What can prediction teach us about violence? Machine learning applications in Indonesia and Colombia*

Samuel Bazzi
Boston University

Robert Blair
Brown University

Christopher Blattman
University of Chicago and NBER

Oeindrila Dube
University of Chicago and NBER

Matthew Gudgeon
Boston University

Richard Peck
Northwestern University

December 15, 2016

Abstract

A large and growing literature aims to identify causal determinants of violent conflict. However, less is known about the predictability of conflict, and much less about what forecasting can teach us about violence. We deploy six machine learning methods to forecast conflict in Indonesia with new geospatial data on violent events at both the subdistrict and village level, beginning in 2000. Our preferred model, an ensemble of the other five, improves on prior attempts to predict local violence in other settings. Combining data from numerous administrative and survey-based sources, we find that lagged conflict, demographic characteristics such as the age distribution, education levels and infrastructure, and natural disasters are the most important predictors of future conflict. However, economic shocks associated with commodity price and weather volatility have limited predictive power. In ongoing work (to be incorporated into a new draft in February 2017), we predict guerrilla and paramilitary attacks in Colombia from 1990–2005 using a similarly large municipality-level panel dataset. In both contexts, the exercise delivers several important findings about local-level violence. First, violence is persistent but cannot be characterized by a simple autoregressive process. Second, violence is highly predictable with a small subset of readily observable variables. Finally, the conflict literature may be focused on a narrow set of economic shocks that are causally well-identified but of limited value in a forecasting sense.

JEL Classifications: C53, C55, D74, P16

Keywords: Conflict, Machine Learning, Prediction

PRELIMINARY AND INCOMPLETE—NOT FOR CITATION

*We thank seminar participants at NEUDC and LACEA 2016. We thank Jan Pierskalla, Audrey Sacks and the Indonesian National Violence Monitoring System for sharing data.

1 Introduction

How predictable and regular is violence, and what can forecasts tell us about the nature and causes of of conflict? Predicting violence has an obvious practical value, helping states prevent conflict before it happens or at least respond more quickly. This is one of the main motives behind a wealth of research that aims to forecast violent conflict at the national level (e.g., Muchlinski et al., 2016; Ward et al., 2013) with recent work extending these methods to the subnational level (Blair et al., 2015; Yonamine, 2014).

But social scientists are also interested in explaining violence rather than simply predicting it. Thirty years of empirical research have identified a number of robust causal determinants of conflict. Economic shocks associated with weather (Burke et al., 2015) and commodity price shocks (Bazzi and Blattman, 2014) feature prominently in this literature. Even within countries at war, climate and trade shocks seem to cause more intense and frequent attacks (Dube and Vargas, 2013; Berman and Couttenier, 2013). It's unclear, however, how much conflict these shocks explain. Conflict regressions have notoriously small coefficients of determination, or R^2 statistics, and these measures of fit tend to move little when causally-identified shocks are added to the model. If shocks have little material influence on the probability of conflict, researchers may want to shift their focus to other factors and search for causal identification there. This is where forecasting methods may have something to add to the science of conflict: they provide clearer and more accurate measures of predictive value than the coefficients of determination, and may focus our attention on new proximate predictors.

To investigate, this paper deploys a suite of established machine learning methods to generate year-ahead predictions of local violence in Indonesia. Indonesia is a useful testing ground because of the quantity and quality of relatively high-frequency, geospatial data on both violence and potential predictors of violence. These data span the years 2002 to 2014, with more than 400 predictors covering over 2000 subdistricts, with diverse measures of violence at the subdistrict level. Since the fall of authoritarian President Suharto in 1998, and the ensuing wave of democratization and decentralization, local level conflict is tragically pervasive and diverse, ranging from armed conflict to election violence to sectarian group clashes. Over the time period we examine, ethno-communal violence and low-level religious violence occur, and we see an increase in resource, particularly land related, disputes as well as violence relating to elections and governance. Each year in our dataset, nearly half of subdistricts experience a violent event, and about ten percent experience five or more violent conflicts. At the same time, there is a sufficiently high capacity in the government, press, civil society and academic community that there are a number of reasonably accurate datasets on violent events. The available predictors, moreover, range from political conditions to economic organization and demographics, and include a wide range of plausibly exogenous climatic and economic shocks. The available data are far more disaggregated, detailed, rich, and micro-level than typically available for forecasts. We use it to ask several questions, including: how predictable or idiosyncratic is violence? What information is needed to predict violence? And do certain variable groups, such as ethnic and religious diversity or economic shocks, matter for violence prediction?

The current version of this paper limits itself to violence data and predictors from 2011

and earlier, holding the 201214 data as a test set. We call the first period the training data and the 2010-14 data the test set, and for the time being hold the test set aside, unexamined. This allows us to maximize model fit on the training set before making a true out-of-sample test on the test set. Cloud-based computing power and machine learning procedures allow researchers to train and fit such complex models, and thus make predictions in a principled manner. Through repeated fitting and testing using random partitions of the training dataset, algorithms are able to find complicated functions of predictorsinteractions and other nonlinear combinationsthat outperform standard regression models that have featured prominently in the forecasting literature.

Our first result is that subdistrict-level violence, especially intense violence, is reasonably predictable. Our leading prediction approach, an ensemble method that averages five machine learning methods, correctly predicts the occurrence of any violent event in three quarters of cases, and correctly predicts over 90 percent of subdistricts with high concentrations of violence. Our preferred metric is the area under the Receiver Operating Characteristic (ROC) curve, a measure of how the true positive and false positive rates vary as forecasters vary model sensitivity. This measure varies between 0.5 (performance no better than chance) to 1 (perfect prediction). For our ensemble model, the area under the curve (or AUC) is 0.81 or 0.94, depending on the measure of violence that we choose to predict. Performance is highest for predicting subdistricts that experience five or more violent events. This outperforms previous attempts to predict subnational violence (e.g., Blair et al., 2015). Among the five component methods, the gradient-boosted machine performs best, itself an ensemble method composed of decision trees. At the same time, the simple LASSO, a constrained logit regression that is relatively familiar to social scientists, comes a close second.

Our second main result is that violence is highly persistent, so that past violence is predictive of current violence. Slow-moving demographic characteristics such as population density are also highly predictive of future conflict. But models that predict current violence using lagged violence alone perform remarkably well, such that the additional covariates do not add much predictive power. This is somewhat surprising. Given that the administrative data that yields these covariates are difficult to collect and do not exist in many parts of the world, it suggests that the practical task of simply predicting violence to forestall it could rely on tracking violent events alone.

