent event (Colombia)



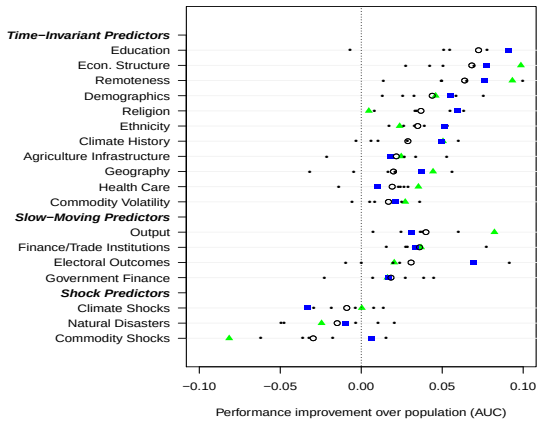
(c) ≥ 5 violent events (Indonesia)



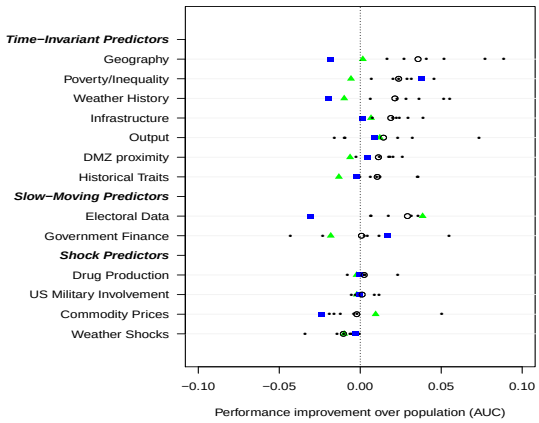
(d) ≥ 5 violent events (Colombia)



(e) ≥ 1 S.D. increase in violent events (Indonesia)



(f) ≥ 1 S.D. increase in violent events (Colombia)



Notes: Performance in individual years appear as small dots. The first (last) year of the sample is represented by a green triangle (blue square) in order to show the change in performance over time, or lack thereof. The large hollow circle is the average of performance across the years. A full breakdown of the variables in each of the different predictor groups can be found in Online Appendix C.3 (Appendix page 18).

5.2.3 Our Models Predict Violence across Locations

So far, several pieces of evidence point to the difficulty of predicting the specific timing of violence, including: the poor performance of time-varying predictors; the interchangeability of histories of violence and largely time-invariant predictors; and the fact that histories of violence predict spikes in violence roughly as well as levels of violence.

An additional exercise in the Online Appendix A.4 (Appendix page 6) shows that our models perform especially poorly if we seek to predict within-location, over-time variation in violence, using the deviation of violent incidents in each period from its historical mean. These results further underscore the difficulty of predicting within-unit changes in violence, given the available data.

Next, we take the opposite approach, forecasting conflict exclusively across locations. To do this, we randomly split subdistricts in Indonesia and municipalities in Colombia into two equal-sized groups. We pool observations over time, and train our algorithms using all location-years of data in one group of locations, generating predictions for a second group of location-years. Table 3 reports these results. Column (1) shows that overall performance when forecasting across locations is strikingly similar to performance when predicting ahead in time.

Table 3: Predicting Across Locations

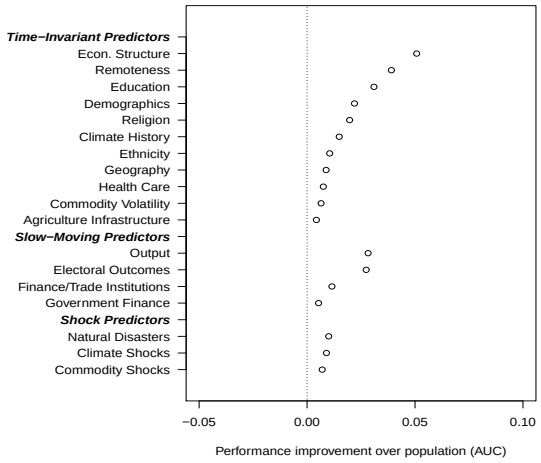
	Area Under the Curve (AUC) with			
	Full Predictors (1)	Only Past Violence Measures (2)	Only Past Violence and Population (3)	Full Excl. Past Violence (4)
	Indonesia (2008–2014)			
Any Incident (AUC)	0.829	0.813	0.820	0.812
$\geq$ 5 Incidents (AUC)	0.941	0.929	0.933	0.918
$\geq$ 1 S.D. Increase (AUC)	0.863	0.837	0.847	0.844
	Colombia (1998-2005)			
Any Incident (AUC)	0.838	0.792	0.817	0.807
$\geq$ 5 Incidents (AUC)	0.917	0.902	0.909	0.876
$\geq$ 1 S.D. Increase (AUC)	0.803	0.746	0.770	0.808

Notes: AUCs for a random test set of locations over time. Algorithms are trained using data from training locations over the entire time span of the datasets. Training data starts in 1991 in Colombia and 2002 in Indonesia. The AUC is the area under the ROC curve, a measure of the trade-off between the true positive rate and false positive rate as we vary the discrimination threshold.

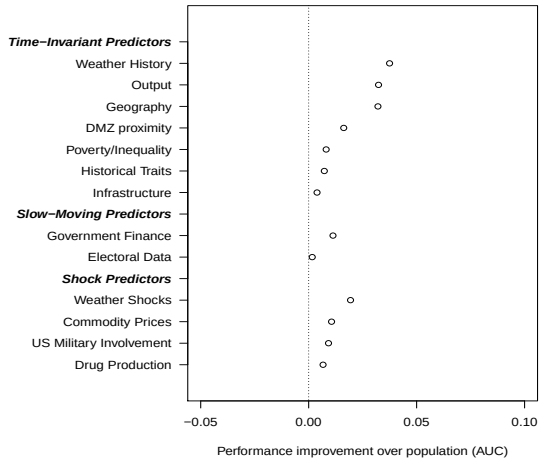
However, some differences emerge in the performance of individual predictor groups. Figure 2 examines which groups of predictors (along with population) best predict out-of-sample violence across locations. Time-invariant predictors remain important. However, time-varying predictors including weather, natural disasters and commodity prices no longer reduce predictive power, and in some cases, improve it substantially. One notable difference between the across-location and year-ahead predictions is that the training set uses all years of available data in the across-location approach. The algorithms therefore observe the entire relevant distribution of weather, disasters and commodity price fluctuations over the duration of the period. These variables may behave very differently year to year. When we predict violence one year ahead, if the training period includes such shocks, while the testing period does not, the lack of common support across these periods may inhibit the predictive power of these variables. Thus the short time series of the training and testing samples, and the difficulty of generating off-support predictions, may explain why time-varying covariates like weather shocks perform worse in our predictions over time.

Figure 2: AUC Improvements from Individual Predictor Groups, Cross-Sectional Prediction

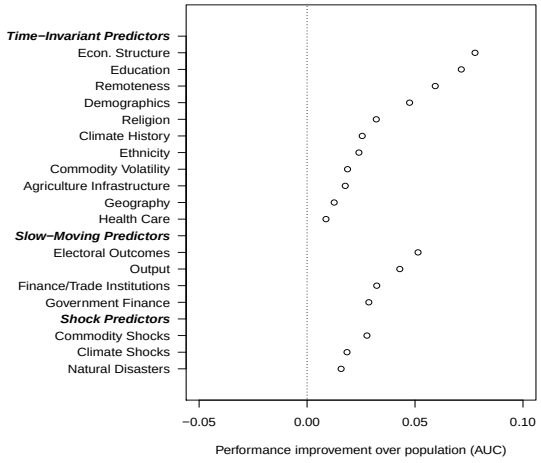
(a) Any violent event (Indonesia)



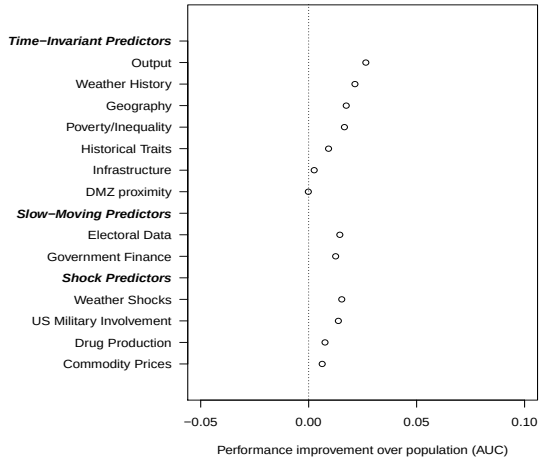
(b) Any violent event (Colombia)



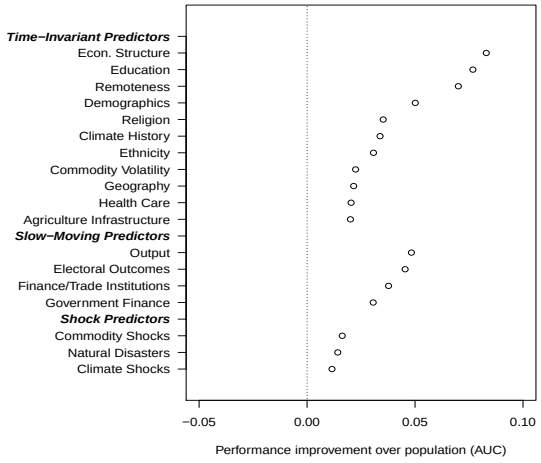
(c) ≥ 5 violent events (Indonesia)



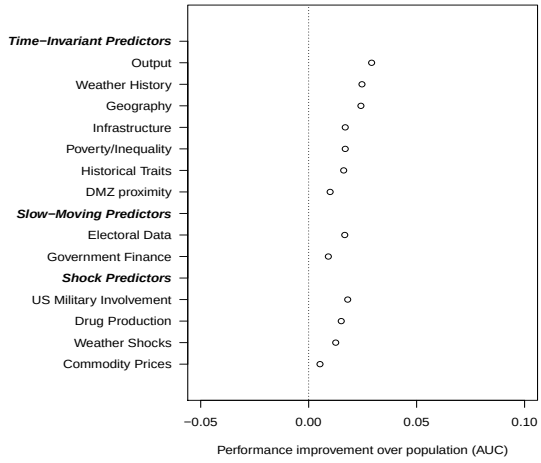
(d) ≥ 5 violent events (Colombia)



(e) ≥ 1 S.D. increase (Indonesia)



(f) ≥ 1 S.D. increase (Colombia)



Notes: Performance (AUC) in the single test sample is reported.

