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 Peck*

Abstract—How feasible is violence early-warning prediction? Colombia and Indonesia have unusually fine-grained data. We assemble two decades of local violent events alongside hundreds of annual risk factors. We attempt to predict violence one year ahead with a range of machine learning techniques. Our models reliably identify persistent, high-violence hot spots. Violence is not simply autoregressive, as detailed histories of disaggregated violence perform best, but socioeconomic data substitute well for these histories. Even with unusually rich data, however, our models poorly predict new outbreaks or escalations of violence. These "best-case" scenarios with annual data fall short of workable early-warning systems.

I. Introduction

ADVANCES in data and computing techniques have kindled hopes that civil society, police, or peacekeepers will be able to 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 macrolevel 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.

Subnational predictions could prove more fruitful. The past decade has seen the study of conflict push down to the microlevel causes, processes, and consequences of violence, and we can avail ourselves of these data to investigate prediction. 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 is unknown, however. Now is a good moment to take stock of what existing methods and the richest available microdata can deliver.

This paper takes advantage of high-quality, extensive annualized 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-quality, local-level panel data in the developing world. This includes a trove of information on annual socioeconomic conditions and other characteristics, plus more than a decade of subdistrict- or municipal-level data on annual local violence.

We chose these two cases because they are among the current best-case scenarios in terms of data availability on both violence and potential predictors of violence in annual panel form. Such annual data are, to date, the most common kind of data available across countries (especially for predictor variables). If conflict prediction proves fruitful in these two instances, they could be models for other prediction efforts. If not, then we must ask where or with what other forms of data we can expect early warning systems to bear fruit (if any). 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 data sets 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 & Miguel, 2010). Each country also has high-quality violence data from which we obtain our outcomes. Using data from 1998 to 2014 in Indonesia, we study conflict related to interethnic and religious tensions, as well as electoral, governance, and resource disputes, among others. In Colombia, our data span more than a quarter century, from 1988 to 2014. We predict clashes between state, guerrilla, and paramilitary forces during a period of protracted civil conflict. These granular violence data also allow us to leverage detailed violence histories, adding to an already rich set of predictors.

We then use 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 four to thirteen years of data and forecast local conflict during the following year. We generate and assess predictions for whether there will be any violence next year, as well as whether there will be a large number of events or an escalation. We further examine predictive power across space as well as time.

Received for publication February 25, 2019. Revision accepted for publication September 10, 2020. Editor: Rema N. Hanna.

*Bazzi: Boston University; Blair: Brown University; Blattman and Dube: University of Chicago; Gudgeon: U.S. Military Academy; Peck: Northwestern University.

For comments, we thank Austin Wright and seminar participants at ESOC, MWIEDC, ISF, NBER Economics of National Security, and NEUDC. Miguel Morales-Mosquera and Sebastian Hernandez provided excellent research assistance.

A supplemental appendix is available online at https://doi.org/10.1162/rest_a_01016.

[1] The Political Instability Task Force's prediction efforts are likely the best known (Goldstone et al., 2010). For other examples of cross-national prediction studies, see Beck, King, and Zeng (2000); Blair and Sambanis (2020); Brandt, Freeman, and Schrodt (2011); Celiku and Kraay (2017); Gleditsch and Ward (2013); Gurr and Lichbach (1986); Harff (2003); Hegre et al. (2013, 2016); Perry (2013); and Ward et al. (2013). For an early exception see Schrodt (2006), who studies violence in the Balkans.

The Review of Economics and Statistics, July 2022, 104(4): 764–779

https://doi.org/10.1162/rest_a_01016

Our results illustrate both the promise and pitfalls of annual 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 combines the best new methods, performs better than previous subnational attempts (Blair, Blattman, & Hartman, 2017; Colaresi, Hegre, & Nordkvelle, 2016; Weidmann & Ward, 2010; Witmer et al., 2017).[2] We view these results as especially important given that such local hot spots can pose a serious risk of regional or national escalation, and some of these locations will not be known to policymakers, especially in large, diverse countries.

Local violence is not merely autoregressive: a model consisting of the lagged dependent variable alone performs consistently poorly. Nonetheless, more nuanced histories of violence tend to be very strong predictors of violence. 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, our covariates also predict hot spots 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 remoteness as well as sectoral shares of the local economy. Time-invariant predictors alone do just as well as detailed violence histories.

Surprisingly, predictive accuracy improves little when we add time-varying factors, including economic output, government finance, communication infrastructure, natural disasters, elections, and fluctuations in rainfall, temperature, commodity prices, drug production, and U.S. military activity. This stands in contrast to a large causal literature on conflict, where an array of findings associates economic and political shocks with intensified violence (Bazzi & Blattman, 2014; Berman & Couttenier, 2015; Blattman & Miguel, 2010; Burke, Hsiang, & Miguel, 2015; Dube & Vargas, 2013; Miguel, Satyanath, & Sergenti, 2004).[3]

Our algorithms' strong performance is mainly driven by forecasts of where, but not when, violence is likely to occur. Our models perform poorly when predicting annual deviations from average levels of violence over the study period. Even nuanced histories of violence, a rich set of covariates, and the most widely studied economic and political shocks do not help us identify what hot spots are likely to get hotter in the coming year.

In contrast, our models perform well when we forecast conflict across space. We use training data from all 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 some of 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 may be a long way off.

We see these patterns consistently across two different country cases, with very different forms of violence. Taken together, our results are both encouraging and disappointing. On the one hand, we are able to predict hot spots for local violence remarkably well, and much more accurately than previous exercises of this sort or simpler benchmark models. 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 annual 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, at least with the data most commonly at hand: human-coded violence data drawn from newspapers and observer reports, annual aggregates of violence counts, and annual predictor variables. This pessimistic conclusion would resonate with warnings that big data and machine learning may not deliver the precision that policymakers long for (Jasny & Stone, 2017, p. 469).[4] 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 are feasible but require longer, more high-frequency data or additional or different risk factors. Longer training samples could give algorithms more variation to train on, and in principle this could help them to identify more complex relationships

[2] Our improved performance has multiple explanations. We exploit a much longer panel and a much wider variety of data sources than Blair et al. (2017) do. We also test a more diverse set of prediction algorithms, and ensemble routines have been shown to surpass the performance of any given model (Montgomery, Hollenbach, & Ward, 2012). Additionally, our event-based data rely on local media and are more comprehensive and less prone to misreporting than the widely used Armed Conflict Location and Event Data (ACLED) data. Finally, we draw on a much wider set of subnational predictors that go beyond lagged violence and country-level predictors.