Our third result is that dropping lagged violence from the predictive models does not reduce the quality of predictions by much. This suggests that much of the information contained in lagged violence is representative of substantive, observable, measured qualities of villages and subdistricts. When violence is excluded, important predictors include measures or religious and ethnic diversity, population density, and natural disasters..

Our fourth main result is that dropping both lagged violence and time-varying economic variables from the prediction models (including shocks) does little to hinder predictive power. Relatively time-invariant characteristics of subdistricts do the majority of work, and achieving AUCs of roughly .8 indicators of any event, and .9 for indicators of five or more events. Indeed, although natural disasters such as earthquakes and floods predict future conflict, common economic shocks (such as commodity price, temperature, and rainfall changes) add little

predictive power to the forecasting models. Dropping them has almost no effect on predictive power.

One final insight comes from the observation that while simply dropping shocks does little to predictive power, partialling out all of the variation and information that is contained in these shocks does affect predictions a great deal. That is, if we partial out economic shocks from all other predictors, and use the residuals of these other predictors to forecast violent events, our predictiive power declines significantly. One possible explanation is that most of the effects of time-varying shocks on conflict comes not from the temporal shock itself, but how persistent volatility shapes other more time-invariant social and economic characteristics over the long run. Economic volatility may engender political instability and violence for structural reasons rather than highly temporal ones such as sudden changes in the relative wages of conflict.

This project is a work in progress. We expect performance and the quality of our conclusions to improve with the introduction of the test data. Furthermore, we are in the process of adding analysis from Colombia, in which we use the same methods to predict municipality-level violence from 1990-2005. To do this, we have undertaken a similar data collection effort, merging together about a dozen data sources to date. The initial results confirm the conclusions from Indonesia: violence is persistent and predictable, and the best predictors of the incidence of violence are time-invariant municipality traits.

In addition to extending this work to Colombia, in both contexts we are currently working on better forecasts of the timing of changes in violence levels, rather than predictions of the incidence of violence, which is the focus of the current paper. These results will be included in a new draft to be completed in February 2017.

The remainder of this paper is organized as follows: Section 2 provides a brief description of Indonesia, and its history of conflict, Section 3 gives an overview of our data sources and variable construction, Section 4 explains the prediction algorithms, Section 5 reports the results, and Section 6 discusses next steps and concludes.

2 Setting: Indonesia

Indonesia is an archipelago of over 6,000 islands inhabited by 255 million people. Indonesia is remarkably diverse with over 1,000 ethnic groups speaking more than 400 languages. Although the vast majority of Indonesia's residents are Sunni Muslim, a substantial Christian minority make up 10% of the population. Around half of the population lives in urban or semi-urban areas today. Commodity production and services dominate the economy, but manufacturing is important in a few of the major cities across the country.

The recent history of Indonesian conflict can roughly be classified into two periods: (i) collective violence during the democratic transition and initial decentralization reforms (19982003), and (ii) routine local violence from 2004present. Violence in the first period included anti-Chinese riots, large-scale interethnic and interreligious violence (e.g., in Maluku), separatist conflicts between the central government in Aceh and Papua, and terrorist acts by fundamentalist Islamic groups .

Since 2004, large-scale conflict has transitioned into more sporadic violence, characterized by fewer fatalities, albeit still concentrated in the same regions with a history of violence. Ethno-communal violence and low-level religious violence remain important, and we see an increase in resource, particularly land related, disputes as well as violence relating to elections and governance (Barron et al., 2009, 2014). Bazzi and Gudgeon (2016) provide further background and empirical evidence on the causes of political violence during this period.

3 Data

3.1 Unit of Analysis

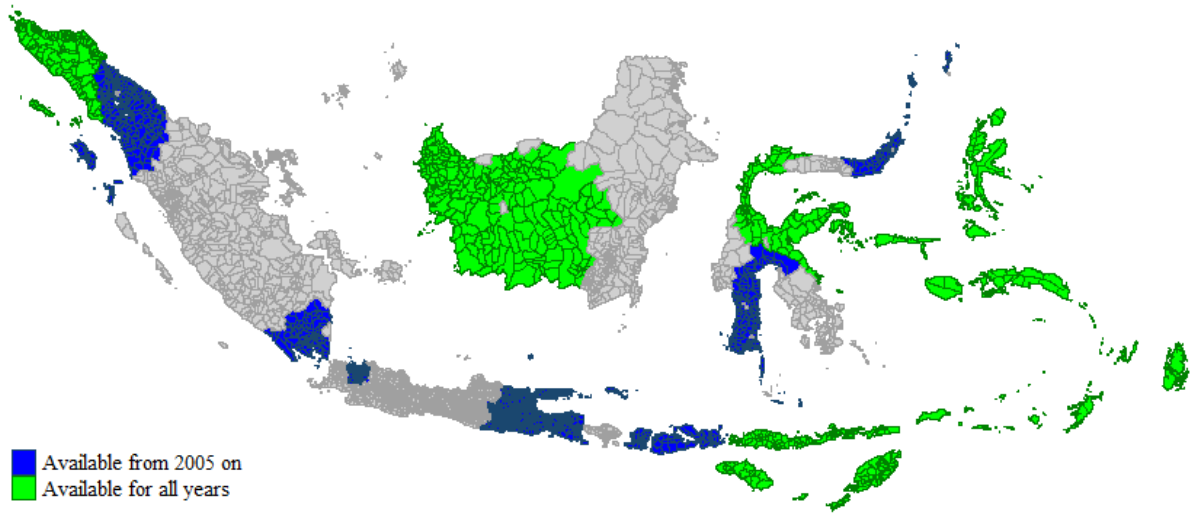
Indonesia has four main tiers of government. The largest tier is the province, of which there were 34 by 2014. Provinces are divided into districts known as *kabupaten* and *kota*. In June 2014, there were 514 districts.[1] Districts are the key autonomous administrative units, responsible for providing the bulk of governmental services. Districts are in turn divided into 7,094 subdistricts (*kecamatan*), the unit of analysis in this paper, which are further subdivided into into more than 80,000 villages, the smallest unit of government. The median subdistrict in our sample has a population of about 22,000.