6 Discussion

A large portion of violence in Indonesia and Colombia appears to be predictable, but that predictability is largely a function of time-invariant, location-specific risk. This is important in and of itself, since hot spots for violence may pose an especially severe risk of further escalation. But the residual variation—year-to-year changes in violence—is difficult to forecast.

There are several possible explanations. For one, it is possible that the time-varying dimensions of violence are simply idiosyncratic and therefore hard to predict. In many cases, conflict is not only inefficient but is an out-of-equilibrium behavior (Fearon, 1995). These deviations from normal, peaceful social competition could be inherently difficult to forecast. Violence may also be hard to predict because it responds endogenously to the strategic calculations of armed actors. For instance, we may observe peace in a particular region precisely because government security forces crudely predicted a high conflict risk there, and allocated resources accordingly. Likewise, a terrorist may decide to attack an area because that is where the attack was least expected.

Measurement problems may also limit model performance. The timing of violence could be a function of factors that are inherently difficult to observe and measure accurately, such as social grievances or the deterioration of communal trust. These variables might improve our models, if only we could measure them. Despite our large set of predictors and our massively interactive models, measurement challenges may reduce predictive performance.

We may also lack a sufficiently long time series to be able to capture time-varying conflict risk. The limited predictive power of shocks in our over-time predictions may reflect a lack of common support in the relatively short training and testing samples. If so, then performance might improve with more years of data. At the same time, our results hold even when the training sample is at its longest, and as the training sample gets longer one might be more concerned about the possibility of structural breaks in the violence generating process. High-frequency data on local conditions and leading indicators of violence are other potential avenues for improvement. For now, few developing country contexts offer such data. But

possibilities in the near future include data from social media, mobile phone meta-data, real-time incident data, and media monitoring. We view these as promising avenues for future research seeking to forecast where violence changes over time.

References

Acemoglu, Daron, Carlos Garcia-Jimeno, and James A. Robinson, "State capacity and economic development: A network approach," *American Economic Review*, 2015.

_ , **James A. Robinson, and Rafael J. Santos**, "The Monopoly of Violence: Evidence from Colombia," *Journal of the European Economic Association*, 2013, *11* (S1), 5–44. https://economics.mit.edu/files/10402.

_ , **Leopoldo Fergusson, James A. Robinson, Dario Romero, and Juan F. Vargas**, "The Perils of High-Powered Incentives: Evidence from Colombia's False Positives," Working Paper 22617, National Bureau of Economic Research September 2016. http://dx.doi.org/10.3386/w22617.

Alesina, A. and E. Zhuravskaya, "Segregation and the Quality of Government in a Cross Section of Countries," *American Economic Review*, 2011, *101* (5), 1872–1911.

Angrist, Joshua D. and Adriana D. Kugler, "Rural Windfall or a New Resource Curse? Coca, Income, and Civil Conflict in Colombia," *The Review of Economics and Statistics*, May 2008, *90* (2), 191–215. https://ideas.repec.org/a/tpr/restat/v90y2008i2p191-215.html.

Barron, Patrick, Kai Kaiser, and Menno Pradhan, "Understanding variations in local conflict: Evidence and implications from Indonesia," *World Development*, 2009, *37* (3), 698–713. https://doi.org/10.1016/j.worlddev.2008.08.007.

_ , **Sana Jaffrey, and Ashutosh Varshney**, "How Large Conflicts Subside: Evidence From Indonesia," *Indonesia Social Development Paper*, 2014, (18). https://asiafoundation.org/resources/pdfs/HowLargeConflictsSubside.pdf.

_ , _ , **and** _ , "When Large Conflicts Subside: The Ebbs and Flows of Violence in Post-Suharto Indonesia," *Journal of East Asian Studies*, 2016, *16* (2), 191–217. https://doi.org/10.1017/jea.2016.6.

Bazzi, Samuel and Christopher Blattman, "Economic shocks and conflict: Evidence from commodity prices," *American Economic Journal: Macroeconomics*, 2014, *6* (4), 1–38. dx.doi.org/10.1257/mac.6.4.1.

_ **and Matthew Gudgeon**, "The Political Boundaries of Ethnic Divisions," *Unpublished Manuscript*, 2017. https://ssrn.com/abstract=3098128.

Beck, Nathaniel, Gary King, and Langche Zeng, "Improving Quantitative Studies of International Conflict: A Conjecture," *The American Political Science Review*, March 2000, *94* (1), 21–35. http://dx.doi.org/10.2307/2586378.

Beger, Andreas, Cassy L. Dorff, and Michael D. Ward, "Irregular leadership changes in 2014: Forecasts using ensemble, split-population duration models," *International Journal of Forecasting*, January 2016, *32* (1), 98–111. dx.doi.org/10.1016/j.ijforecast.2015.01.009.

Berger, Daniel, Shankar Kalyanaraman, and Sera Linardi, "Violence and Cell Phone Communication: Behavior and Prediction in Cote D'Ivoire," *Unpublished Manuscript*, 2014. https://ssrn.com/abstract=2526336.

Berman, Nicolas and Mathieu Couttenier, "External shocks, internal shots: the geography of civil conflicts," *Review of Economics and Statistics*, 2013, (0). https://doi.org/10.1162/REST_a_00521.

_ , _ , **Dominic Rohner, and Mathias Thoenig**, "This Mine is Mine! How Minerals Fuel Conflicts in Africa," *American Economic Review*, forthcoming. dx.doi.org/10.1257/aer.20150774.

Blair, Robert A., Christopher Blattman, and Alexandra Hartman, "Predicting local violence: Evidence from a panel survey in Liberia," *Journal of Peace Research*, March 2017, *54* (2), 298–312. http://dx.doi.org/10.1177/0022343316684009.

Blattman, C. and E. Miguel, "Civil war," *Journal of Economic Literature*, 2010, pp. 3–57. dx.doi.org/10.1257/jel.48.1.3.

Brandt, Patrick T., John R. Freeman, and Philip A. Schrodt, "Real Time, Time Series Forecasting of Inter- and Intra-State Political Conflict," *Conflict Management and Peace Science*, 2011, *28* (1), 41–64. https://doi.org/10.1177/0738894210388125.

Burke, Marshall, Solomon Hsiang, and Edward Miguel, "Climate and Conflict," *Annual Review Economics*, 2015, *7*, 577–617. https://doi.org/10.1146/annurev-economics-080614-115430.

Cederman, Lars-Erik and Nils B. Weidmann, "Predicting armed conflict: Time to adjust our expectations?," *Science*, 2017, *355* (6324), 474–476. http://dx.doi.org/10.1126/science.aal4483.

Celiku, Bledi and Aart Kraay, "Predicting conflict," *World Bank Policy Research Working Paper 8075*, 2017. https://openknowledge.worldbank.org/handle/10986/26847.

Chacon, Mario, "In the Line of Fire: Political Violence and Decentralization in Colombia," *Working Paper*, 2014. https://dx.doi.org/10.2139/ssrn.23866677.

Colaresi, Michael, Håvard Hegre, and Jonas Nordkvelle, "Early ViEWS: A prototype disaggregated, open-source Violence Early-Warning System," *Presented to the American Political Science Association annual convention, Philadelphia*, 2016. http://www.pcr.uu.se/digitalAssets/653/c_653796-l_1-k_earlyviewsapsa2016.pdf.

Dube, Oeindrila and Juan F. Vargas, "Commodity price shocks and civil conflict: Evidence from Colombia," *The Review of Economic Studies*, 2013, *80* (4), 1384–1421. https://doi.org/10.1093/restud/rdt009.

_ **and Suresh Naidu**, "Bases, Bullets, and Ballots: The Effect of US Military Aid on Political Conflict in Colombia," *The Journal of Politics*, 2015, *77* (1), 249–267. https://www.journals.uchicago.edu/doi/10.1086/679021.

Fearon, James D., "Rationalist explanations for war," *International organization*, 1995, *49* (3), 379–414. https://www.jstor.org/stable/2706903.

Fischer, Gunther, Freddy Nachtergaele, Sylvia Prieler, Harrij van Velthuizen, Luc Verelst, and David Wiberg, "Global Agro-ecological Zones Assessment for Agriculture (GAEZ 2008)," Technical Report, GAEZ 2008. http://www.fao.org/nr/gaez/publications/en/.

Freund, Yoav and Robert E. Schapire, "A Short Introduction to Boosting," in "In Pro-

ceedings of the Sixteenth International Joint Conference on Artificial Intelligence" Morgan Kaufmann 1999, pp. 1401–1406. https://dl.acm.org/citation.cfm?id=1624417.

Gartzke, Erik, "War is in the Error Term," *International Organization*, 1999, *53* (3), 567–587. https://doi.org/10.1162/002081899550995.

Gleditsch, Kristian Skrede and Michael D. Ward, "Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes," *Journal of Peace Research*, 2013, *50* (1), 17–31. https://doi.org/10.1177/0022343312449033.

Goldstone, Jack A., Robert H. Bates, David L. Epstein, Ted Robert Gurr, Michael B. Lustik, Monty G Marshall, Jay Ulfelder, and Mark Woodward, "A Global Model for Forecasting Political Instability," *American Journal of Political Science*, January 2010, *54* (1), 190–208. https://doi.org/10.1111/j.1540-5907.2009.00426.x.

Gurr, Ted Robert and Mark Lichbach, "Forecasting Internal Conflict," *Comparative Political Studies*, 1986, *19* (1), 3–38. https://doi.org/10.1177/0010414086019001001.

Harff, Barbara, "No Lessons Learned from the Holocaust? Assessing Risks of Genocide and Political Mass Murder Since 1955," *American Political Science Review*, 2003, *97* (1), 57–73. https://www.jstor.org/stable/3118221.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning* Springer Series in Statistics, New York, NY, USA: Springer New York Inc., 2001.