[3] We find similar associations of commodity price and weather shocks with conflict in Colombia and Indonesia using our data sets (see appendix C4). While these shocks may cause conflict and help us forecast conflict hot spots, they add little to our ability to forecast conflict over time.

[4] Of course, conflict may be more difficult to predict than other policy-relevant outcomes. For example, Guha and Ng (2019) strike a more hopeful note in the context of predicting retail sales in a high-frequency, subnational panel.

with time-varying predictors. If true, this implies that there will be large gains to collecting longer conflict time series, for example, by delving into past historical archives. That said, our longest training samples are already decades long, and longer time series can introduce their own challenges, like potential structural breaks in the violence-generating process. There could also be gains from using higher-frequency data or more data on new risk factors—data from mobile phones, social media, local price fluctuations, and so forth.

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, Kalyanaraman, and Linardi (2014) use cell phone call patterns to predict temporal variation in conflict. 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 offers several results that future work should leverage to explore the promise of machine learning methods for modeling and predicting conflict. Overall, conflict can be forecast well across space but not over time with annualized panel data. 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. Cross-sectional hot spot prediction systems are probably feasible in a wide range of countries, even if temporal early warnings system may not be.

II. Settings

A. *Indonesia*

Following the 1998 collapse of Suharto's authoritarian regime, Indonesia experienced large-scale collective violence.[5] 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. After 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 deaths. Deadly conflict 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 & Gudgeon, 2021; Pierskalla & Sacks, 2017), with economic inequality (Indra, Hartono, & Sumarto, 2019), and with natural disasters, weather shocks, and commodity price fluctuations (Barron et al., 2009; Wright & 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.

B. *Colombia*

Colombia's long-running civil war has resulted in 220,000 deaths and 25,000 disappearances, and the forced displacement of over 5 million civilians (Historical Memory Group, 2013). During our primary analysis period, 1988 to 2005, the conflict mainly involved left-wing guerrilla groups, the government military, and right-wing paramilitary groups. The insurgency was 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, though residual violence (including killings of civil society leaders) continues.

Scholars have linked regional variation in conflict in Colombia to a host of political and economic factors, including shocks to drug production (Angrist & Kugler, 2008), fluctuations in commodity prices (Dube & Vargas, 2013), revenue decentralization (Chacon, 2014), collusion between paramilitaries and politicians (Acemoglu, Robinson, & Santos, 2013), American military aid (Dube & Naidu, 2015), and military incentives in the targeting of civilians (Acemoglu et al., 2016). As in Indonesia, it is unknown whether these causal factors are also good predictors of violence.

III. Data

An important contribution of this study is the two data sets 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 can be used for purposes of prediction.

A. *Indonesia*

Our units of analysis are Indonesia's third-tier administrative divisions, known as subdistricts (*kecamatan*). The

[5] For detailed accounts see Barron, Jaffrey, and Varshney (2014, 2016); Tadjoeddin (2014).

country had 7,094 subdistricts in 514 districts in 2014. These subdistricts had a median population of around 22,000.[6] While districts are the key autonomous administrative units, responsible for providing major public goods, subdistricts are also important sites of political organization. They are also the most granular level at which violence can be systematically tracked over time.

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 fifteen provinces plus parts of three provinces in greater Jakarta beginning in 2005. The data are not formally representative of Indonesia, but by 2005, they span all major island groups and cover 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 and appointments, governance, resource violence, popular justice, separatist, and other (could not be classified). Appendix table C1 defines each of these.

SNPK offers uniquely rich data on violence at the microlevel. In 2014, the architects of SNPK wrote, “As far as we know, the [SNPK] is the largest dataset of violence created for any single country” (Barron et al., 2014). Barron et al. (2014, 2016) provide additional detail on the multiyear process involved in creating this data set, including the source selection process and various quality control measures.[7]

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.[8]

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.

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—a “hot spot.” 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 appendix A.2; 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.[9] 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.[10]

Covariates. In addition to detailed violence histories from SNPK and *Podes*, we assemble a set of 482 subdistrict-level predictors from multiple data sources.[11] In order to assess how predictive performance differs across related covariate sets, we group covariates into the following predictor groups: (a) population, (b) religion, (c) ethnicity, (d) demographics (e.g., fraction male, fraction young), (e) education (e.g., mean years of schooling, school presence), (f) health (e.g., doctor and facility presence, rates of self-reported health problems), (g) geography (e.g., soil quality, ruggedness), (h) remoteness (e.g., distance to the capital, road presence, transportation terminals), (i) sector shares (e.g., population share in industry or agriculture), (j) agricultural features (e.g., irrigation access, share of households with agricultural land), (k) Public goods (e.g., safe water, garbage disposal, electricity and gas sources), (l) output (e.g., night light intensity, district gross domestic product), (m) distributional measures (e.g., inequality and poverty), (n) communication (e.g., access to telephones, cell signal), (o) government revenues and expenditures, (p) electoral outcomes (e.g., vote share polarization), (q) natural disasters, (r) weather histories and shocks, and (s) commodity shocks (e.g., food, cash crop, and mineral price shocks).

We use time-varying covariates whenever possible. Some, like geographic features, are time invariant by nature. Others, like ethnic and religious shares from the 2000 Population Census, were only observed once, at the beginning of the study period, and so they are time invariant in our panel. Covariates derived from *Podes* are time varying but only observed triennially. Other variables, like education and health

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

[7] The data and supporting documentation can be accessed at https://microdata.worldbank.org/index.php/catalog/2626.

[8] 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).

[9] In appendix A.5, we consider prediction of future violent events in places that are not currently experiencing violence.

[10] In appendix A.6, we consider alternative magnitudes and time horizons of escalations. We generally find the measure of a 1 standard deviation increase in incidents to be most predictable, but the differences when predicting other definitions of escalation are not substantial.

[11] Unless specified otherwise in appendix C.1, these measures are available at the subdistrict level or finer and are aggregated to their subdistrict boundaries in 2000.

infrastructure, are slow moving by nature. The violence data, night lights, economic shocks, as well as district-level revenues, GDP and unemployment, and poverty and inequality, among others, vary annually. Our models also include two lags of all violence predictors as well as several lags of our other time-varying predictors subject to availability.[12] We assess how predictive performance varies when using only time-invariant or time-varying covariates.

Details on sources and variable construction can be found in appendix C.1. Summary statistics for each covariate, organized by predictor group, can be found in appendix C.3, while details on sources and variable construction can be found in appendix C.2.

B. Colombia

Our unit of analysis for Colombia is the municipality (*municipio*). The country had 1,023 municipalities in our study period, averaging 37,000 residents—about two-thirds larger than the average Indonesian subdistrict. These are the main administrative units in the country, and they play a key role in the allocation and contestation of public resources.