3.2 New Geospatial Conflict Data (SNPK)

Our dependent variable is drawn from new monthly data on conflict from the Indonesian National Violence Monitoring System. Hereafter, we refer to the data by its Indonesian acronym, SNPK (*Sistem Nasional Pemantauan Kekerasan*). Like other geospatial conflict databases such as the Armed Conflict Location & Event Data (see e.g., Michalopoulos and Papaioannou, forthcoming), the SNPK data is based on reports of violence in local media, in this case over 120 newspapers. Multiple regional newspapers are collected for each province to ensure district coverage. 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. Thus, the data is not formally representative of Indonesia, but by 2005 it spans all major island groups and covers a majority of the Indonesian population. Table 1 displays the coverage. Green subdistricts are available from 1998 onward, and blue subdistricts are available beginning in 2005.

[1] *Kabupaten* tend to be more rural than *kota*. We treat them similarly in the empirical analysis and refer to both as districts unless otherwise specified.

Figure 1: Data Coverage



Note that we have merged subdistricts to reflect their 2000 boundaries.

In most cases, teams recorded conflict locations at the 2011 subdistrict level. Coders recorded the subdistrict whenever the newspaper report included sufficient location information to determine it. Specifically, 82% of incidents for the areas that are in the data from 1998-2014 have subdistrict codes. [2] Hence, we drop about 16% of incidents that are not given a subdistrict code. These are more likely to come from the early years than the later years, but not differential across crime and non-crime.[3]

Data collection occurred in 2008, when teams first went to collect the newspaper archives. Papers are either classified at the provincial level or sub-provincial level, with the sub-provincial level being a collections of districts or urban zones. We include data on the number of newspapers used in each province-year.[4]

Coders read articles and then use a standardized template to code the incident based on the underlying trigger. The incidents are first coded as domestic violence, violent crime, violence during law enforcement, or conflict. Within conflict, the coders further sort into identity, elections/appointments, governance, resource violence, popular justice, separatist, and other (could not be classified). Table 1 reports definitions for each of these categories.

In our baseline specification, we predict an indicator of any conflict, excluding domestic violence and crimes.[5] We group the remaining categories of violence to guard against possible miscoding of ambiguous events or events that span multiple classifications. For example a conflict over land-use may involve individuals or groups of different religions. Nevertheless,

[2]This number increases to 84% if only look at 2005-onwards. 87% of the incidents recorded in districts that enter in 2005 have subdistrict codes.

This is increasing over time reaching 90% for the post-2010 years.

[3]In ongoing work, we explore the prevalence and account for any systematic missingness.

[4]We have data on number of papers used in each month at the province or sub-province level. For now we aggregate to the province level for simplicity. For the annual panel, this is further aggregated up to the year.

[5]We exclude from domestic violence and crime, because we regard it as simpler to distinguish from the other forms of violence, and domestic violence is likely to have different causes than the other categories. Crime accounts for a large portion of the incidents therefore overwhelms trends in other forms of violence, so we exclude it. We have done the analysis including domestic violence and crime, and it does not substantively change the results

Table 1: Violence Definitions

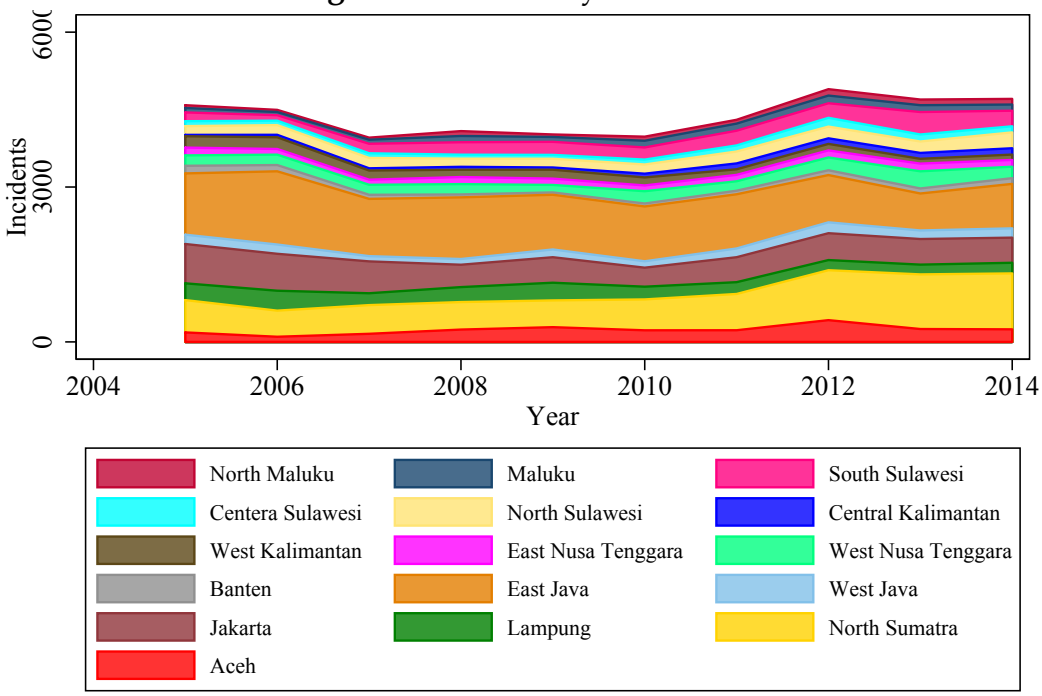
Violent Crime	Criminal violence not triggered by prior dispute or directed towards specific targets.
Domestic Violence	Physical violence perpetrated by family member(s) against other family member(s) living under one roof/same house including against domestic workers and violence between cohabitating 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.

we believe these subcategories to be interesting, and plan to investigate disaggregations of this variable in ongoing work.

For the initial analysis we restrict the period to 2005 to 2009. We are holding the last five years of data for a true out-of-sample test in the future in order to add validity to our hypothesis testing and guard ourselves against over-fitting, since our current results involved many rounds of training and testing on the same data. We choose the starting data of our panel based on availability of other predictor variables. Thus, in order to have a balanced panel with several lagged measurements of violence for each observation, we further restrict the analysis to the subdistricts with data available from 2005 onward (subdistricts with data before 2005 are colored green in Figure 1. Those with data from 2005 onward are colored blue.).