Hegre, Håvard, Joakim Karlsen, Håvard Mokleiv Nygård, Håvard Strand, and Henrik Urdal, "Predicting Armed Conflict, 2010–2050," *International Studies Quarterly*, 2013, *57* (2), 250–270. https://doi.org/10.1111/isqu.12007.

Hegre, Håvard, Halvard Buhaug, Katherine V Calvin, Jonas Nordkvelle, Stephanie T Waldhoff, and Elisabeth Gilmore, "Forecasting civil conflict along the shared socioeconomic pathways," *Environmental Research Letters*, 2016, *11* (5), 054002. dx.doi.org/10.1088/1748-9326/11/5/054002.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil, "Measuring Economic Growth from Outer Space," *American Economic Review*, 2012, *102* (2), 994–1028. dx.doi.org/10.1257/aer.102.2.994.

Historical Memory Group, *"Enough Already!" Colombia: Memories of War and Dignity*, The National Center for Historical Memory, 2013. http://www.centrodememoriahistorica.gov.co/micrositios/informeGeneral/descargas.html.

Jasny, Barbara R. and Richard Stone, "Prediction and its Limits," *Science*, 2017, *355* (6324), 468–469. dx.doi.org/10.1126/science.355.6324.468.

Miguel, E., S. Satyanath, and E. Sergenti, "Economic shocks and civil conflict: An instrumental variables approach," *Journal of Political Economy*, 2004, *112* (4), 725–753. https://www.journals.uchicago.edu/doi/10.1086/421174.

Montgomery, Jacob, Florian Hollenbach, and Michael Ward, "Improving Predictions using Ensemble Bayesian Model Averaging," *Political Analysis*, 2012, *20*, 271–291. https://doi.org/10.1093/pan/mps002.

Mueller, Hannes and Christopher Rauh, "Reading Between the Lines: Prediction

of Political Violence Using Newspaper Text," *American Political Science Review*, 2017, pp. 1–18. dx.doi.org/10.1017/S0003055417000570.

Mullainathan, Sendhil and Jann Spiess, "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, May 2017, *31* (2), 87–106. http://dx.doi.org/10.1257/jep.31.2.87.

Perry, C., "Machine learning and conflict prediction: a use case," *Stability: International Journal of Security and Development*, 2013, *2* (3), 56. http://doi.org/10.5334/sta.cr.

Pierskalla, Jan H. and Audrey Sacks, "Unpacking the Effect of Decentralized Governance on Routine Violence: Lessons from Indonesia," *World Development*, 2017, *90*, 213–228. https://doi.org/10.1016/j.worlddev.2016.09.008.

Restrepo, Jorge, Michael Spagat, and Juan Vargas, "The Dynamics of the Columbian Civil Conflict: A New Dataset," *Homo Oeconomicus*, 2004, *21*, 396–429. https://ssrn.com/abstract=480247.

Sappington, J. Mark, Kathleen M. Longshore, and Daniel B. Thompson, "Quantifying Landscape Ruggedness for Animal Habitat Analysis: A Case Study Using Bighorn Sheep in the Mojave Desert," *The Journal of Wildlife Management*, 2007, *71* (5), 1419–1426. http://dx.doi.org/10.2193/2005-723.

Schrodt, Philip A., "Forecasting Conflict in the Balkans using Hidden Markov Models," in Robert Trappl, ed., *Programming for Peace: Advances in Group Decision and Negotiation, Vol. 2*, Springer, 2006, pp. 161–184. dx.doi.org/10.1214/10-STS330.

Shmueli, Galit, "To explain or to predict?," *Statist. Sci.*, 08 2010, *25* (3), 289–310. dx.doi.org/10.1214/10-STS330.

Tadjoeddin, Zulfan, *Explaining collective violence in contemporary Indonesia: from conflict to cooperation*, Springer, 2014.

Tibshirani, Robert, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 1994, *58*, 267–288. https://www.jstor.org/stable/2346178.

Ward, M., N. Metternich, C. Dorff, M. Gallop, F. Hollenbach, A. Schultz, and S. Weschle, "Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction," *International Studies Review*, 2013, *15* (4), 473–490. https://doi.org/10.1111/misr.12072.

Weidmann, Nils and Michael Ward, "Predicting Conflict in Space and Time," *Journal of Conflict Resolution*, 2010, *54* (6), 883–901. https://doi.org/10.1177/0022002710371669.

Witmer, Frank D.W., Andrew M Linke, John O'Loughlin, Andrew Gettelman, and Arlene Laing, "Subnational violent conflict forecasts for sub-Saharan Africa, 2015–65, using climate-sensitive models," *Journal of Peace Research*, 2017, *54* (2), 175–192. http://dx.doi.org/10.1177/0022343316682064.

Wright, Austin L. and Patrick Signoret, "Climate Shocks, Price Dynamics, and Human Conflict," *Working paper*, 2016. https://www.austinlwright.com/climate-shocks/.

Appendix for Online Publication

Table of Contents

A Additional Results

A.1 Other Measures of Model Performance

In the main text, we measure performance with the AUC. However, practitioners may be interested in other performance statistics. Here we consider choosing various discrimination thresholds, each of which essentially picks a point on the ROC curve. In Table A.1 we report a number of alternative performance statistics. Each of our three main predictands are reported in a separate panel. Note that we do not change the performance metric that we use to choose hyper-parameters. We merely report alternative performance metrics for the same models that are developed in the main results to maximize AUC.

Maximal Accuracy. In the top portion of each panel, we report performance when we set a discrimination threshold to maximize accuracy. We choose this discrimination threshold in our cross validation routine. We report accuracy (the proportion of all cases correctly predicted), sensitivity (the proportion of incidents correctly predicted), and specificity (the proportion of non-incidents correctly predicted).[1]

Maximal Sensitivity. In the next section of each panel, we choose a different discrimination threshold to maximize sensitivity. Of course, if we predicted violence everywhere, we would achieve a sensitivity of 1, but that would not be a useful prediction. Instead, we choose to maximize sensitivity subject to the constraint that we keep accuracy above 0.5 in the cross-validation process. We observe, that if needed, these models could identify all of the places that experience violence, but this coverage comes at a high cost in terms of accuracy and specificity (false positives).[2]

Mean Squared Error. We report the mean squared error. Generally, the MSE is closely correlated with the AUC, but the correspondence is not exact.[3] Using the MSE does not meaningfully change the results of this paper.

Area Under the Precision-Recall Curve. Finally, we report the area under the Precision-Recall Curve. This metric measures the trade-off between precision (share of true positives in the sample of positive predictions) and recall (the true positive rate). In contexts such as ours in which data is imbalanced and violence is relatively rare, high PR-AUC's can be difficult to achieve. If there are many non-events, then mis-classifying a small share as positive can have large, deleterious effects on precision. Indeed, we see for the rarest events (the > 1 standard deviation spikes) the PR-AUC is quite low. However, we also see that holding the predictand fixed, comparisons of model performance based on the area under the Precision-Recall curve are largely consistent with the comparisons based on the area under the ROC curve.

[1] Observe that as the outcome becomes rarer, accuracy generally increases while sensitivity plummets. This is a mechanical phenomenon. When an outcome is rare, it is easy to predict that it never happens (an uninformative prediction) and achieve high accuracy.

[2] A practitioner might have preferences between these two extremes and consequently choose a point on the ROC that differs from these two. This is precisely why we use the AUC as our benchmark.

[3] This is because the AUC penalizes errors in relative rankings of different locations, while the MSE penalizes errors of prediction according to the magnitude of the difference between actual outcomes and predicted probabilities. While these two types of errors are related, they are distinct. However, observe that when the two yield differing comparisons of two models, the differences are small. Assessing model performance according to MSE only changes comparisons for borderline cases, which we do not put much emphasis on.

Table A.1: Out-of-Sample Performance of Prediction Models

	Indonesia (social conflicts)					**Colombia** (attacks and clashes)				
	LASSO (1)	Random Forest (2)	Adaptive Boosting (3)	Neural Network (4)	EBMA (5)	LASSO (6)	Random Forest (7)	Adaptive Boosting (8)	Neural Network (9)	EBMA (10)
	(a) Indicator of any violent event									
EBMA Weight	0.265	0.261	0.212	0.262		0.258	0.254	0.235	0.253	
	Threshold maximizes accuracy									
Accuracy	0.725	0.728	0.728	0.705	0.731	0.773	0.776	0.778	0.764	0.779
Sensitivity	0.749	0.697	0.729	0.676	0.75	0.609	0.633	0.609	0.601	0.645
Specificity	0.699	0.767	0.732	0.741	0.713	0.872	0.864	0.883	0.863	0.86
	Threshold maximizes sensitivity, while accuracy is above 50%									
Accuracy	0.546	0.541	0.546	0.54	0.574	0.526	0.522	0.515	0.439	0.56
Sensitivity	0.999	0.999	0.999	1	0.994	0.976	0.976	0.98	0.983	0.966
Specificity	0.017	0.006	0.019	0.003	0.084	0.243	0.241	0.226	0.096	0.302
MSE (Brier Score)	0.18	0.178	0.178	0.195	0.177	0.155	0.154	0.152	0.167	0.152
P-R AUC	0.838	0.845	0.849	0.807	0.848	0.794	0.794	0.792	0.767	0.8
AUC	**0.806**	**0.810**	**0.814**	**0.780**	**0.814**	**0.845**	**0.846**	**0.848**	**0.826**	**0.851**
Dep Var. Mean	0.531					0.361				
	(b) Indicator of ≥ 5 violent events									
EBMA Weight	0.262	0.278	0.179	0.281		0.252	0.251	0.246	0.251	
	Threshold maximizes accuracy									
Accuracy	0.927	0.928	0.926	0.915	0.928	0.922	0.922	0.922	0.916	0.923
Sensitivity	0.599	0.574	0.572	0.449	0.614	0.399	0.369	0.405	0.251	0.443
Specificity	0.974	0.978	0.976	0.981	0.973	0.98	0.983	0.979	0.989	0.976
	Threshold maximizes sensitivity, while accuracy is above 50%									
Accuracy	0.399	0.355	0.371	0.49	0.54	0.481	0.498	0.495	0.214	0.566
Sensitivity	0.985	0.992	0.991	0.963	0.978	0.974	0.983	0.975	0.977	0.967
Specificity	0.316	0.264	0.283	0.422	0.478	0.424	0.445	0.442	0.129	0.52
MSE (Brier Score)	0.058	0.057	0.057	0.065	0.055	0.058	0.058	0.059	0.067	0.058
P-R AUC	0.707	0.754	0.752	0.675	0.763	0.624	0.601	0.602	0.54	0.619
AUC	**0.922**	**0.927**	**0.932**	**0.91**	**0.935**	**0.914**	**0.909**	**0.910**	**0.88**	**0.916**
Dep. Var. Mean	0.127					0.083				
	(c) ≥ 1 S.D. increase in violent events									
EBMA Weight	0.276	0.299	0.125	0.300		0.250	0.250	0.249	0.251	
	Threshold maximizes accuracy									
Accuracy	0.965	0.965	0.961	0.965	0.964	0.948	0.947	0.948	0.948	0.946
Sensitivity	0.008	0.02	0.041	0.007	0.047	0.012	0.01	0.023	0	0.053
Specificity	0.999	0.999	0.994	1	0.997	0.999	0.998	0.998	1	0.994
	Threshold maximizes sensitivity, while accuracy is above 50%									
Accuracy	0.392	0.236	0.332	0.352	0.344	0.455	0.406	0.481	0.233	0.495
Sensitivity	0.965	0.975	0.959	0.95	0.961	0.925	0.949	0.91	0.916	0.914
Specificity	0.372	0.21	0.31	0.33	0.322	0.428	0.375	0.457	0.193	0.471
MSE (Brier Score)	0.03	0.034	0.046	0.033	0.032	0.046	0.047	0.049	0.05	0.046
P-R AUC	0.207	0.17	0.185	0.157	0.209	0.197	0.168	0.184	0.157	0.19
AUC	**0.851**	**0.797**	**0.82**	**0.803**	**0.841**	**0.803**	**0.787**	**0.792**	**0.747**	**0.798**
Dep. Var. Mean	0.033					0.045				