Municipality-level violence data. The Conflict Analysis Resource Center (CERAC) provides data on armed confrontations from 1988 to 2005. We also extend beyond 2005 by using an additional source of conflict data collected by Universidad del Rosario through 2014 (see appendix A.8).

The CERAC data set contains over 21,000 events from the Colombian civil war, drawn from periodicals published by two Colombian NGOs, the Center for Research and Popular Education/Peace Program (CINEP) and Justicia y Paz. These periodicals are based on two underlying sources: reports of political violence and human rights abuses that appear in 25 printed media outlets with local and national coverage and a broad network of priests from the Catholic church with representation in almost every Colombian municipality, who report conflict episodes to CINEP. The priests are seen as neutral actors, often serving as negotiators between the two sides. Thus, their accounts are viewed as both credible and indispensable for attaining a comprehensive picture of conflict events in rural areas. Reliance on a large number of media sources as well as the priests leads to the inclusion of every part of the Colombian territory: violence events are reported in over 950 of 1,000 Colombian municipalities over this eighteen-year period.

All events in the CERAC data are hand-coded and checked extensively to ensure accuracy, with the precise coding procedures documented in Restrepo, Spagat, and Vargas (2004). For example, all large events associated with double-digit casualties and a random sample of smaller events are cross-checked against the archives of the leading Colombian newspaper, *El Tiempo*, to ensure that the data have been entered accurately, without double counting. The events are also cross-checked against databases from the National Police, Human Rights Watch, and Amnesty International.

CERAC also screens out events unrelated to the civil war (e.g., incidents of domestic violence or crime). The data set instead hones in on war-related actions carried out by politically motivated armed groups. Events are coded as either bilateral clashes between sides or unilateral attacks by any one side against another. The data separately categorize violence by the military, various paramilitary organizations, and several guerrilla groups (the largest being the FARC). Clashes occur among all three kinds of actors, though government versus paramilitary clashes are rare, as are clashes between guerrillas or between paramilitaries. The CERAC data set typically does not include kidnapping of individuals under event categories, since kidnapping is commonly used by criminal groups and tends to be underreported, which makes it difficult to measure comprehensively with accuracy (Restrepo et al., 2004). The exception is when kidnapping results directly from war-related actions, for example, alongside other actions by an armed group that result in the classification of the event as an attack.

We use CERAC data only after 1992 for the training sample since a consistent set of covariates is unavailable before then. We further examine the possibility of prediction in a longer panel by combining the CERAC series through 2005 with the Universidad del Rosario data set from 2006 to 2014. We present this in the appendix rather than the main analysis in part because there are some coding differences between the two data sets, which do not make them perfectly comparable. For example, the Universidad del Rosario data set includes additional political events, like the kidnapping of political actors, under the attack classification. In addition, it categorizes attacks and clashes differently for complex events involving the military. Nevertheless, predictive performance with the combined series is similar to the baseline using only CERAC (see appendix A.8).

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.

In Colombia, our indicator of any conflict occurs in about one-third of municipalities each year. "Hot spots" 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.

Covariates. As in Indonesia, we assemble a broad set of predictors from multiple sources: over 310 in all. First, we include detailed violence histories capturing incident types, actors, and outcomes. Beyond these violence histories, we incorporate the following predictor groups of related covariates: (a) population, (b) geography (e.g., terrain ruggedness), (c) remoteness (e.g., road presence), (d) distributional

[12] For the *Podes* variables, we use the two most recent measures of each conflict variable. Appendix C.1 details the number of lags included for each time-varying predictor.

measures (e.g., poverty and inequality), (e) historical traits (e.g., colonial population and infrastructure, following Acemoglu, Garcia-Jimeno, & Robinson, 2015), (f) demilitarized zone proximity, (g) Municipality Revenues and Expenditure, (h) electoral outcomes, (i) U.S. Military presence and spending (Dube & Naidu, 2015), (j) drug production and drug price shocks, (k) weather histories and shocks, and (l) commodity production and price shocks (as in Dube & Vargas, 2013).

As in Indonesia, we strive to use time-varying data when available. Electoral outcomes, municipality revenues and spending, drug production, U.S. military involvement, and commodity and weather shocks all vary annually. Our models include two lags of all violence predictors as well as lags of other time-varying predictors subject to availability. Summary statistics for each covariate, organized by predictor group, can be found in appendix C.3, while details on sources and variable construction can be found in appendix C.3.

C. *Comments on Data Quality*

It is worth discussing why we selected Colombia and Indonesia and why scholars regard their data as unusually high quality, especially the violence data. First, as we have already mentioned, it is rare to have such a large number of covariates systematically available in so many subnational units for such a long period. We know of few close comparisons in countries with histories of violence.

Second, both countries offer state-of-the-art subnational violence data. The SNPK, for example, was an explicit attempt to address shortcomings of earlier violence data sets within Indonesia (Barron et al., 2014). Both sources are meticulously documented and subject to various quality control measures (Restrepo et al., 2004; Barron et al., 2016). Each has been effectively deployed in the academic literature (Bazzi & Gudgeon, 2021; Dube & Vargas, 2013).

Third, the data draw on a large number of high-quality, local-language newspapers. This is rare, as most news-coded data sets draw mainly from international news services. In many low-income countries, the number and quality of local newspapers are extremely poor, and there is little news coverage in conflicted places. Hence, many events are never reported.

In fact, both countries' conflict data offer significantly more comprehensive coverage than other popular event-based data. The widely used Uppsala Conflict Data Program (UCDP) Georeferenced Event Data (GED; Sundberg & Melander, 2013), which relies on international media, misses over 99% of events in the SNPK and 61% of events in CERAC, many of which involve substantial casualties. We detail these comparisons with UCDP-GED in appendix C.1.3 for Indonesia and appendix C.2.3 for Colombia. Another popular source, the Armed Conflict Location and Event Data Project (ACLED) data, is available for a more limited range of years, beginning only in 2015 for Indonesia and 2019 for Colombia.

Nevertheless, we do not wish to overstate the quality of the data. The main violence data sets undoubtedly omit or misclassify some events or get the timing or location wrong, thus impeding accurate prediction. Coding procedures also change over time. In Colombia, for instance, the local sources and coding methods for the 2005–2014 data differ slightly from the original 1988–2005 data, which is why we only include the later period in the appendix. Our results are not particularly sensitive to the way in which we create a longer series by tying them together, but the issue is indicative of the challenges one can expect from quarter-century-long conflicts and data collection.