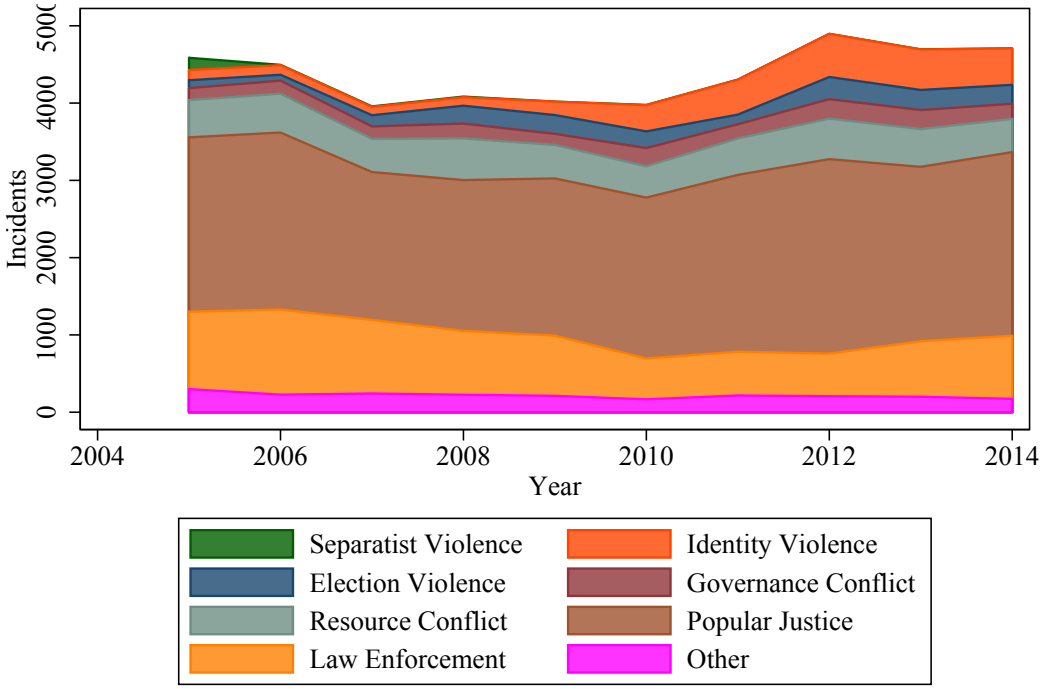
In Figure 2, we show the trends in violence by region over the period (excluding crime and domestic violence).

Figure 2: Violence by Province



We also break these trends down by violence category in Figure 3.

Figure 3: Violence by Category



In addition to any non-crime violence, we also consider an indicator for at least five non-crime events as an alternative prediction variable meant to capture higher intensity episodes of conflict. We choose to predict indicators for violence, rather than conflict counts, in order to simplify the interpretation of performance the models either correctly predict the incidence of violence or they do not.

3.3 Potensi Desa (Podes)

We have a second set of conflict indicators from the Village Potential Statistics (Podes). The Podes survey occurs at the village level roughly every three years, and village leaders are asked about a long list of village characteristics, which we also employ as predictors, including recent violent events. Since these conflict indicators are based on surveys of town leaders, there may be some downward bias in this measure, as town leaders may be reluctant to report violent events. For the analysis in this draft, we use the Podes data as predictors.[6] Since the Podes measurement is at the village level, we aggregate it to the subdistrict level and construct an indicator for each subdistrict of any violent event in its constituent villages. This way, the data will be directly comparable to the Podes measurement. In addition, Podes reports the the numbers of security personnel and posts in each town, which we also use as predictors of violence.

[6]We plan to predict village level violence using the Podes data in future work.

Table 2: Violence Rates

	SNPK: Any Incident	SNPK:≥ 5 Incidents	Podes: Any Conflict
2005	0.540	0.130	0.221
2006	0.537	0.122	
2007	0.509	0.114	
2008	0.550	0.125	0.281
2009	0.527	0.120	
2010	0.536	0.119	
2011	0.531	0.127	0.306

Notes: Podes data are only collected in 2005, 2008, and 2011.

Table 2 reports the incidence rates for the two measures. The correlation between the two measures is low0.110. For now, we use SNPK as our preferred measure given the richness and high frequency nature of the data.

3.4 Additional Predictors

Indonesia is a data-rich environment, and we have assembled a large set of subdistrict level characteristics from ten different data sources. A detailed reporting of the data sources and the variable construction can be found in the Data Appendix. The long list of potential predictors capture the myriad factors associated with conflict in prior literature. The machine learning approach allows us to remain agnostic ex ante about whether and how each measure should enter the specification. In addition to lagged violence measures from SNPK and Podes, we have:

Religion. From the 2000 census, we have data on the share of the population that belong to major religionsIslam, Catholicism, Protestantism, Buddhism, Confucianism, and Hinduism. Using these variables we construct an index of religious fractionalization as well as group shares.

Ethnicity. Similarly, we have data on indigenous ethnicities as well as ethnic Arab and Chinese, and we construct indices of ethnic polarization and fractionalization.

Education. Also from the 2000 census, we have data on educational attainment, and Podes reports the number of educational institutions per capita.

Remoteness. Using GIS data we have distance to provincial capitals, and distance to district capitals comes from Podes. We also include measures of transportation infrastructure such as road access and presence of major bridges or seaports from Podes and the 2000 census.

Demographics. We have annual population from the National Socioeconomic Survey (Susenas) and gender and age breakdowns from the 2000 census.

Health Infrastructure. We have data on the presence of hospitals and medical clinics from the Podes.

Natural Disasters. Podes also provides village-level indicators of mudslides, floods, earthquakes, and fires. The Regional Disaster Management agency (BPBD) also provides a a rich set of historical disaster data, including floods, earthquakes, fires, tsunamis, and volcano eruptions. These data are available annually, and we include both counts of disasters as well as death tolls from each disaster type.

Elections. We use the 1999 and 2004 national parliamentary election data by district as well as 1999 local parliamentary election by subdistrict to capture the share of votes obtained by mainstream parties as well as extremist parties. We also construct an index of political polarization, and include an indicator for years in which direct local elections take place (every five years, staggered across districts).

Agricultural Infrastructure. From Podes and the 2000 census, we have data on the quality of irrigation and the presence of rice mills.

Finance and Trade Institutions. Podes provides indicators of the presence of markets and banking institutions at the village level. We construct measures of the share of contituent villages with each of these institutions for each subdistrict.