Notes: Each model is trained on all data available preceding the out-of-sample prediction year. Accuracy is the proportion of subdistricts correctly predicted. Sensitivity is the proportion of subdistricts that actually experience violence for which we predicted violence. Specificity is the proportion of subdistricts that do not actually experience violence, where we accurately predict non-violence. The AUC is the area under the ROC curve, a measure of the trade-off between the true positive rate and false positive rate as we vary the discrimination threshold.

A.2 Alternative Benchmarks: Autoregression and Fixed Effects

Is this simple recurrence? Table A.2 reports the performance for the full model (Column 1) and the model that includes all predictors aside from past violence (Column 2). We compare these models to simpler ones, such as auto-regressive and fixed effect models that rely purely on the persistence of violence.

Column 3 reports the simplest recurrence model, restricting the predictors to two lags of the dependent variable. When predicting more than one or five events, the AUCs generally exceed 0.74, but performance is nevertheless considerably worse than for the full model. Standard deviation changes in violence are not particularly auto-correlated, however, and, hence, for this outcome the simple recursive model performs considerably worse than our full models. The AUC near 0.5 for the standard deviation increase for Indonesia indicates that this recursive model is about as good as chance.

Is violence prediction a function of fixed effects? Table A.2 also considers whether our predictions, which use mostly time-invariant covariates, simply approximate fixed effects. Column 4 reports a benchmark case of ordinary least squares (OLS) regression with fixed effects. We see that in all cases, prediction using our full set of covariates moderately outperforms the fixed effects model.[4]

The relative outperformance of all covariates is greatest in Indonesia and in the cases in which the dependent variable is rarer. This is intuitive from the perspective of estimator variance. These fixed effects are noisily estimated, and our prediction algorithms are better able to estimate the relationship between measured fixed factors and conflict.[5]

Column 5 attempts to remedy this imprecision by estimating fewer fixed effects at a higher level of aggregation—the department in Colombia and the district in Indonesia.[6] In this case, performance is close to that of the model with past violence in Indonesia and is slightly superior to that model in Colombia.

Finally, Column 6 reports a hybrid model, closely related to those in Columns 3 and 5. Here, we include two lags of the dependent variable, location fixed effects, and log population. In general, performance improves relative to Columns 3 and 4, but these predictions still fall short of those of the main models in Columns 1 and 2.

In general, these results show that the full models perform somewhat better than several models which fully rely on the persistence of violence. For the most part, the model that employs all variables except past violence also outperforms these benchmark cases, or at least does just as well. We infer that these predictions do improve upon a simple a recurrence model. However, the magnitudes are not large in all cases.[7] Given the large effort to collect data and employ sophisticated algorithms, which use days of computing power, these may be relatively modest returns.

[4] Recall that a .05 AUC gain is akin to a 10% improvement in model performance.

[5] The fixed effects model is required to estimate 1,023 parameters in Colombia and 2,009 parameters in Indonesia. As the variation in the dependent variable decreases, which happens as it becomes rarer, this estimation becomes increasingly difficult. In all but the case of predicting any incident in Colombia, the performance of the fixed effect model falls short of the model that uses all predictors except for past violence (Column 2).

[6] In Colombia, there are 33 departments, and there are 168 districts in Indonesia. Therefore, the number of parameters drops considerably, as does the imprecision in their estimation.

[7] The improvements range from 0.02–0.05, which is roughly 4–10% of the difference between a random prediction and perfection; hence, it is not negligible.

Table A.2: Out-of-Sample (One Year Ahead) Performance Versus Benchmarks

	Area under the curve (AUC) for Models with					
	Full Predictors	No Lagged Violence	Only Lagged Indicator AR(2)	OLS Mun./Subdist. FE	OLS Dept./Dist. FE	OLS Dept./Dist. FE + population + AR(2)
	(1)	(2)	(3)	(4)	(5)	(6)
	Indonesia (2008–2014)					
Any Incident (AUC)	0.814	0.798	0.745	0.774	0.752	0.803
$\geq$ 5 Incidents (AUC)	0.935	0.912	0.854	0.880	0.871	0.914
$\geq$ 1 S.D. Increase (AUC)	0.841	0.820	0.527	0.694	0.776	0.821
	Colombia (1998–2005)					
Any Incident (AUC)	0.851	0.827	0.795	0.828	0.721	0.829
$\geq$ 5 Incidents (AUC)	0.916	0.878	0.798	0.849	0.742	0.870
$\geq$ 1 S.D. Increase (AUC)	0.798	0.778	0.591	0.683	0.712	0.756

Notes: Each model is trained on all data available preceding the out-of-sample prediction year. Training data starts in 1991 in Colombia and 2002 in Indonesia. Out-of-sample prediction begins in 1998 in Colombia and 2008 in Indonesia. The AUC is the area under the ROC curve, a measure of the trade-off between the true positive rate and false positive rate as we vary the discrimination threshold. We report average performance over the out-of-sample years above. Past violence measures include breakdowns of events by actors and outcomes such as deaths and damages whereas the lagged indicator refers to only the variable that is being predicted. Population includes population growth rates and density.

A.3 Predicting Event Counts

We chose to predict indicators instead of counts in our main analyses for several reasons: they are easy to interpret, performance statistics are also easy to interpret, and they are the most common type of outcome predicted by the conflict literature to date. In this section, however, we examine the predictability of event counts.[8]

Table A.3 reports performance of the ensemble for a number of predictor sets similar to those that we consider in the paper. We report two statistics–the mean squared error and a deviance-based R^2. The deviance-based R^2 is similar to a typical R^2 but is adapted to count data.[9]

Column (1) reports overall performance of the full model. Our predictions explain a decent share of the variation in counts in both countries. Column (2) shows that, as in our prediction of indicators, violence histories perform just as well as the full set of predictors. Column (3) adds population measures to Column (2) and shows a small increase in performance.

Column (4) considers all of the predictors that do not directly measure violence. These predictors perform significantly worse than the full model. This drop in performance is worse than when we are predicting indicators. Columns (5) and (6) further clarify the reason for this. Of the

[8]Predicting event counts requires some changes to the algorithms. In the cases of Lasso and Gradient Boosted Machines, we change the loss function to a Poisson loss function to accommodate the count data. In theory, we could do the same for Random Forests and Neural Networks, but it is technically more challenging since the existing R packages do not accommodate count data. Instead, we use a Gaussian loss function. This presents a problem for the Neural Network algorithm, as its output is linear and therefore cannot take values below zero. While this occurs for some location-years, it is not terribly common and we deal with this by left censoring the predictions at zero. For Random Forests, this is not an issue because the algorithm can only make predictions in the convex hull of the predictand values in the training set. We aggregate these predictions using an ensemble model, in this case with weights based on model likelihood according to a Poisson distribution.

[9]While the typical R^2 measures the share of variance that for which the fitted values account, R^2_{dev} measures the share of deviance that the fitted values explain. Specifically, this measure is $R^2_{dev} = 1 - \frac{\text{Residual Deviance}}{\text{Null Deviance}}$

non-violence predictors, the time-varying predictors perform more poorly–so much so that their inclusion actually reduces performance relative to time-invariant predictors alone.

This result is in line with what we see in the main text. In the case of predicting indicators, we saw that some of the time-varying predictors actually decreased performance. In this case we see much larger decreases, but that is because the outcome is unbounded and so is the size of possible errors. Therefore, when extrapolating the effect of commodity price shocks, for example, imprecise estimation can lead to large differences between actual and predicted violence in a few cases. And these large errors play an outsized role in contributing to our performance statistics.