Another issue is that the data are typically aggregated to the level of a year or to a subdistrict or municipality. This introduces sources of measurement error in timing or location that could, in principle, reduce predictive power. It is possible that we need to reinvent the way that conflict scholars collect and code data in order to facilitate early warning. But that would be a huge undertaking. This illustrates the purpose of this paper: to evaluate how prediction works when applied to unusually rich existing data, before radically changing data approaches to data collection. We are careful to limit our conclusions to the specific kind of data that are currently available and one-year-ahead prediction exercises.

IV. Prediction Methods

A. *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 subannual frequency and because disaggregation would exacerbate a class imbalance problem (i.e., the fact that there are far more nonevents 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, and violence measures up to and including period t are the training outcomes. That is, violence in period t is matched to predictors in period $t - 1$, period $t - 1$ violence is matched to period $t - 2$ predictors, and so on, and we have $t - 1$ observations per location.
2. We use five-fold cross-validation to choose optimal tuning parameters specific to each machine learning algorithm (see section IVB). We choose tuning parameters to maximize out-of-sample area under the receiver operating characteristic, or ROC, curve (AUC), a metric detailed in section IVC. Five-fold cross validation simulates out-of-sample prediction. First, the data are randomly partitioned into five equal-sized subsamples. A model is then fit to four subsamples and used to predict violence in the fifth. This is repeated for each of the five 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 ten "optimal" tuning parameters, and then we take the average over these ten trials.[13]
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 in 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 $t + 1$; and so on. For Indonesia, this procedure generates seven predictions per algorithm. The first is trained on four years of data to forecast conflict in 2008, and the last is trained on ten years of data to forecast conflict in 2014. For Colombia, we generate eight predictions per algorithm. The first is trained on six years of data to forecast conflict in 1998, and the last is trained on thirteen years of data to forecast conflict in 2005.

B. *Machine Learning Algorithms*

We apply several machine learning methods. 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, Dorff, & Ward, 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 considers the full potential of the 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 0. This algorithm is the simplest of the four that we test, and arguably 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, being able to model complicated interactions between variables. Random forests are also relatively straightforward in terms of tuning parameter selection and have been used (albeit sparingly) in the conflict forecasting literature (Blair et al., 2017; Blair & Sambanis, 2020; Muchlinski et al., 2016).
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 data set, 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 & Schapire, 1999). Gradient boosted machines can improve on 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 0 and 1. 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, Tibshirani, & Friedman, 2001). The optimization problem is to choose appropriate weights in each linear combination.[14] 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 data sets, carefully chosen network architecture, and computing resources to achieve high performance.

Appendix B reports further details about hyperparameter 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, they

[13] In appendix B.2, we assess the extent to which multiple cross-validation rounds add stability to the hyperparameter choice. We conclude that for the most part, one single cross-validation exercise is sufficient to choose a good hyperparameter value. Nevertheless, this step is parallelizable and not costly in terms of computing time, so we use ten cross-validation runs throughout the paper in order to be conservative.

[14] 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 thirty principal components in Indonesia and twenty in Colombia. Predictive performance is not sensitive to these particular choices. In appendix B.3, we explore alternative neural network architectures.

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, 2008–2014)				
Any violent event	0.819	0.818	0.823	0.792	0.823
Five or more violent events	0.940	0.935	0.942	0.910	0.941
1 SD (or more) increase in events	0.866	0.817	0.852	0.825	0.860
	Colombia (attacks and clashes, 1998–2005)				
Any violent event	0.845	0.847	0.849	0.825	0.850
Five or more violent events	0.914	0.911	0.910	0.886	0.915
1 SD (or more) increase in events	0.802	0.787	0.796	0.741	0.801

Each model is trained on all available data preceding the out-of-sample prediction year. Training data start with 1991 data used to predict 1992 violence in Colombia and 2003 data for 2004 violence 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.

have been used (albeit infrequently) for purposes of forecasting in economics and political science, and 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 nonlinear 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.

C. *Performance Metrics*

To evaluate our models, we focus on the area under the ROC curve, known as the "area under the curve" (AUC). Other performance metrics such as the mean squared error, area under the precision-recall curve, and maximal accuracy and sensitivity are reported in appendix A.1. ROC curves plot the trade-off between true and false positives for a given model. The AUC captures the probability that a randomly 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 specifying a probability threshold above which we predict violence will occur. Selecting a specific threshold requires making a trade-off between accuracy, sensitivity (the proportion of incidents correctly predicted), and specificity (the proportion of nonincidents 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 we 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.[15]

V. Results

A. *Next Year's Violence Is Predictable*

Table 1 shows that all of the machine learning methods we test have strong predictive performance. For the ensemble average (EBMA), the AUC is above 0.82 for predicting one or more events, above 0.91 for five of more events, and above 0.80 for escalations of 1 or more standard deviations. In general, AUCs of 0.8 and above are considered very good, and AUCs of 0.9 and above are considered excellent.

To fix ideas, given a random pair of Indonesian subdistricts in which one location experiences five or more incidents and the other does not, there is a 0.941 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.

These models exceed the performance of other recent subnational conflict forecasting exercises. 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 from 1992 to 1995, 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 to 2012.[16] Gains in the range of 0.05 or

[15] Appendix A.1 reports the MSE, the precision-recall-AUC, 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 similar when we compare models using these alternative performance metrics rather than the AUC. Appendix section A.1 also reports alternative models that are trained to minimize MSE. Again, the results are qualitatively similar.

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

TABLE 2.—OUT-OF-SAMPLE PERFORMANCE VERSUS BENCHMARKS

	Baseline EBMA (1)	OLS (2)	Lagged Predictand (3)	Location FE Only (4)	Department/District FE Only (5)
	Indonesia (social conflict, 2008–2014)				
Any violent event	0.823	0.790	0.687	0.774	0.752
Five of more violent events	0.941	0.911	0.808	0.880	0.871
1 SD (or more) increase in events	0.860	0.759	0.527	0.694	0.776
	Colombia (attacks and clashes, 1998–2005)				
Any violent event	0.850	0.821	0.743	0.828	0.721
Five of more violent events	0.915	0.858	0.748	0.849	0.742
1 SD (or more) increase in events	0.801	0.744	0.521	0.683	0.712

Each model is trained on all data available preceding the out-of-sample prediction year. Training data start with 1991 data used to predict 1992 violence in Colombia and 2003 data for 2004 violence 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. Lagged predictand includes only a single lag of the violence indicator of interest. Fixed-effects models are estimated by OLS.

0.10 represent 10% to 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, which is notable given its relative simplicity. The tree-based models also perform well, with gradient boosted machines edging out random forests in all instances. The neural networks are generally the worst performers. They may be ill suited for this prediction task, which uses a relatively modest amount of data. They are also the most difficult to tune among our candidate algorithms. Finally, the ensemble (EBMA) is generally the best performer, giving us confidence in focusing on the ensemble as an indicator of the overall potential of these methods. That said, the gains from model averaging are not large relative to the individual top performers.