Government Finance. We include district revenues from the World Bank's Database for Policy and Economic Research (DAPOER) which are available on an annual basis. Revenues are disaggregate by sourcelocal taxes, revenue sharing with other districts, and income from natural resources. The 2000 census also contains government finance information, including both revenues and expenditures by category.

Economic Structure. The 2002 Agricultural Census provides the share of agricultural production in cash crops and food crops. The 2000 census provides the share of the population employed in sectors such as agriculture, industry, services, and transportation. SNL provides mine locations, and we construct a measure of the distance to the nearest mine for major minerals such as bauxite, coal, copper, gold, iron, silver, nickel, tin, and zinc.

Output. Susenas provides annual measures of household consumption expenditures. We also constuct a proxy for annual agricultural gdp using production weights from the 2002 Agricultural Census and commodity price movements. DAPOER provides district level output measures. Finally, we use nighttime light intensity as a proxy for output at the subdistrict level.

Inequality. Using household survey data from Susenas, we construct measures of consumption inequality. Susenas also allows us to construct wage inequality (difference between the 75th and 25th percentile) for rural workers, urban workers, agricultural workers, and non-agricultural workers.

Commodity Shocks. We use 2002 agricultural production shares to weight log changes in commodity prices for cash crops, major cash crops, and food crops. We also construct historical volatility for each of these commodity price shocks. Using the mine location data, we construct mineral price shocks by multiplying log changes in world mineral prices by the inverse distance to the nearest mine producing that mineral.

Weather Shocks. Lastly, we compute rainfall deviations from long-run district averages, as well as rainfall and temperature deviations from long-run subdistrict averages. Again, we compute historical volatility for these shocks as well as means.

4 Prediction Methods

4.1 Models

We train and test five different machine learning algorithms, as well as an ensemble model that combines the predictions of the five models into a singe prediction. We chose our models to span the range of machine learning approaches and offer a mix of assumptions (or lack thereof) about functional form. LASSO and neural networks have been established tools of the trade since the 1990's, while Random Forests and the closely related Adaptive Boosting, as well as Support Vector Machines have increased in popularity in the past 10 years. Ensemble methods are also becoming increasingly popular (Montgomery et al., 2012).

LASSO

The LASSO regression, first introduced in Tibshirani (1994), is closest to a logistic regression model. As in logistic regression, LASSO posits that the probability that the natural logarithm of the odds ratio of violence is a linear combination of the predictor variables. Unlike logistic regression, LASSO penalizes large weights in the linear combination. Thus the objective is to minimize the sum of the standard logistic loss function and the sum of the absolute values of the estimated coefficients. This penalty biases coefficients toward zero, and because the penalty takes the form of a hypercube centered at zero in the coefficient space, many coefficients will be estimated to be exactly zero (along the edge of the hypercube). Hence, we can think of LASSO as a more parsimonious logit, in which only the most important predictors have non-zero coefficents. This algorithm is the simplest of the five, and is least susceptible to overfitting. However, it is unable to uncover more complex relationships between the predictors and future violence.

Random Forests

Random Forests are composed of many decision trees. Each decision tree is a sequence of rules that splits the sample into subsets (called leaves) based on certain variable cutoffs. The prediction for each leaf is the mean outcome for the trained observations on that leaf, and trees are fit so as to minimize mean squared error.

Each tree in a Random Forest 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.

Adaptive Boosting (AdaBoost)

This is a variant of Random Forests, in which trees are not fit 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. Therefore, later trees are fit with a larger weight on observations that previous trees found difficult to predict. In this manner, the addition of each tree offers a slight improvement of the model (Freund and Schapire, 1999).

Support Vector Machines

Support Vector Machines fit a hyperplane onto a higher dimensional mapping of the predictor space that best separates the outcomes into classes ("violence" and "no violence"). This boundary is then defined solely by the observations that are near to it (the "support vectors"). While the boundary is linear in the higher dimensional mapping of the predictor space, it can be highly nonlinear when mapped back onto the predictor space. The degree of boundary complexity (nonlinearity) is determined by the mapping and its associated kernel, which a function in the predictor space that is equivalent to the dot product in the higher dimensional mapping (Hastie et al., 2001).

Neural Networks

Neural Networks consist of a system of "nodes," which are each functions of the predictors. The functions themselves are sigmoidal activation functions that take as an input a linear combination of the predictors, and take a value between zero and one. The output of these nodes are then further combined to give a single output. The organization here is meant to invoke the image of a brain. Signals pass through a system of nodes which are activated in different combinations to produce a final signal (Hastie et al., 2001).

The optimization problem is then to choose the appropriate weights in each linear combination, and since there is a separate set of weights for each node, the number of free parameters can get large quickly. Since we do not have that many observations relative to our predictor dimensionality, we must first cut down the number of predictors by taking principal components of the predictors. As a result, we consider the first five principal components of the set of predictors.

Ensemble Bayesian Model Averaging

Since each of these models has its own strengths and drawbacks, we have an aggregate model which is a weighted average of the five. Starting with the prior that each algorithm is equally appropriate, we use cross validationresampling training and test observations from our training sampleto update our weights based on the accuracy of each model (Montgomery et al., 2012).

4.2 Training and Testing Methods

We index observations by the subdistrict i in year t. A given observation consists of the predictors measured in t as well as lagged predictors measured in $t-1$, and $t-2$ where available, as well as time invariant predictors such as the census variables which are only measured once in 2000. In year t, when we want to forecast where violence will happen in year $t+1$ we do the following.

1. For each model, we take observations from t_0 to $t-1$ as our training set.
2. We use 5-fold cross validation to choose optimal tuning parameters. 5-fold cross validation is a method by which researchers can simulate out of sample predictions. In this

procedure, the data is randomly partitioned into 5 equal sized subsamples. For each subsample, a model is fit to the other four subsamples and used to predict results for the remaining sample. This is repeated for each of the five subsamples, so that there is an "out of sample" prediction of each observation in the sample. This is conducted for each tuning parameter value in the parameter search space. The best performing parameter is chosen.

3. We repeat step 2 10 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 in observations in year t to predict violence in year $t + 1$.

We want to have two lagged measurements of violence. Therefore, we can fit models starting from 2006 onward, meaning we generate true year ahead, out of sample predictions for years 2008-2009.

4.3 Assessing variable importance

One important feature of prediction exercises is the ability to identify key variables driving the forecastability of conflict. We discuss here several approaches to doing so.

4.3.1 Typical approaches to variable importance

In an initial inquiry into the importance of various predictors, a researcher might be tempted to look at model parameters.