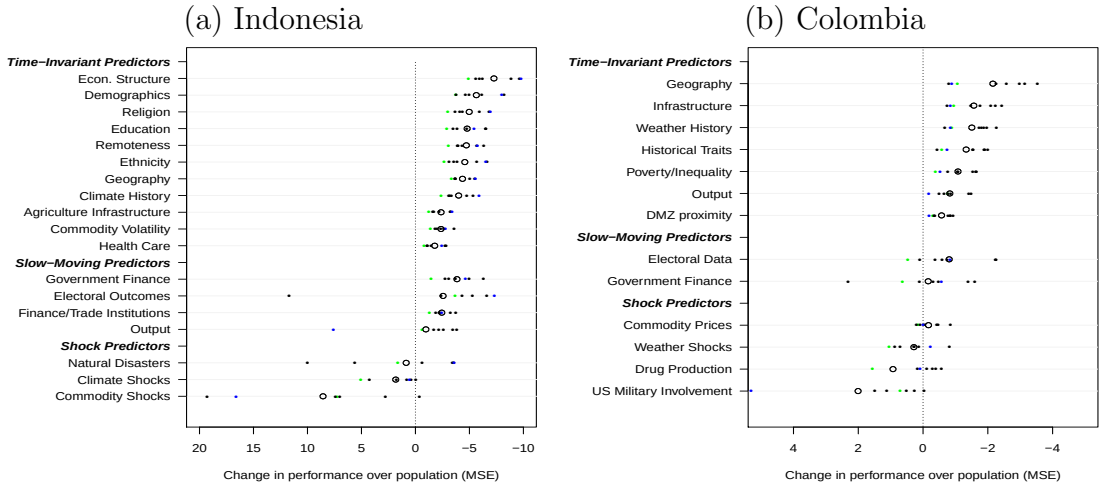
Finally, figure A.1 reports a breakdown of specific predictor groups and shows similar patterns to those we see for the prediction of indicators.

Table A.3: Out-of-sample (one year ahead) performance of the ensemble (EBMA) method, varying predictor sets

	Predicting the count of violent events					
	Full Set (1)	Only Past Violence (2)	Past Viol. plus Pop. (3)	No Lagged Violence (4)	Fixed Predictors (5)	Varying Predictors (6)
	Indonesia (2008-2014)					
MSE	7.33	7.59	7.58	12.41	9.41	19.97
R^2_{dev}	0.65	0.66	0.66	0.48	0.63	0.26
	Colombia (1998-2005)					
MSE	6.36	6.41	6.29	8.17	7.95	9.69
R^2_{dev}	0.53	0.52	0.54	0.44	0.45	0.29

Notes: Each model is trained on all data available preceding the out-of-sample prediction year. Training data starts in 1991 in Colombia and 2002 in Indonesia. Out-of-sample prediction begins in 1998 in Colombia and 2008 in Indonesia. We report average performance over the out-of-sample years above.

Figure A.1: MSE Improvements, Predicting Counts



Notes: Performance in individual years appear as small dots. The first (last) year of the sample is colored green (blue) in order to show the change in performance over time, or lack thereof. The large hollow circle is the average of performance across the years.

A.4 Predicting Within-Location Risk

We construct a new outcome measure that isolates within-location, over-time variation in violence. Specifically, we measure the deviation of the number of incidents in period t from the average number of incidents per year from the start of the panel to period $t-1$. By taking deviations from the historical mean, we remove the average difference in violence across locations, leaving purely within-location variation. Given the continuous outcome, we evaluate performance using the out-of-sample mean squared error (MSE) instead of the AUC.

The results, reported in Table A.4, suggest that our models struggle to predict within-location variation in violence in both Indonesia and Colombia.[10] The first row reports the variance of the dependent variable in each context. In Indonesia, the MSE is only slightly lower than the variance of the dependent variable in the test set, indicating we are able to predict very little of this within-location variation. In Colombia, the MSE is actually higher, on average, than the variance of the predictand, yielding an out-of-sample R^2 below zero.

[10]These differences arise from the difficulty of predicting within-unit deviations and not changes in the performance metric or the shift to counts. Online Appendix A.1 shows that our baseline performance is similar using MSE instead of AUC, while Online Appendix A.3 shows that our benchmark model is able to predict a large share of the variation in incident counts.

Table A.4: Predicting Demeaned Number of Violent Events

	Indonesia	Colombia
	(1)	(2)
Var(Dependent Variable)	8.622	7.541
EBMA mean squared error (MSE)	7.975	7.772
out-of-sample-R^2	**0.0129**	**-0.0276**

Notes: Each model is trained on all available data preceding the out-of-sample prediction year. Training data starts in 1991 in Colombia and 2002 in Indonesia. Out-of-sample prediction begins in 1998 in Colombia and 2008 in Indonesia. We report average mean squared error over the out-of-sample years above.

A.5 Returns to More Detailed Violence Measurement

Our benchmarking exercise found that a model using detailed violence histories performs almost as well as one that includes additional social, economic, and political covariates. This leads us to ask what is the payoff to using richer and more detailed violence data. More granular, accurate, or disaggregated data could improve predictions. For example, a history of small-scale ethnic cleansing could conceivably presage larger scale violence, whereas other kinds of inter-group hostilities might not. If a violence measures conflates these two kinds of violence then its predictive performance will falter. Yet, collecting and coding richer data is costly for policymakers and researchers. Hence, it is useful to explore the returns in terms of predictive performance.

With Indonesia, which has fairly granular violence data, we can conduct this 'experiment' by using more versus less granular violence data, and observe changes in performance. We report results in Table A.5. Column (1) reports the full model with all predictors for comparison. Column (2) reports performance for the lagged dependent variables alone. Column (3) reports performance using measurements of prior aggregate conflict from SNPK, including total number of incidents, total deaths, total injuries, and total property damage. Column (4) further breaks down these incident measures by violence category (e.g., identity violence, resource conflict). And finally, Column (5) includes the lagged total number of killings and indicators for mass unrest reported in *Podes* and the Disaster Information Management System.

The AUCs increase with each successive column. Perhaps unsurprisingly, the largest increase comes from the move from columns 2 to 3. The AUCs increase more when disaggregating the subdistrict-level violence categories than when adding additional measures of the same broad episodes of violence as reported in *Podes*. This highlights the potential predictive value of having detailed information on the nature of prior conflict in terms of the key outcomes of contestation.

Table A.5: Out-of-Sample (One Year Ahead) Performance of the Ensemble (EBMA) Method, Varying Data Granularity

	Full Predictors	Only Lagged Indicator AR(2)	Add Intensity	Disaggregate Violence	All Lagged Violence Data
	(1)	(2)	(3)	(4)	(5)
	Indonesia, 2008–2014				
Any Incident (AUC)	0.818	0.745	0.776	0.797	0.807
≥ 5 Incidents (AUC)	0.939	0.854	0.925	0.933	0.936
≥ 1 S.D. Increase (AUC)	0.840	0.527	0.832	0.842	0.853
Predictor set:					
Two lags of dependent variable	Yes	Yes	Yes	Yes	Yes
Total incidents, deaths, injuries, damage	Yes		Yes	Yes	Yes
Disaggregate by violence type	Yes			Yes	Yes
Total village-level killings reported	Yes				Yes
Podes and DIMS	Yes				Yes
Economic and social characteristics	Yes				

Notes: Each model is trained on all data available preceding the out-of-sample prediction year. Training data starts in 2002 in Indonesia. Out-of-sample prediction begins in 2008 in Indonesia. The AUC is the area under the ROC curve, a measure of the trade-off between the true positive rate and false positive rate as we vary the discrimination threshold. We report average performance over the out-of-sample years above.

B Methodological Details

Each of the machine learning methods that we use involves a number of hyper-parameter choices. In general, the hyper-parameters govern the degree of flexibility afforded to each algorithm. The optimal choice of these parameters balances the models' ability to uncover complex relationships against the risk that the algorithm over-fits to noise in the data. Moreover, limits to computing power also prevent the use of certain parameter values. Since we run these algorithms several times with a number of different subsets of covariates, we needed a set of algorithms that would run in a manageable amount of time. The baseline set of predictions for Indonesia takes about two hours, and the predictions take about an hour and a half for Colombia. We choose these parameters using a mixture of rules-of-thumb and five-fold cross validation. In this section, we detail these choices.

LASSO. We implement a logistic LASSO where predictors are standardized to have a mean of zero and unit variance before they are fed into the fitting algorithm. LASSO and ridge regression are two closely-related penalized regression techniques. While LASSO penalizes the sum of the absolute value of the regression coefficients, ridge regression penalizes the sum of squares. We follow best practices as in Blair et al. (2017) and use a weighted average of the two penalties, where the weight for the LASSO penalty is $\alpha = 0.95$ and the weight on the ridge penalty is $1 - \alpha$.

For each country in each year of estimation, we find the optimal penalty parameter λ by searching over a grid of candidate values and testing the penalties using 5-fold cross validation. We repeat this process 10 times, and take the average optimal value.

Random Forests are collections of many trees which are fit to random subsets of the data and then are averaged together. The underlying logic is that the individual trees may be overfit to their respective dataset, but since each tree is fit using a different set of predictors, the overfitting averages out over the entire forest. We choose mean-squared error as the loss function for the random forest. Beyond the choice of the loss function, there are three important hyper-parameters

to choose.

First, we must specify a rule governing how large trees can be. We could specify the minimum number of observations per terminal node or the maximum number of terminal nodes. In general, larger trees afford more complexity, and we err on the side of larger trees as the risks of overfitting are mitigated by the averaging over the entire forest. However, larger trees also add to computation time. As a result, we limit our trees to 60 terminal nodes.

Second, we choose the number of trees in the forest. Since each tree is independent, additional trees simply reduce variance of estimates and do not add to bias. As a result, more trees are always better. At the same time, performance gains from additional trees generally diminish quickly, while the computing time costs of fitting individual trees does not. Therefore, we choose to fit 100 trees in each forest.

Finally, we must choose the number of covariates to be considered at each branch in the trees. We follow the rule-of-thumb of using one third of the potential covariates at each split. (Hastie et al., 2001)

Gradient Boosted Machines are comprised of decision trees that are fit sequentially to the residual variation in the predictand that was not predicted by previous trees. Unlike random forests which leverage many overfit trees, gradient boosted machines are meant to learn slowly, with each tree explaining a small amount of additional variation. Therefore, the key parameter is the shrinkage parameter which limits the extent to which each tree can contribute to the machine's overall prediction. Best practice sets these shrinkage parameters as low as possible. However, as the parameter gets lower, the number of trees required to get a good fit increases, as does computing time. We choose a shrinkage parameter $\lambda = 0.1$.