Machine learning outperforms simpler benchmarks. Table 2 compares our machine learning approaches to simpler benchmarks. Column 1 reproduces the results from our main EBMA specification in column 5 of table 1. Column 2 reports the performance of an OLS model using all predictor variables. Using our expansive data set of conflict predictors, OLS alone does a good job of predicting conflict. However, our ensemble machine learning model outperforms simple OLS. The gains are moderate in the case of predicting any incident and more significant for the rarer outcomes. Linear regression models appear to overfit the data: the generated predictions underperform those of the less flexible LASSO models in column 1 of table 1. Moreover, the linear regression is not able to model interactions between variables, which may explain the outperformance of random forests and gradient boosted machines.

Columns 3 to 5 consider alternative benchmarks that use fewer predictors. Column 3 reports the performance of a simple autoregressive model (AR1) in which positive cases in period t are predicted to remain positive cases in $t + 1$. Our prediction models outperform this simple AR1 model as well. Unsurprisingly, the lagged predictand model is particularly poor at predicting rapid increases in the number of conflicts in both Indonesia and Colombia.

Our models include enough time-invariant covariates, however, that in principle the machine learning performance could simply approximate fixed effects. Therefore, column 4 examines a simple OLS fixed-effects model. In each year, we regress all previous years' outcomes on fixed effects for each location. Then we take that fixed effect as our prediction for the following year. We see that in all cases, prediction using our full set of covariates outperforms the fixed-effects model. The improvements for any violent event are moderate, ranging from 0.025 to 0.05, which is roughly 5% to 10% of the difference between a random prediction and perfection. Improvements are larger when predicting hot spots and escalations. The fact that the relative outperformance of our baseline model is greatest in Indonesia and in cases where the dependent variable is rarer is intuitive from the perspective of estimator variance.[17] These fixed effects are noisily estimated, and our prediction algorithms are better able to estimate the relationship between fixed factors and conflict.

The specification in 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.[18] In this case, performance for the rarest event, a 1 standard deviation increase in violence, improves but still falls well short of our benchmark model.

Together, these results show that the full models perform better than several simpler alternatives. While performance varies across countries and outcomes, it is clear that there are gains to machine learning approaches.

[17] 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.

[18] There are 33 departments in our Colombia sample and 168 districts in our Indonesia sample. Therefore, the number of parameters drops considerably, as does the imprecision in their estimation.

TABLE 3.—OUT-OF-SAMPLE (ONE-YEAR-AHEAD) PERFORMANCE OF THE ENSEMBLE (EBMA) METHOD, VARYING PREDICTOR SETS

	Full Predictors (1)	All Past Violence Measures (2)	All Past Violence and Population (3)	Full Excl. Past Violence (4)	Time-Invariant Predictors (5)	Time-Varying Predictors (6)
	Indonesia (social conflict, 2008–2014)					
Any violent event	0.823	0.805	0.815	0.810	0.817	0.789
Five of more violent events	0.941	0.939	0.942	0.922	0.931	0.902
1 SD (or more) increase in events	0.860	0.845	0.856	0.847	0.856	0.815
	Colombia (attacks and clashes, 1998–2005)					
Any violent event	0.850	0.812	0.838	0.828	0.832	0.763
Five of more violent events	0.915	0.905	0.912	0.878	0.880	0.808
1 SD (or more) increase in events	0.801	0.765	0.788	0.780	0.781	0.748

Each model is trained on all available data preceding the out-of-sample prediction year. Training data start with 1991 data used to predict 1992 violence in Colombia and 2003 data for 2004 violence 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.

B. *We Predict Time-Invariant Risk over Space Rather than Time*

In this section, we clarify the nature and sources of predictability and conclude that our predictions mainly capture time-invariant risks of violence. Table 3 reports results. For purposes of comparison, column 1 reproduces the baseline EBMA results from column 5 of table 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. Instead, they comprise the number and severity of incidents (e.g., number of deaths, destruction of property) and the actors involved, and, in Indonesia, distinguish between each of the ten different violence categories.[19]

Columns 2 and 3 of table 3 consider models that use 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.

We find that prediction models using violence histories alone perform almost as well as our full model in column 1, suggesting that additional covariates yield, at most, modest improvements beyond these rich conflict histories. The addition of population in column 3 leads to little improvement. The model with violence histories also vastly outperforms a simple lagged dependent variable alone (column 3 of table 2). Thus, performance is not simply driven by the autoregressive properties of conflict but by the rich set of conflict measurements provided by our data sources. Appendix A.4 provides further evidence on the predictive returns to more detailed violence data.

Time-invariant predictors are most effective in our models. If detailed, past violence predicts future violence, do we need other predictors at all? We develop a number of tests for parsing the sources of predictability. Column 4 of table 3 shows results for a model that uses only 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 they 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 on their own. If remote areas are at risk for conflict in one year, then they continue to be at risk in the following years because they continue to be remote.