For LASSO, this is typically done by inspecting the predictor coefficients (Blair et al., 2015). LASSO will set the coefficents on the least important variables to zero, and only retain a set of predictors that is most useful for prediction. However, LASSO will tend to select predictors that are uncorrelated with one another. Therefore, if we include, for example, mean years of schooling and some measure of the literacy rate, the two are likely to be correlated, and LASSO would typically pick only one. This is fine if we just aim to trim the set of predictors, but this is less useful if we hope our model can inform theories. In this example, LASSO would assign zero importance to one of the variables, although both would, in truth, have similar importance since they measure a similar concept. While this distinction seems trivial in this example, one could imagine less obvious correlations between predictors that would generate similarly misleading results. Furthermore, comparing predictor importance within the set of predictors with non-zero coefficients can be a complicated process, as coefficients must be interpreted as conditional on keeping other predictors fixed, when in reality, the predictors may covary.

For Support Vector Machines, some researches suggest looking at the equation that defines the separating hyperplane (Hastie et al., 2001). If a hyperplane is orthogonal to a predictor's axis in the predictors space, then that predictor is very important. However, when one uses

a kernel to map the predictor space to a higher dimensional space, the hyperplane becomes much more difficult to interpret in the original predictor space. On the other hand, Neural Networks simply involve far too many parameters for sensible interpretation.

For Random Forests and AdaBoost, researchers have taken a different tack (Strobl et al., 2009). After the model is fitted, one can randomly permute each variable, one at a time, and measure the decrease in predictive performance. The downside of this strategy is that it keeps the model fixed. If one were to omit the variable completely and refit the model, one could imagine that the model might be able to find a suitable stand-in for the omitted variable, and performance would not suffer. In this way, this manner of measuring variable importance does not measure the *unique* contribution of each variable to the model's predictive performance.

Moreover, with all of these methods, by simply looking at the importance of individual variables, it is difficult to see if entire classes of variables are important. For example, if you had 50 variables about demographics, and 1 variable about economic shocks, it is likely that that single economic variable would be more important than any single demographic variable, even though the demographic variables as a whole might be much more important. Hence, in this section, we will look at model performance when we exclude classes of predictors.

And finally, we should note the disadvantage that the idiosyncrasies of these importance measures vary across models, which can pose difficulties for cross-model comparisons.

4.3.2 Alternative approaches

In a first attempt at assessing variable importance, we might simply ask how well individual subgroups of variables predict violence. However, since population predicts violence for mechanical reasons (more populous subdistricts simply have more people to conflict with one another), we will see that variable groups most correlated with population and population density perform best. Therefore, in order to "control" for the population effects, we will start with a baseline of a set of predictors that only measure population and population density, and then add individual groups of predictors to that set and measure the increase in predictive performance. In doing this, we measure the total information contained in each variable set beyond information about population.

The above approach has the limitation that, because many groups of predictors may be correlated and hence contain similar information, we will be measuring total information content rather than information that is unique to that variable group. Therefore, an alternative approach will be to start with the full set of predictors, and then remove individual groups and measure the reduction in predictive performance. However, since we posit that economic and demographic characteristics of subdistricts are good predictors of violence, they should also be highly correlated with past violence. Therefore, we expect past violence to already incorporate much of the information contained in the other predictor variables. Indeed, as we will see in the results, past measures of violence alone are good predictors of violence, and when we restrict the predictor set to only past violence measures, performance suffers, but is relatively close to the performance of the full models. This difference in performance between the full set of predictors and the restricted set provides a narrow window in which we can see the marginal contributions of other variables. To widen this window, we consider a different

baseline universe of full information U', in which we have our full set of predictors, but have no knowledge of past violence. We will then proceed as in the first, removing subsets of the predictors and measuring the decrease in performance.

Finally, we offer an alternative method of dealing with the correlation between variable sets. We aim to address this by removing not only the group of variables in question, but also the correlation of the remaining variables with the predictors in the excluded group. More formally, if we want to exclude the predictor set X, we regress each z_i in $(U - X)$ on the predictors in X, set z_i^* equal to the residual, and use the set of z_i^*'s as predictors. Thus, the remaining predictors are uncorrelated with X. The resulting drop in model performance is the most liberal estimate of the importance of the excluded variables. Note that this is analogous to asimple partialling out of regressors in an OLS context.

5 Results

5.1 How well can we predict violence?

In Table 3 we report performance metrics for each model for each out-of-sample year individually. For the 2008 predictions, we fit each model to 2007 violence, and for the 2009 predictions, we train each model with 2007 and 2008 data, and so on. Each model generates a predicted probability, rather than a binary indicator of violence. Therefore, one of the parameters chosen in the cross-validation procedure above is a discrimination threshold above which we classify a subdistrict as predicted to have a violent event in the upcoming year. We choose this threshold to maximize accuracythe share of subdistricts where the incidence of violence is accurately predicted.

Table 3: Out-of-Sample performance

	LASSO (1)	Random Forest (2)	Ada Boost (3)	SVM (4)	Neural Network (5)	EBMA (6)
2008						
Accuracy	0.713	0.706	0.712	0.672	0.710	0.706
Sensitivity	0.670	0.617	0.654	0.753	0.800	0.657
Specificity	0.766	0.815	0.783	0.575	0.601	0.766
AUC	**0.782**	**0.794**	**0.789**	**0.742**	**0.781**	**0.792**
DV Mean	0.550					
2009						
Accuracy	0.736	0.715	0.731	0.719	0.702	0.742
Sensitivity	0.778	0.733	0.707	0.734	0.804	0.749
Specificity	0.690	0.694	0.758	0.701	0.589	0.735
AUC	**0.813**	**0.801**	**0.805**	**0.784**	**0.789**	**0.816**
DV Mean	0.527					
2010						
Accuracy	0.724	0.702	0.732	0.718	0.702	0.729
Sensitivity	0.771	0.673	0.750	0.778	0.771	0.777
Specificity	0.671	0.736	0.710	0.648	0.622	0.673
AUC	**0.805**	**0.792**	**0.809**	**0.781**	**0.779**	**0.809**
DV Mean	0.536					
2011						
Accuracy	0.723	0.701	0.726	0.708	0.715	0.730
Sensitivity	0.773	0.689	0.734	0.795	0.779	0.769
Specificity	0.667	0.714	0.718	0.609	0.642	0.686
AUC	**0.806**	**0.783**	**0.807**	**0.781**	**0.787**	**0.809**
DV Mean	0.531					