As for random forests, GBM requires the implementer to specify the complexity of trees and the number of trees in the ensemble. We specify the number of terminal nodes as 7 in each tree, a standard parameter value for these models. (Hastie et al., 2001) The number of trees to include in the model is the key hyper-parameter that drives the overfitting versus complexity trade-off. If there are too few trees, the predictions will be a very simple function of the predictors, whereas if the there are too many trees, the later trees will be fit to noise generated by idiosyncrasies of the the training sample. As a result, we choose the number of trees by 5-fold cross validation over a grid of candidate sizes. We average the results of 10 such trials to get an optimal number of trees in each year.

Neural Networks. Neural networks are built from weighted combinations of features, and an activation function which through which these combinations are passed. We use a single hidden layer neural network, and the network is trained via back-propagation. Our neural networks use a sigmoid activation function.

The major parameter governing complexity is the number of nodes to allow in this single layer. We choose the number of nodes by searching over a grid of values and employing 5-fold cross validation to test each candidate number of nodes. As with the other algorithms, we repeat this process 10 times, and choose the average parameter.

Since each predictor has a weight for each node in the hidden layer, training of the neural network can involve the computation of thousands of parameters. The computation costs are magnified during the grid search process. To alleviate this pressure, we preprocess the data by standardizing the predictors and calculating principal components of the predictor set. We use these principal components instead as predictors to be used in the neural network. We use 30 principal components in Indonesia and 20 in Colombia. This rotation of the predictor space dramatically increases speed without throwing away much important variation.

Our **Ensemble Bayesian Model Average** is computed by generating a 5-fold cross-validaton

set of probabilistic predictions for each algorithm using the parameters chosen above. We take these predictions and calculate posterior likelihoods that each model is correct given their predictions and the observed levels of violence. These likelihoods, when normalized to sum to 1 give us weights for our model average. We repeat this process 10 times to get ten sets of weights and average them to aggregate our predictions.

C Data Appendix

C.1 Indonesia Data

C.1.1 Administrative Divisions

Our unit of analysis for the SNPK data is the 2000 subdistrict. Subdistricts have increased in number over time and are mapped back to the larger 2000 units.

Concurrent with the wave of decentralization, the Indonesia government created many new districts through a process of redistricting known colloquially as *pemekaran* or blossoming. After remaining steady from 1980 to 1998, the number of new districts ballooned from 302 in 1999 to 514 in 2014. The proliferation of districts occurred across the entire archipelago, with the exception of most districts on the island of Java. New districts are formed when existing subdistricts break off from their original district and create their own local government, complete with a new capital, district head, parliament, and government apparatus. On occasion, one district can mushroom into multiple new districts. Subdistricts have also split (and a few amalgamated) over time.

The number of districts and the number of subdistricts have ballooned over time. Despite the increase in the number of districts, the number of subdistricts per district actually increase. The number of villages has remained relatively more steady and thus the number of these per-district and per-subdistrict decline over time. The unit of analysis in this paper is the subdistrict amalgamated to its 2000 borders.

C.1.2 Violence Data

The conflict data comes from the Indonesian National Violence Monitoring System (known by its Indonesian acronym SNPK). The data are reported at the 2011 subdistrict level and include incident dates. Sub-district codes are non-missing in 84% of cases. We aggregate incidents to the 2000 subdistrict borders in each year. Our main conflict measures are binary indicators for any conflict in a given subdistrict–year. Table C.1 presents the violence definitions in the SNPK.

Table C.1: Violence Definitions in the SNPK

Violent Crime	Criminal violence not triggered by prior dispute or directed towards specific targets.
Domestic Violence	Physical violence perpetrated by family member(s) against other family member(s) living under one roof/same house including against domestic workers and violence between cohabiting couples.
Violence during law-enforcement	Violent action taken by members of formal security forces to perform law-enforcement functions (includes use of violence mandated by law as well as violence that exceeds mandate for example torture or extrajudicial-shooting).
Resource Conflict	Violence triggered by resource disputes (land, mining, access to employment, salary, pollution, etc.).
Governance Conflict	Violence is triggered by government policies or programs (public services, corruption, subsidy, region splitting, etc.).
Popular Justice Conflict	Violence perpetrated to respond to/punish actual or perceived wrong (group violence only).
Elections and Appointment	Conflict Violence triggered by electoral competition or bureaucratic appointments.
Separatist Conflict	Violence triggered by efforts to secede from the Unitary State of the Republic of Indonesia (NKRI).
Identity-based Conflict	Violence triggered by group identity (religion, ethnicity, tribe, etc).
Other Conflicts	Violence triggered by other issue.

Table C.2 reports the rates at which each of these indicators occur. While around half of subdistricts experience some conflict in a given year, only around 12% experience more than five incidents, and even fewer experience a large increase in the number of events relative to the prior year.

Table C.2: Annual Rates of Conflict in Indonesian Subdistricts

	any conflict (1)	≥ 5 conflict incidents (2)	≥ 1 std. dev. increase in incidents (3)
2005	0.540	0.130	
2006	0.537	0.122	0.029
2007	0.509	0.114	0.028
2008	0.550	0.125	0.030
2009	0.527	0.120	0.031
2010	0.536	0.119	0.034
2011	0.531	0.127	0.040
2012	0.573	0.144	0.045
2013	0.541	0.127	0.034
2014	0.539	0.129	0.034

Notes: Conflict incidents above exclude crime and domestic violence.

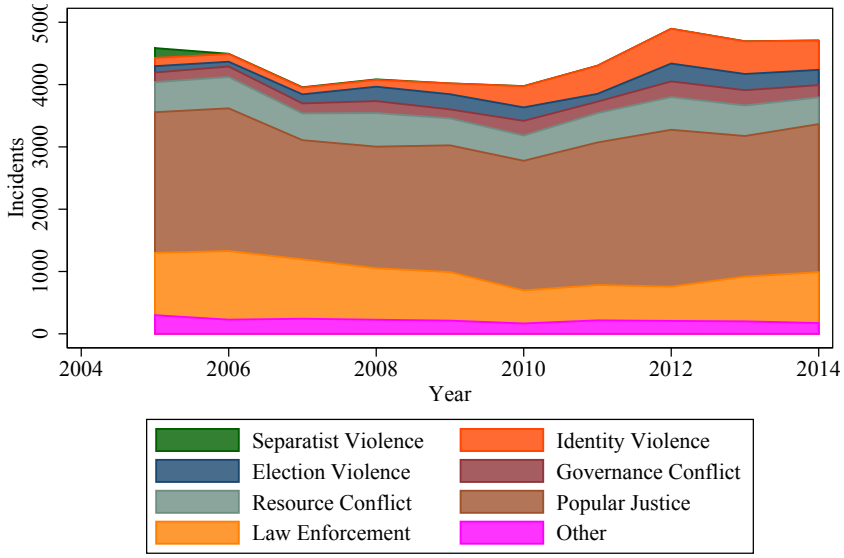
C.1.3 Covariates

2000 Population Census

We use the following predictors constructed at the 2000 subdistrict level from Indonesia's 2000 universal Population Census.

- **Employment Shares:** Fraction of persons in agriculture, forestry and fishing, industry, services, trade, and transportation. We compute the fraction of persons in self-employment and the fraction that are employers.

Figure C.1: Violence by Category, Indonesia



Notes: The graph above plots violent incident counts according to the SNPK. Crimes and domestic violence are not included. The panel is balanced from 2005 onward.

- **Demographic Variables:** Share of people in each of the following religions; Muslim, Catholic, Protestant, Hindu, Buddhist, Confucian, and other; share of the population that is Chinese and the share Arab; share of men and the share married. Finally, share of persons under 10 and from 10 to 30 years old.
- **Education:** Share of individuals who completed no school and the average years of schooling.
- **Rural Population Share:**

  Share of villages within the subdistrict classified as rural.
- **Ethnic Fractionalization:** Ethnic fractionalization in district d is given by $F = \sum_{g=1}^{M_e} \pi_g(1-\pi_g)$, where M_e is the number of ethnic groups in the district, and π_g is the population share of group g as reported in the 2000 Census. We observe over 1000 ethnicities and sub-ethnicities speaking over 400 languages.
- **Ethnic Polarization:** Same as ethnic fractionalization but own group share is emphasized. Specifically $P = \sum_{g=1}^{M_e} \pi_g^2(1-\pi_g)$.
- **Religious Fractionalization:** Religious polarization, $R = \sum_{g=1}^{M_r} \sum_{h=1}^{M_r} \pi_g \pi_h$, where M_r is the number of religious groups, and π_g (π_h) is the population share of group g (h). There are seven religions recorded in the Census, but in most districts, there is a single cleavage between a Muslim and a non-Muslim group.
- **Religious Polarization:** Same as ethnic Polarization but own group share is emphasized. Specifically $RP = \sum_{g=1}^{M_e} \pi_g^2(1-\pi_g)$.
- **Ethnic Residential Segregation:** Following Alesina and Zhuravskaya (2011), we use the 2000 census to compute Ethnic segregation by comparing ethnic fractionalization at the village level to that of the subdistrict level. Specifically we compute:

$$S = \frac{1}{M-1} \sum_{m=1}^{M} \sum_{s=1}^{S} \frac{t_s}{T} \frac{(\pi_{sm} - \pi_m)^2}{\pi_m}$$

M is the number of ethnic groups, T is the total population of the subdistrict, t_s is the population in village s, π_m is the fraction of group m in the district, and π_{sm} is the fraction of group m in subdistrict s. We drop the smallest 1% of ethnic groups so that M remains reasonable (< 25).

Night Lights

- **Light Intensity:** Annual night light data to proxy for GDP (Henderson et al. (2012)). We use mean stable light intensity at the village level, which ranges from 0 to 63. This attempts to filter out background noise and unstable sources of light. We compute the (population weighted) average light intensity across villages at the 2000 subdistrict boundary level.

Potensi Desa (Podes)

We construct the following variables using the quasi-triennial administrative village census *Potensi Desa*, abbreviated Podes. We use the 2000, 2003, 2005, 2008 and 2011 rounds.