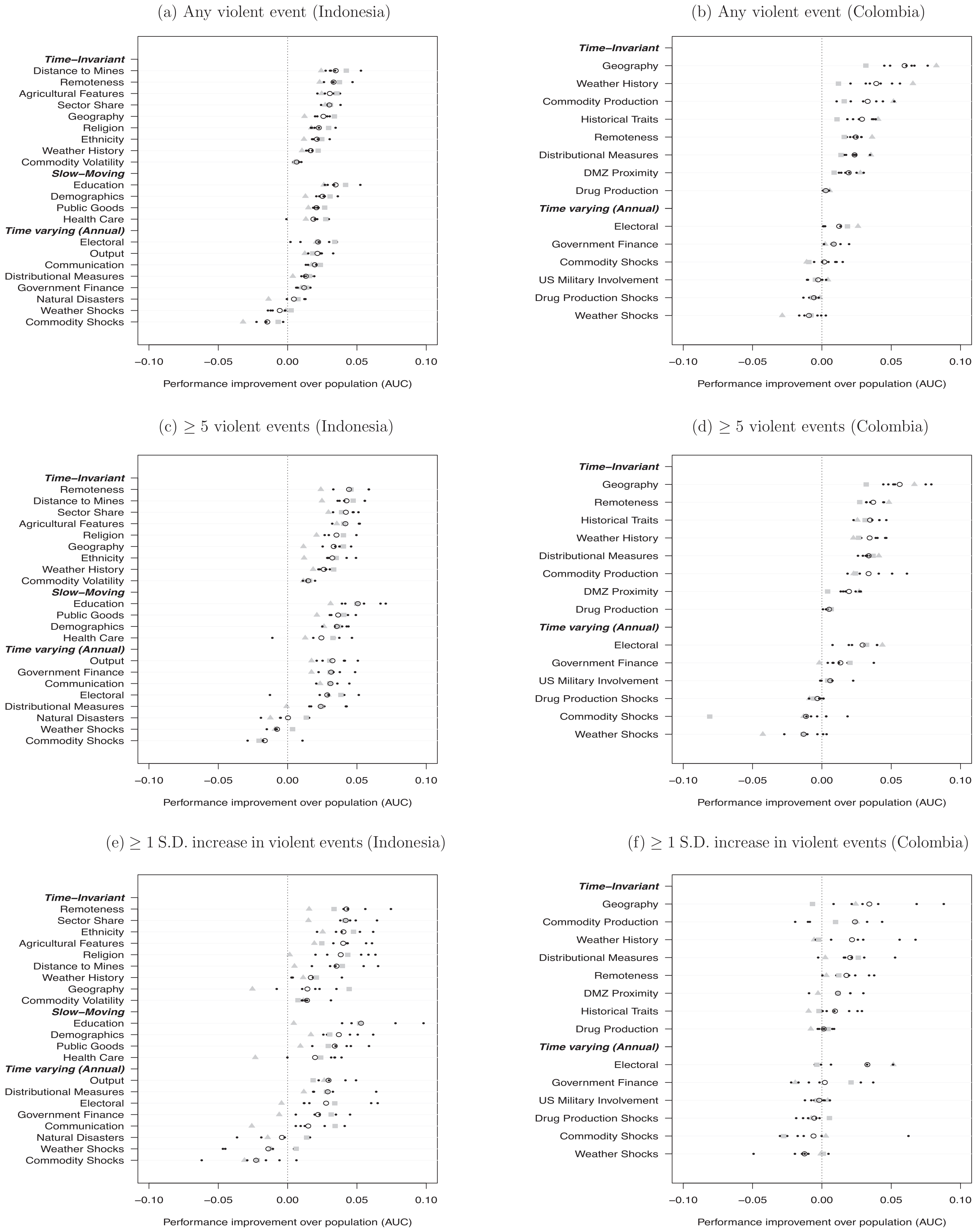
To examine the relative performance of time-varying and time-invariant predictors, we compare models composed entirely of one or the other. Column 5 uses only time-invariant traits to predict violence, and performance roughly matches or outperforms the model in column 4. Column 6 uses only time-varying predictors, and performance diminishes.[20] 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

[19] See appendix C.3 for the full list of past violence predictors in each country.

[20] This is true even for the 1 SD (or more) increased outcome. 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.

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)

Dots represent performance in individual years. A triangle (square) denotes the first (last) year of the sample. The hollow circle is average performance across the years. Appendix C.3 lists the variables in each predictor group.

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

Figure 1 plots the change in model performance from adding each subgroup. Each out-of-sample year is indicated with a small dot to give a sense of the range of performance changes. The first out-of-sample year is marked with a triangle, and the last is marked with a 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 and slow-moving predictors appear to add the most to predictive performance.[21] This is generally the case 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 road access, and geographic traits like terrain ruggedness, are generally the best predictors. Time-invariant measures of economic structure such as sectoral shares, agricultural features, or mineral presence are also important predictors.

In contrast, the covariates with the most year-to-year variation seldom improve predictive power, even when predicting escalations. This is particularly evident for natural disasters, commodity price shocks, drug price shocks, and weather shocks. In some instances, adding these predictors seems to decrease performance. However, the direction in which they affect performance varies, and the range of performance contributions over out-of-sample years is large, which limits our ability to definitively conclude how these predictors affect model performance.[22]

Some of the time-varying variables do improve performance. Output and communication variables also seem to generally help in Indonesia. However, it is worth noting that even for these time-varying measures, it could still be the case that the majority of their variation is cross-sectional. Indeed, the within over between variance for many of the variables in these predictor groups is below 1, indicating that much of the variation lies in the cross-section rather than within-location over time.[23] Electoral data generally appear to improve performance in both countries, though the effect varies for Colombia. 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. Overall, we conclude that time-invariant predictors contribute more consistently to our model's performance in predicting conflict over time.

Our models predict violence across locations. So far, several pieces of evidence point to the difficulty of predicting the specific timing of violence: the relatively poor performance of time-varying predictors, the interchangeability of histories of violence and time-invariant predictors, and our finding that histories of violence predict spikes in violence roughly as well as levels of violence.

In appendix A.3, we further show that our models perform especially poorly when predicting within-location, over-time variation in violence. In this exercise, we attempt to predict deviations in the number of violent incidents in each location from its historical mean. Performance is very poor. 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 4 reports these results. Strikingly, column 1 shows that overall performance when forecasting across locations is similar to performance when predicting ahead in time.

When forecasting across locations, 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, they improve it substantially.