(a) Indicator of any non-crime event

	LASSO (1)	Random Forest (2)	Ada Boost (3)	SVM (4)	Neural Network (5)	EBMA (6)
2008						
Accuracy	0.922	0.933	0.928	0.906	0.922	0.927
Sensitivity	0.519	0.599	0.599	0.294	0.520	0.602
Specificity	0.977	0.981	0.976	0.994	0.980	0.972
AUC	**0.921**	**0.934**	**0.930**	**0.887**	**0.929**	**0.929**
DV Mean	0.120					
2009						
Accuracy	0.922	0.925	0.928	0.916	0.924	0.927
Sensitivity	0.519	0.606	0.598	0.456	0.548	0.602
Specificity	0.977	0.969	0.973	0.979	0.975	0.972
AUC	**0.921**	**0.934**	**0.935**	**0.902**	**0.932**	**0.929**
DV Mean	0.120					
2010						
Accuracy	0.934	0.932	0.931	0.932	0.920	0.939
Sensitivity	0.575	0.633	0.596	0.558	0.596	0.629
Specificity	0.983	0.972	0.977	0.983	0.964	0.981
AUC	**0.942**	**0.936**	**0.939**	**0.920**	**0.919**	**0.945**
DV Mean	0.119					
2011						
Accuracy	0.938	0.937	0.936	0.931	0.929	0.937
Sensitivity	0.655	0.647	0.635	0.569	0.600	0.643
Specificity	0.979	0.979	0.980	0.984	0.977	0.980
AUC	**0.941**	**0.936**	**0.944**	**0.925**	**0.930**	**0.944**
DV Mean	0.127					

(b) Indicator of ≥ 5 non-crime events

Notes: Each model is trained on all data available preceding the out-of-sample prediction year. Accurary is the proportion of subdistrict correctly predicted. Sensitivity is the proportion of subdistrict 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 few clear patterns emerge. First, model performance increases from 2008 to 2009, reflecting the increase in the training sample size.[7] Each of the five models performs reasonably well with accuracy above 70% for predictions of any event and above 90% for predictions of five or more events. This similar performance accross models is striking. Each method uses very different assumptions, yet there are few differences in performance. The average of the models (EBMA), performs best in each year, although the differences are not large. We also report the sensitivity of the the model the percent of violent cases that are accurately predicted and specifity the percent of non-events that are correctly predicted.

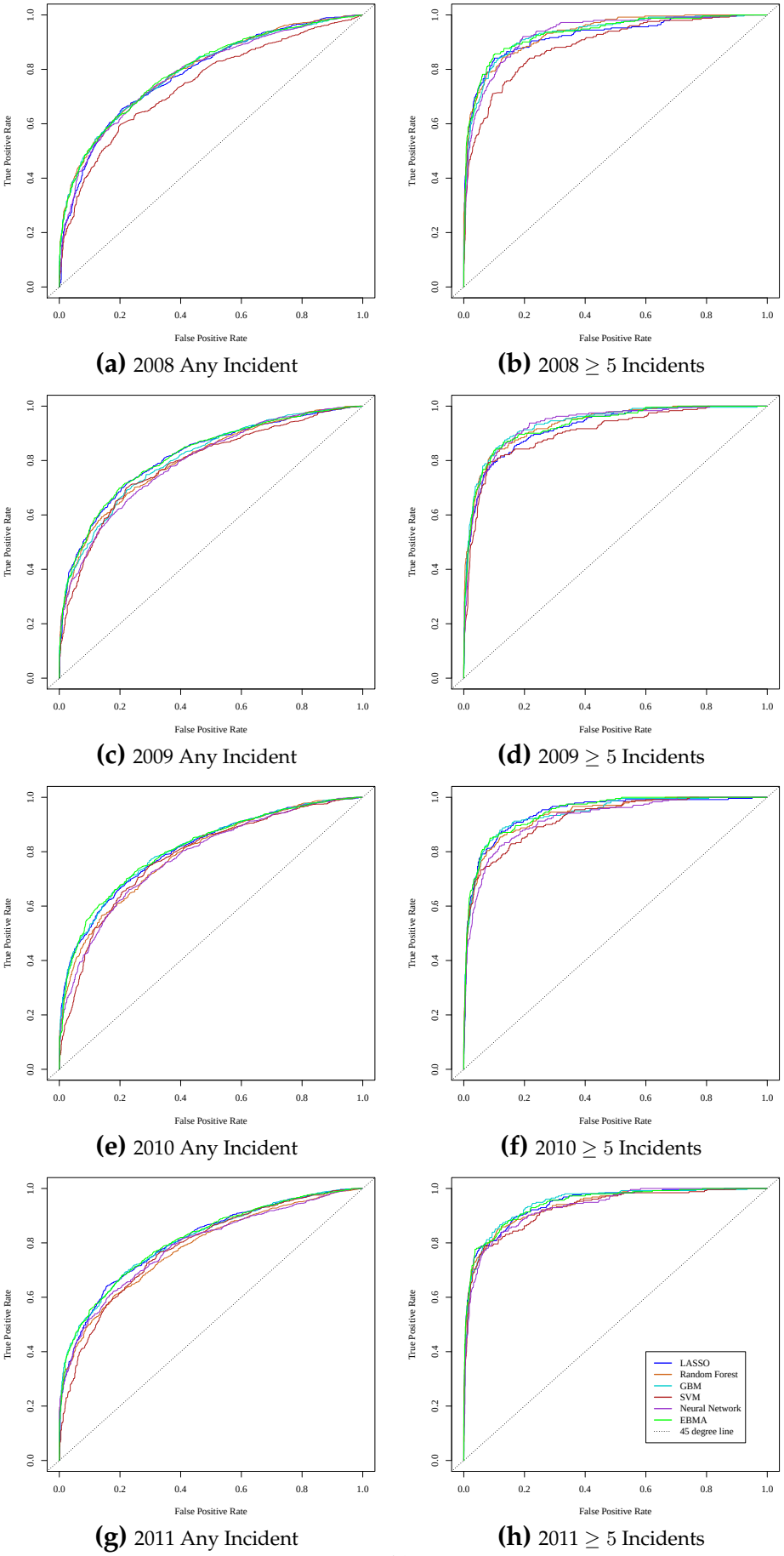
However, we should note that these accuracy, sensitivity, and specificity rates are somewhat arbitrary. Each of these model generates a probability of violence between 0 and 1, rather than a binary prediction. Through cross validation, we have chosen discrimination thresholds for these predicted probabilities that maximize accuracy. However, a policy maker may judge the costs of false positives and false negatives differently, and if she wants to prevent all violence, she might choose a lower discrimination threshold that increases sensitivity at the cost of specificity and accuracy. Conversely, if the policy maker has scarce security resources and only wants to allocate them to subdistricts that are certain to experience violence, she might raise the discrimination threshold, increasing specificity at the cost of sensitivity and accuracy.