- **Police and Security:** The population weighted (across villages) distance to the nearest police post in all rounds. In addition we compute the share of villages with any security or police post in each round.
- **Distance to the District Capital:** The population weighted (across villages) distance to the district capital in 2000, 2003, 2008, and 2011. It can vary due to the creation of new districts capital as a result of district proliferation.
- **Population:** Subdistrict population in each round.
- **Health Care:** For the 2000 wave, we record of the number of health care facilities, which we classify into hospitals (hospital, maternity hospitals, and policlinics) and middle care (puskesmas and supporting puskesmas).
- **Schooling:** For the 2000 wave, we compute the number of junior high schools and high schools per capita. We also include the percentage of villages with a madrasa or pesantran.
- **Conflict:** In each wave we record the share of villages that experienced any conflict. We do the same for each type of conflict recorded (Theft, Robbery, Thuggery, Arson, Rape, Drug-related, Murder, Child Trafficking). We also include the total number of deaths and injuries reported.
- **Natural Disasters:** Each wave records the share of villages that experienced a mudslide, flood, fire, or earthquake in the past three years.

Global Precipitation Climatology Project (GPCP)

The Global Precipitation Climatology Project (GPCP) provides annual rainfall at the district level. We calculate historical averages and calculate annual deviations from the historical average.

University of Delaware Global Climate Resource Database

The University of Delaware Global Climate Resource Database provides monthly rainfall and temperature data at the subdistrict level from 1900 onward. We calculate historical averages and annual deviations from the average.

Crop Shocks

- **Food Price Shocks:** By using the 2003 Agriculture Census we calculate the quantity of each crop produced within the 2000 subdistrict borders. We use the UN Food and Agriculture price series for each crops to construct log changes in crop prices and weight these changes by the 2003 production share. We group into cash crops and food crops.

- **Agricultural GDP:** By using the 2003 Agriculture Census we calculate the quantity of each crop produced within the 2000 subdistrict borders and then use the UN Food and Agriculture price series to construct total agricultural value in each year.

Mineral Shocks

Geocoded mine data from the SNL Mine Production Data. For each mineral, we calculate the distance from the center of the subdistrict to the nearest active mine producing that mineral. We multiply log mineral price changes from the World Bank's GEM Commodity Price Database by the inverse of the distance to the nearest mine.

Disaster Information Management System

The Disaster Information Management System lists the major disasters such as floods or volcano eruptions at the district level.

National Socioeconomic Survey (*Susenas*)

The National Socioeconomic Survey (*Susenas*) is an annual survey of about 200,000 national representative households. The survey measures household income and expenditures. We use the expenditure data to construct average expenditures, an expenditure Gini, the ratio of the 80th percentile to the 20th percentile of inequality, and we use the wage data to construct median wages in agriculture and non-agriculture, as well as differences between the 75 percentile and the 25th percentile. We also use this survey to get an annual population measure.

Database for Policy and Economic Research (DAPOER)

The World Bank's Indonesia Database for Policy and Economic Research (DAPOER), which in turn obtains data from the Indonesia ministry of finance data, to keep track of total **district** revenue in each year.

- **Total District Revenue Per Capita:** District revenue figures come from the World Bank's Indonesia Database for Policy and Economic Research (DAPOER), which in turn obtains data from the Indonesia ministry of finance data. They are given for each district at the time of existence. We aggregate up to the 2000 district boundary and separately also consider only parents. Population data is taken from the same dataset. All figures are inflation adjusted using 2010 as the base year.
- **DAU/DAK Revenue Per Capita:** District revenue in Indonesia is divided into a general allocation grant (*Dana Alokasi Umum*, DAU), some shared taxes, shared natural resource rents, and the special allocation grant (*Dana Alokasi Khusus*, DAK), as well as limited own revenue. DAU/DAK revenue focuses on the portion of revenue not due to natural resources or shared taxes.
- **Initial Resource Revenue:** Natural resource revenue such as that from oil/gas and mines is first transferred to the center and then partly returned to the district (and to a lesser extent nearby districts) based on percentages that vary by product and over the course of the study period. We use the level in 2000 to proxy, albeit imperfectly, for the presence and value of natural resources in the original district.

Political Data

- **Direct Election Data:** Direct local elections for district heads were phased in beginning 2005 across districts and occur every 5 years. We record the date of each of these elections.
- **Vote Share Polarization within 2000 subdistrict:** Data on vote share by party and subdistrict in the 1999 district parliamentary (DPRDII) elections were used—the first of the post-Suharto era—to construct a measure of vote share polarization at the subdistrict level. Forty-eight parties competed in these elections.

- **Time-varying vote Share Polarization at District Level:** To construct time-varying vote share polarization measures we use national parliamentary vote share data in 1999, 2004, 2009 recorded at the district level.
- **Party Shares:** Using the national vote data we also retain the votes for certain parties. We keep Golkar and PDIP shares, as well as vote shares for Islamist parties.

Topographical Variables

- **Slope and Elevation Data** Topographical variables were created using raster data from the *Harmonized World Soil Database* (HWSD), Version 2.0 (Fischer et al., 2008). The raster files are compiled from high-resolution source data and aggregated to 30 arc-second grids. The terrain, slope, and aspect database provided by HWSD researchers was compiled from a high-resolution digital elevation map constructed by the Shuttle Radar Topography Mission (SRTM).

  Elevation data were computed for each village as the average elevation over the entire village polygon, using raster data from HWSD. Slope and aspect data were also recorded for each village and calculated similarly. Variables equal to the average share of each village corresponding to each slope class (0-2 percent, 2-4 percent, etc.) were constructed using ArcView.
- **Ruggedness** A 30 arc-second ruggedness raster was computed for Indonesia according to the methodology described by Sappington et al. (2007), and village-level ruggedness was recorded as the average raster value. The authors propose a Vector Ruggedness Measure (VRM), which captures the distance or dispersion between a vector orthogonal to a topographical plane and the orthogonal vectors in a neighborhood of surrounding elevation planes.

  To calculate the measure, one first calculates the x, y, and z coordinates of vectors that are orthogonal to each 30-arc second grid of the Earth's surface. These coordinates are computed using a digital elevation model and standard trigonometric techniques. Given this, a resultant vector is computed by adding a given cell's vector to each of the vectors in the surrounding cells; the neighborhood or window is supplied by the researcher. Finally, the magnitude of this resultant vector is divided by the size of the cell window and subtracted from 1. This results in a dimensionless number that ranges from 0 (least rugged) to 1 (most rugged).8.
- **Soil Quality** We also make use of the HWSD data for soil quality measures. HWSD provides detailed information on different soil types across the world. The HWSD data for Indonesia is taken from information printed in the FAO-UNESCO Soil Map of the World (FAO 1971-1981), a map printed at a 1:5,000,000 scale. For each subdistrict, we we use the following measures of soil types: percentage of land covered by coarse, medium, and fine soils, percentage of land covered by soils with poor or excessive drainage, average organic carbon percentage, average soil salinity, average soil sodicity, and average topsoil pH.

While each of the above datasets covers the entire country, there are inevitably minor missing data issues as we combine so many sources. Rather than exclude entire predictors or observations, we impute these missing predictors via regression on contemporaneous predictors for which data is available. Violence measurements are not used for the imputation of other variables, nor are any violence measures imputed. The sample is restricted to observations where we have full violence data.

C.2 Colombia Data

C.2.1 Administrative Divisions

The municipality is the second level of administrative authority in Colombia (the first is the Department) and is the fundamental territorial entity in the political-administrative division of the State. It has political, fiscal and administrative autonomy within the framework of the Colombian law.

As of 2015 Colombia has 1101 municipalities in 32 departments. The departments are composed by municipalities and are also a territorial entity with administrative autonomy. They must perform administrative and coordination functions complementing the municipal action and should serve as intermediaries between the Nation and the municipalities.

C.2.2 Violence Data

Our conflict data in Colombia comes from the Conflict Analysis Resource Center (CERAC) which contains data on military confrontation from 1988 to 2005. The data are reported at the event level and episodes are characterized either as bilateral clashes between sides or unilateral attacks from one side against another. We aggregate incidents by municipality-year in our main specification including events involving all three conflict actors: the guerrillas, the paramilitaries, or the government. However, we also consider a specification where the aggregation excludes government attacks or clashes (results available on request). Our main dependent variables are binary indicators for any event in a given municipality-year; more of than five events and more than 1 standard deviation increase in violence.

Table C.3 presents a descriptive analysis of the dependent variable (total onsets, any incident, more than five incidents or 'high', and greater than a 1 SD increase in violent events or 'spike') for Colombia. We cover the period between 1992 to 2005 at the municipality level and calculate the overall, within and between variation for each dependent variable.

Table C.3: Annual Rates of Conflict in Colombian Municipalities

	any conflict	≥ 5 conflict incidents	≥ 1 std. dev. increase in incidents
	(1)	(2)	(3)
1992	0.352	0.081	0.039
1993	0.323	0.049	0.021
1994	0.340	0.063	0.043
1995	0.276	0.057	0.018
1996	0.317	0.066	0.042
1997	0.301	0.056	0.031
1998	0.375	0.079	0.045
1999	0.374	0.073	0.040
2000	0.417	0.106	0.073
2001	0.421	0.122	0.064
2002	0.454	0.145	0.082
2003	0.376	0.101	0.038
2004	0.335	0.099	0.053
2005	0.308	0.068	0.020

Notes: Conflict incidents above include paramilitary attacks, guerrilla attacks, government attacks, and bilateral clashes between these groups.

C.2.3 Covariates

Demography and Economic Activity

- **Demographics:** Information on population by municipality-year was included from the National Administrative Department of Statistics (referred as DANE). We also consider information on population density and binary variables classifying large municipalities (e.g., more than 250 thousand inhabitants) and metro areas.
- **Commodities and international prices:** Commodity data compiled by (Dube and Vargas, 2013) was used in this study. We consider both the local production and international price of coffee, oil, coal, silver, platinum, and precious metals.
- **Local government finance:** Fiscal revenue data from the National Planning Department (NPD) was collected. Specifically, information on municipal income, spending, deficit and transfers.
- **Wellbeing:** Some welfare measures were included. In particular, variables like the Gini coefficient for land inequality; unmet basic needs (1993) and life quality index.
- **Infrastructure:** We consider both paved and unpaved roads for main and secondary roads in thousand of kilometers (1995). Also the density of those roads measures as kilometers over kilometers square (Km/Km^2).
- **Illicit production:** We obtain data on coca cultivation for Colombia from two sources: Direccion Nacional de Estupefacientes (DNE) and from the United Nations Office of Drug Control (UNODC). Additionally, we also incorporate information on coca cultivation for Bolivia, Peru and the world, along with exportation to certain destinies and prices.