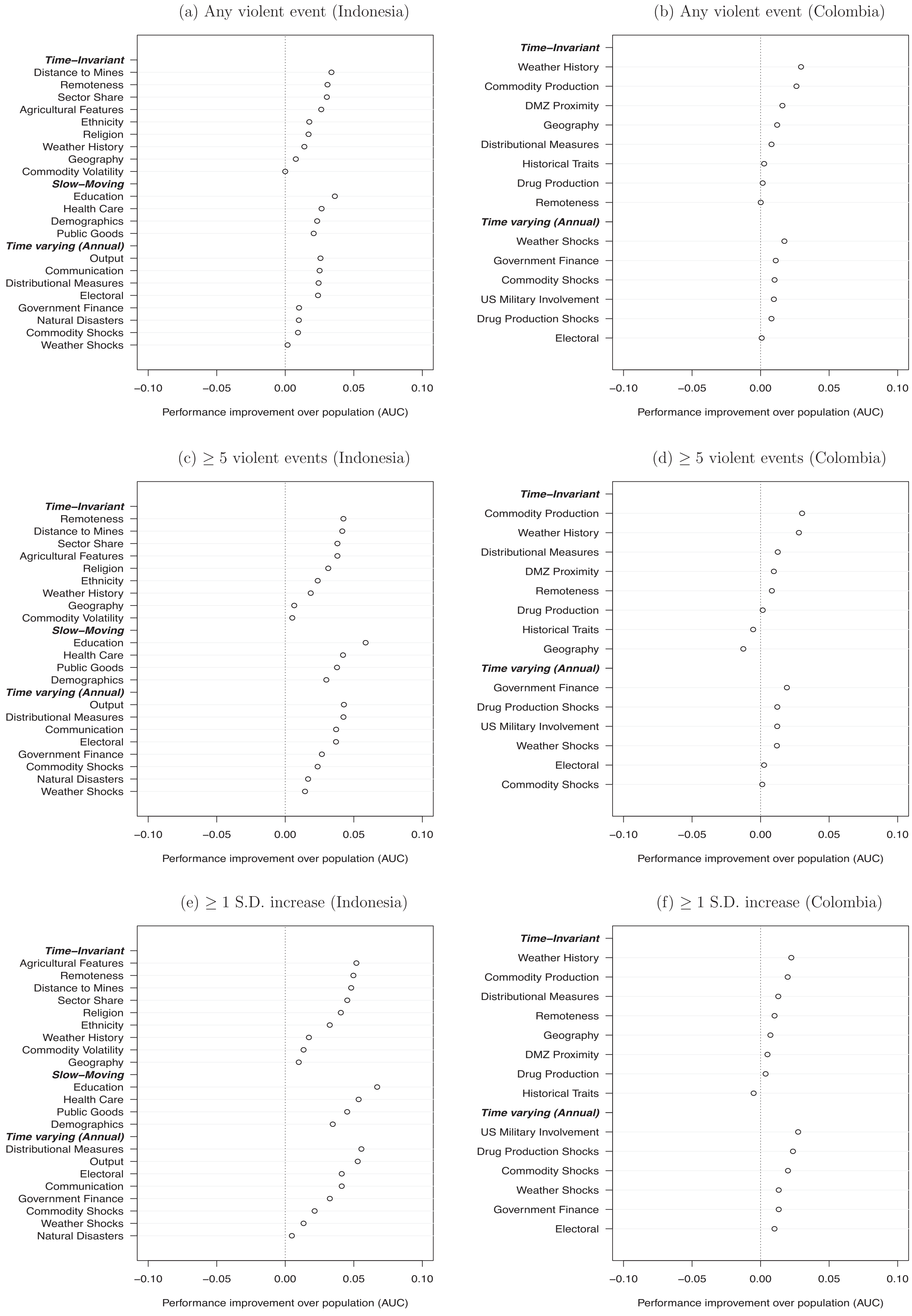
One notable difference between the across-location and year-ahead predictions is that the training set uses all years of available data in the across-location approach. The algorithms therefore observe the entire relevant distribution of weather, disasters, and commodity price fluctuations over the duration 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

[21] We classify variables as "slow-moving" in Indonesia if they vary triennially, originating from the *Podes* survey. These predictors measure characteristics that tend to vary more between locations than over time within location (see the summary statistics in appendix C.3, which show the within over between variance). Indeed, a prior draft used a purely time-invariant version of these predictor groups from just the 1999 *Podes*, and their predictive performance was extremely similar (Bazzi et al., 2019).

[22] 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 shocks have robust significant effects on conflict intensity and sometimes conflict onset. See, for example, Miguel et al. (2004), Bazzi and Blattman (2014), Berman and Couttenier (2015), Berman et al. (2017), 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 that the objective of prediction differs fundamentally from the objective of parameter estimation (Mullainathan & Spiess, 2017).

[23] In appendix A.7, we consider alternative groupings of time-varying predictors that shed light on this distinction between the spatial and temporal variation in time-varying predictors.

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)

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

TABLE 4.—PREDICTING ACROSS LOCATIONS

	Full Predictors (1)	All Past Violence Measures (2)	All Past Violence and Population (3)	Full Excl. Past Violence (4)
	Indonesia			
Any violent event	0.827	0.807	0.817	0.810
Five of more violent events	0.941	0.935	0.938	0.915
1 SD (or more) increase in events	0.864	0.844	0.850	0.855
	Colombia			
Any violent event	0.840	0.795	0.823	0.798
Five of more violent events	0.931	0.919	0.926	0.862
1 SD (or more) increase in events	0.834	0.786	0.812	0.796

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 data sets. Training data start with 1991 data used to predict 1992 violence in Colombia and 2003 data for 2004 violence 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.

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.

VI. Discussion

Using an unusually long and wide array of annual data, we show that local violence in Indonesia and Colombia appears to be predictable with relatively high levels of accuracy. 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. Machine learning approaches can help to identify these hot spots that would have remained more obscure with simpler forecasting methods. However, the residual variation—year-to-year changes in violence—remains difficult to forecast.

There are several possible explanations for this latter result. 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.

A variety of measurement problems may also limit model performance. Human-coded violence data, or data from news reports, is state-of-the-art in that it is often the best or only source of information available, but it is nonetheless prone to misclassification or errors of omission. The timing of violence could be a function of factors that are inherently hard to observe and measure, such as social grievances or the deterioration of communal trust. Both issues will bedevil prediction exercises of all kinds.

We might also lack a sufficiently long time series to be able to capture time-varying conflict risk, though several signs suggest otherwise. The limited predictive power of shocks in our over-time predictions may reflect a lack of common support in the training and testing samples. If so, then performance could improve with more years of data. Yet 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. If models need more than a quarter-century of rich conflict and risk factor data to yield better predictions, then the practical prospects for early warning are surely limited.

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 metadata, 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, Leopoldo Fergusson, James A. Robinson, Dario Romero, and Juan F. Vargas, "The Perils of High-Powered Incentives: Evidence from Colombia's False Positives," NBER working paper 22617 (September 2016). 10.3386/w22617

Acemoglu, Daron, Carlos Garcia-Jimeno, and James A. Robinson, "State Capacity and Economic Development: A Network Approach," *American Economic Review* 105:8 (2015), 2364–2409. 10.1257/aer.20140044

Acemoglu, Daron, James A. Robinson, and Rafael J. Santos, "The Monopoly of Violence: Evidence from Colombia," *Journal of the European Economic Association* 11:S1 (2013), 5–44. https://economics.mit.edu/files/10402. 10.1111/j.1542-4774.2012.01099.x

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

Barron, Patrick, Sana Jaffrey, and Ashutosh Varshney, "How Large Conflicts Subside: Evidence from Indonesia," Indonesia Social

Development Paper, World Bank (2014), https://asiafoundation.org/resources/pdfs/HowLargeConflictsSubside.pdf.

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

Barron, Patrick, Kai Kaiser, and Menno Pradhan, "Understanding Variations in Local Conflict: Evidence and Implications from Indonesia," *World Development* 37:3 (2009), 698–713. 10.1016/j.worlddev.2008.08.007

Bazzi, Samuel, and Christopher Blattman, "Economic Shocks and Conflict: Evidence from Commodity Prices," *American Economic Journal: Macroeconomics* 6:4 (2014), 1–38, 10.1257/mac.6.4.1

Bazzi, Samuel, Robert Blair, Christopher Blattman, Oeindrila Dube, Matthew Gudgeon, and Richard M. Peck, "The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia," NBER working paper 25980 (2019).

Bazzi, Samuel, and Matthew Gudgeon, "The Political Boundaries of Ethnic Divisions," *American Economic Journal: Applied Economics* 13:1 (2021), 235–266. 10.1257/app.20190309

Beck, Nathaniel, Gary King, and Langche Zeng, "Improving Quantitative Studies of International Conflict: A Conjecture," *American Political Science Review* 94:1 (2000), 21–35. 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* 32:1 (2016), 98–111, 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," this REVIEW 97:4 (2015), 758–776. 10.1162/REST_a_00521, PubMed: 25948523

Berman, Nicolas, Mathieu Couttenier, Dominic Rohner, and Mathias Thoenig, "This Mine Is Mine! How Minerals Fuel Conflicts in Africa," *American Economic Review* 107:6 (2017), 1564–1610, 10.1257/aer.20150774

Blair, Robert A., and Nicholas Sambanis, "Forecasting Civil Wars: Theory and Structure in an Age of 'Big Data' and Machine Learning," *Journal of Conflict Resolution* 64:10 (2020), 1885–1915. 10.1177/0022002720918923

Blair, Robert A., Christopher Blattman, and Alexandra Hartman, "Predicting Local Violence: Evidence from a Panel Survey in Liberia," *Journal of Peace Research* 54:2 (2017), 298–312. 10.1177/0022343316684009

Blattman, Christopher, and Edward Miguel, "Civil War," *Journal of Economic Literature* 48:1 (2010), 3–57, 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* 28:1 (2011), 41–64. 10.1177/0738894210388125

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

Cederman, Lars-Erik, and Nils B. Weidmann, "Predicting Armed Conflict: Time to Adjust Our Expectations?" *Science* 355:6324 (2017), 474–476. 10.1126/science.aal4483, PubMed: 28154047

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), 10.2139/ssrn.23866677

Colaresi, Michael, Håvard Hegre, and Jonas Nordkvelle, "Early ViEWS: A Prototype Disaggregated, Open-Source Violence Early-Warning System," paper 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 Suresh Naidu, "Bases, Bullets, and Ballots: The Effect of US Military Aid on Political Conflict in Colombia," *Journal of Politics* 77:1 (2015), 249–267, https://www.journals.uchicago.edu/doi/10.1086/679021.

Dube, Oeindrila, and Juan F. Vargas, "Commodity Price Shocks and Civil Conflict: Evidence from Colombia," *Review of Economic Studies* 80:4 (2013), 1384–1421. 10.1093/restud/rdt009

Fearon, James D., "Rationalist Explanations for War," *International Organization* 49:3 (1995), 379–414, https://www.jstor.org/stable/2706903. 10.1017/S0020818300033324

Freund, Yoav, and Robert E. Schapire, "A Short Introduction to Boosting" (pp. 1401–1406), in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence* (San Mateo, CA: Morgan Kaufmann, 1999), https://dl.acm.org/citation.cfm?id=1624417.

Gartzke, Erik, "War Is in the Error Term," *International Organization* 53:3 (1999), 567–587. 10.1162/002081899550995, PubMed: 10692747

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* 50:1 (2013), 17–31. 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* 54:1 (2010), 190–208. 10.1111/j.1540-5907.2009.00426.x

Guha, Rishab, and Serena Ng, "A Machine Learning Analysis of Seasonal and Cyclical Sales in Weekly Scanner Data," NBER technical report (2019).

Gurr, Ted Robert, and Mark Lichbach, "Forecasting Internal Conflict," *Comparative Political Studies* 19:1 (1986), 3–38. 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* 97:1 (2003), 57–73, https://www.jstor.org/stable/3118221. 10.1017/S0003055403000522

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning* (New York: Springer, 2001).

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* 11:5 (2016), 054002, 10.1088/1748-9326/11/5/054002

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

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

Indra, Suahasil Nazara, Djoni Hartono, and Sudarno Sumarto, "Roles of Income Polarization, Income Inequality and Ethnic Fractionalization in Social Conflicts: An Empirical Study of Indonesian Provinces, 2002–2012," *Asian Economic Journal* 33:2 (2019), 165–190. 10.1111/asej.12179

Jasny, Barbara R., and Richard Stone, "Prediction and Its Limits," *Science* 35:6324 (2017), 468–469. 10.1126/science.355.6324.468

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti, "Economic Shocks and Civil Conflict: An Instrumental Variables Approach," *Journal of Political Economy* 112:4 (2004), 725–753, https://www.journals.uchicago.edu/doi/10.1086/421174. 10.1086/421174

Montgomery, Jacob, Florian Hollenbach, and Michael Ward, "Improving Predictions Using Ensemble Bayesian Model Averaging," *Political Analysis* 20 (2012), 271–291. 10.1093/pan/mps002

Muchlinski, David, David Siroky, Jingrui He, and Matthew Kocher, "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data," *Political Analysis* 24:1 (2016), 87–103. 10.1093/pan/mpv024

Mueller, Hannes, and Christopher Rauh, "Reading between the Lines: Prediction of Political Violence Using Newspaper Text," *American Political Science Review* 108:1 (2017), 1–18, 10.1017/S0003055417000570

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

Perry, Chris, "Machine Learning and Conflict Prediction: A Use Case," *Stability: International Journal of Security and Development* 2:3 (2013), 56, 10.5334/sta.cr

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

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

Schrodt, Philip A., "Forecasting Conflict in the Balkans Using Hidden Markov Models" (pp. 161–184), in Robert Trappl, ed., *Programming for Peace: Advances in Group Decision and Negotiation*, vol. 2 (Berlin: Springer, 2006), 10.1214/10-STS330

Shmueli, Galit, "To Explain or to Predict?" *Statist. Sci.* 25:3 (2010), 289–310, 10.1214/10-STS330

Sundberg, R., and E. Melander, "Introducing the UCDP Georeferenced Event Dataset," *Journal of Peace Research* 50:4 (2013), 523–532. 10.1177/0022343313484347

Tadjoeddin, Zulfan, *Explaining Collective Violence in Contemporary Indonesia: From Conflict to Cooperation* (New York: Springer, 2014).

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

Ward, Michael, Nils Metternich, Cassy Dorff, Max Gallop, Florian Hollenbach, Anna Schultz, and Simon Weschle, "Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction," *International Studies Review* 15:4 (2013), 473–490. 10.1111/misr.12072

Weidmann, Nils, and Michael Ward, "Predicting Conflict in Space and Time," *Journal of Conflict Resolution* 54:6 (2010), 883–901. 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, 2015–2065: Using Climate-Sensitive Models," *Journal of Peace Research* 54:2 (2017), 175–192. 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/.