Geographic and Weather

- **Geographic:** The main features we expoit are: surface area; proportion of non-habitate land; suitable land for sugar cane or palm; surface; terrain stepness; water availability; main; secondary and tertiary rivers longitude; flat land; hills; mountains; valleys and water bodies as a percentage of municipal land; proportion of hilly; montanous and rugged terrain; maximun slope.
- **Weather:** The University of Delaware Global Climate Resource Database provides monthly rainfall and temperature data at the municipal level from 1900 onward. We calculate historical averages and annual deviations from the average for each municipality-year.

Historical Variables

From Acemoglu et al. (2015), we include measures of colonial institutions and infrastructure, such as number of crown employees, presence of colonial cities and royal roads, population in 1843, slave share of the population in 1843, number of indians in 1560, number of encomiendas in 1560, colonial gold mines, foundation dates, and population in 1843.

Political Data:

- **Election Data:** Information on mayoral and congressional elections (lower and upper house) were used from the entire period of analysis. Universidad de los Andes compiled a database of electoral results since 1958 and has been updating it until 2014. The original data comes from the Registraduria Nacional del Estado Civil ("National Registry")[1]. In particular, we generate parties' vote shares, turnout, and time dummies for electoral periods.

[1]Universidad de los Andes CEDE makes the database publicly available through its' database website (https://datoscede.uniandes.edu.co).

- **Vote Share Polarization within municipality:** Each of this elections were meant to represent different levels of political power in Colombia, at the local, regional and national level. More specifically, we consider measures of concentration, polarization and fractionalization index for competitiveness of the elections; margin between winner and runner-up; party's vote share and political leaning.

Distance to DMZ (Delimitarized Zone)

The Caguan DMZ was a delimitarized zone of 42.000 square kilometers in southern Colombia authorized by the goverment to negotiate a peace process with the FARC-EP. The region was made up by the municipalities of Vista Hermosa, La Macarena, La Uribe, Mesetas, and San Vicente del Caguan. The DMZ started in January of 1999 and ended in February 2002. Its existence coincides with the scalation of the conflict in Colombia, therefore we calculate the distance of each municipality in Colombia to the DMZ as a covariate.

US Military Aid

We use the dataset created by (Dube and Naidu, 2015) of US military aid and Colombian military bases. One feature of US military aid is that it is disbursed to particular Colombian military brigades, each of which is attached to and operates out of a particular government military base. We consider the natural log of US military and antinarcotics aid to Colombia interacted with a dummy that defines if a particular municipality has a military base. In total, we covered 34 municipalities with military bases, of which 32 appear in the sample for which the conflict data is available.

As for Indonesia, we restrict our sample to observations with full violence data. There are inevitably minor missing data issues as we merge other covariates. Rather than exclude entire predictors or observations, we impute these missing predictors via regression on contemporaneous predictors for which data is available. Violence measurements are not used for the imputation of other variables, nor are any violence measures imputed.

C.3 Predictor Groupings

Tables C.4 and C.5 report the variables that are included in each of the groups displayed in the predictor breakdown graphs.

Table C.4: Indonesia Predictor Groups

SNPK

- Incident Counts: total, total excluding crime, resource, governance, election, identity, popular justice, law enforcement, crime, domestic violence, separatist, other.
- Deaths: total, total excluding crime, resource, governance, election, identity, popular justice, law enforcement, crime, domestic violence, separatist, other.
- Injuries: total, total excluding crime, resource, governance, election, identity, popular justice, law enforcement, crime, domestic violence, separatist, other.
- Damaged Buildings: total, total excluding crime, resource, governance, election, identity, popular justice, law enforcement, crime, domestic violence, separatist, other.
- Destroyed Buildings: total, total excluding crime, resource, governance, election, identity, popular justice, law enforcement, crime, domestic violence, separatist, other.
- Total active newspapers used in data collection

PODES Violence Data

- Percent of villages with any: theft, robbery, looting, thuggery, arson, rape, murder, fights among citizens, fights with security officers, fights among students, fights among tribes, any conflict.
- Deaths, injuries
- Distance to nearest police post, number of security posts, number of police posts, number of security officers

DIMS Violence Data

- Terrorism: deaths, counts
- Unrest: deaths, counts

Religion

- Fractionalization, Polarization
- Population Share: Muslim, Catholic, Protestant, Hindu, Buddhist, Confucian, Christian
- Percent of villages with any Islamic boarding school

Ethnicity

- Fractionalization, Polarization, Segregation
- Population Share: Chinese, Arab

Education

- Mean years of schooling
- Share with no schooling
- Junior high schools, high schools

Demographics

- Male share, married share, share below 10 years old, share from 10 to 30 years old

Remoteness

- Rural population share, distance to district capital, distance to province capital, land area
- Main road electrified, road access year-round, telephone signal strength
- Has: airport, bridge, terminal, seaport

Health Infrastructure

- Sophisticated health institutions, mid-level health institutions

Disasters

- Number of mudslides, floods, earthquakes, fires, other disaster (from PODES)
- Count: land abrasions, disease outbreaks, droughts, earthquakes, floods, forest fires, industrial and transportation accidents, landslides, tornadoes, tsunamis, volcanoes, total
- Deaths: land abrasions, disease outbreaks, droughts, earthquakes, floods, forest fires, industrial and transportation accidents, landslides, tornadoes, tsunamis, volcanoes, total

Elections

- 1999 local elections: mainstream party vote share, Islamist party vote share, fractionalization, polarization, voter turnout
- Election year indicator
- National elections: mainstream party vote share, Islamist party vote share, fractionalization, polarization, voter turnout

Agricultural Infrastructure

- Mean access to irrigation, small rice mills per village, large rice mills per village

Finance and trade institutions

- Banks per capita, rural cooperative banks per capita
- Percent of villages with permanent market

Economic Structure

- Output share: major cash crop, cash crop, major food crop
- Share of households with any agricultural land, share with more than 0.1 hectare of agricultural land
- Percent of villages where rice is the primary commodity
- Population share in: agriculture, forestry, industry, trade, service, transport
- Output share: trading, self-employment, agriculture, services, employers
- Distance to nearest mine: bauxite, coal, copper, gold, iron ore, nickel, tin, zinc, silver

Government Finance

- Revenue from: total, village resources, taxes, social organizations, ROSCAs, other villages, higher government administrative unit, central government, provincial government, district government, natural resource tax-sharing, non-natural resource tax-sharing, total self-generated
- Expenditures: routine, development projects

Output

- Agricultural GDP constructed from 2002 output weights and current commodity prices
- District GDP, district agricultural GDP
- Unemployment
- Nighttime light intensity

Commodity Price Shocks

- Cash crop, major cash crop, major food crop, bauxite, coal, copper, gold, iron ore, nickel, tin, zinc, silver

Historical Commodity Price Shocks Volatility

- Cash crop, major cash crop, major food crop

Climate Shocks

- Temperature and rainfall deviations from historical means

Historical Climate Traits

- Mean: temperature and rainfall
- Standard deviation: temperature and rainfall

Table C.5: Colombia Predictor Groups

Violence

- Attacks: guerilla, paramilitary, government, total
- Clashes, total attacks and clashes, total casualties
- Massacres: guerilla, paramilitary
- Infrastructure paramilitary attacks, non-infrastucture paramilitary attacks

Production

- Oil production (1988), coal reserves (1978)
- Department coal production, department gold production

Government Finance

- Income: total, capital, tax, non-tax, transfer, land taxes, commerce taxes, other taxes
- Expenditure: total, capital spending, total functional spending, personnel spending, transfers paid
- Municipal expenditure, national expenditure
- Budget deficit, credit

Wellbeing

- Land ownership gini
- Unmet basic needs index, life quality index

Infrastructure

- Main, secondary, tertiary paved roads: length and density
- Main, secondary, tertiary unpaved roads: length and density
- Main, secondary, tertiary dirt roads: length and density

Elections

- Mayoral election: indicator, turnout, winner's party
- Lower and upper house elections: indicator, turnout, concentration, fractionalization, polarization, margin between 1st and 2nd party, party leaning of winner, party leaning of runner up, party of winner and runner-up
- Presidential election: indicator

U.S. Military

- U.S. military aid, U.S. narcotics aid, U.S. combined aid
- Colombian government military base presence
- Interaction between base presence and U.S. spending

DMZ

- Demilitarized zone indicator, distance to demilitarized zone

Illicit Production

- Poppy producing hectares, any poppy production, coca producing hectares, any coca production
- Eradicated hectares, any eradication
- Coca production in: Bolivia, Colombia, Peru, world total, non-colombia
- Cocaine production in: Bolivia, Colombia, Peru
- Coca eradicated in: Bolivia, Colombia, Peru
- Cocaine exports to Switzerland: Bolivia, Colombia, Peru
- Coca producing hectares in: Bolivia, Colombia, Peru, world total, non-colombia
- Change in coca price: wholesale U.S., wholesale Europe, retail Europe
- Coca hectares times U.S. cocaine price

Colonial History

- Employees: crown, non-military crown
- Colonial state presence index
- Colonial city status dummy, distance to local roads
- Slave population share, slave presence indicator
- Indian population
- Encomiendas: number, indicator
- Colonial gold mine presence indicator
- Foundation date
- Population 1843

Geography

- Area
- Percent suitable/non-suitable/sub-optimal/optimal: sugar, palm
- Slope: max, mean, standard deviation
- Percent flat, slightly-sloped, sloped, strongly-sloped, moderately steep, strongly steep
- Water availability: mean, standard deviation
- Main, secondary, and tertiary rivers: length and density
- Percent flat, hills, mountains, valleys, water bodies.

Commodity Price Shocks

- Coffee, oil, gold, coal, silver, platinum

Climate Shocks

- Temperature and rainfall deviations from historical means

Historical Climate Traits

- Mean: temperature and rainfall
- Standard deviation: temperature and rainfall

Climate Shocks

- Temperature and rainfall deviations from historical means

Historical Climate Traits

- Mean: temperature and rainfall
- Standard deviation: temperature and rainfall