In light of this, we show receiver operating characteristic curves (ROCs). The ROC curve plots the true positive and false positive rates as the discrimination threshold changes. In the bottom left hand corner, the we see that with a sufficiently high discrimination threshold, all subdistricts are classified as non-violent and the true postive and false positive rates are zero. Likewise, when the discrimination threshold is very low, the true positive rate is high, but so is the false positive rate. Ideally, we want to increase the true positive rate without increasing the false positive rate, so we want to see a large area under the ROC. Finally, note that any point on the 45 degree line is attainable by random chance. For example, if one was to forecast violence in 60% of subdistricts and choose those subdistricts randomly, we would expect both the true and false positive rates to be .6. The final line of Table 3 reports the area under the ROC curves (AUCs) for each model. The AUC for the ensemble in 2011 is particularly good 0.809 for the indicator of any event and 0.944 for the indicator of at least 5 events.[8]

[7] Recall, that we are holding back an additional 5 years of data, so we would expect performance to increase even further.

[8] Another interpretation of the AUC is that it measures the probability that, if we randomly select 2 observations, they will be correctly ordered in their predicted risk for violence, i.e. the probability that the subdistrict at greater risk for violence is assigned a higher probability of violence.

Figure 4: ROC Performance



(a) 2008 Any Incident

(b) 2008 $\geq$ 5 Incidents

(c) 2009 Any Incident

(d) 2009 $\geq$ 5 Incidents

(e) 2010 Any Incident

(f) 2010 $\geq$ 5 Incidents

(g) 2011 Any Incident

(h) 2011 $\geq$ 5 Incidents

Finally, it is worth noting that LASSO is nearly the best performing model in each year. This is important given LASSO's simplicity. It is the most accessible of these algorithms to social scientists who are generally familiar with logistic regression. This suggests that despite other machine learning advances since LASSO's development in the 1990s (Tibshirani (1994)), LASSO remains the best starting place for social scientists interested in predictor or dealing with high dimensional data.

5.2 What do we need to predict violence?

The above results are encouraging from a policy maker's perspective violent events can be predicted with some degree of accuracy. We might think that if violence is highly autocorrelated, this impressive performance may simply follow from predicting recurrence in places that have experienced violence in the past. Table 4 tests this theory. To ease display, we make a few simplifications to the results we pool the 2008-2011 out of sample predictions and look at performance metrics over all years, we report only the EBMA results, as the results are broadly consistent across machine learning methods, and we report only the AUC, as it is our preferred measure of performance. Finally, we report the p-values of the difference between the limited models and the full model. We calculate the p-value via bootstrapping in which we resample subdistricts (grouping both years) 100 times to estimate the standard error of the difference.

Table 4: EBMA AUC

	Full Predictors (1)	Only Lagged Indicator (DV) (2)	Only Lagged Violence (3)	Lagged Violence and Population (4)	No Lagged Violence (5)
Any Incident (AUC)	0.807	0.735	0.793	0.805	0.784
reduction from (1)		-0.072	-0.014	-0.002	-0.023
p-value (bootstrap)		0.000	0.000	0.420	0.002
p-value (HM test)		0.000	0.000	0.532	0.000
≥ 5 Incidents (AUC)	0.936	0.856	0.928	0.940	0.920
reduction from (1)		-0.080	-0.009	0.003	-0.017
p-value (bootstrap)		0.000	0.000	0.041	0.000
p-value (HM test)		0.000	0.005	0.311	0.000

Notes: For each specification, we pool the 2008-2011 out-of-sample estimates and report the area under the ROC curve (AUC) for the pooled sample. P-values are derived from a bootstrapping procedure in which we resample subdistricts to estimate the standard error of the difference in AUC between models.

Column (1) reports performance when we use the full set of predictors. Our first concern might be that violence is simply persistent, and our models' impressive performance is purely the result of the autoregressive properties of violence. To test this simple recurrence model, column (2) reports performance when we restrict the predictors to two lags of the dependent variable. We see that performance is still good, but it is considerably worse than our full model, indicating that our rich set of predictors is actually offering useful information for predicting violence. However, we have measurements of past violence that go far beyond the lagged dependent variable. In colomn (3), we include measures of deaths, damage, and incident counts, broken out be category of violence, as well as conflict indicators from Podes. We include two lags of each of these measures. When the predictors are restricted, performance is still worse

than that of the full model, but not by a large amount. Finally, we should note that there is a simple, mechanical relationship between violence and population, population density, and grownmore populous subdistricts have more people who can engage in violent activity and hence are more likely to experience violence. Therefore, we add measures of population to lagged violence measures (the broader set, beyond the lagged dependent variable) in column (4). The performance of past violence plus population is indistinguishable from that of the full model.

Finally, given the impressive performance of a model that includes only information about past violence, we might ask "How crucial is knowledge of past violence to making predictions about future conflict?" Column (5) reports performance when we exclude all measures of past violence from the predictor space. Without information on past conflicts, the models still perform remarkably well, indicating that past conflicts are reflective of the same economic, social and polictical fundamentals that are driving current conflict and that our set of predictors is fairly comprehensive.

5.3 Which predictors of violence matter most?

While past conflicts will obviously be a valuable predictor of future conflicts, we are also interested in which economic, social and political variables are the best predictors of conflict. In this section, we investigate a number of approaches to assess variable importance.

5.3.1 Performance of individual variable groups

To begin to investigate variable importance, we first examine how subgroups of our predictors perform alone. We group these categories of predictors into "economic" and "noneconomic" variables. Figure 5 also shows these performances graphically, with AUC's plotted along the horizontal axis. Variable groups are sorted by performance.

Table 5: AUCs Removing Broad Categories

	Only Agric. Infrastructure (1)	Only Finance and Trade Institutions (2)	Only Economic Stucture (3)	Only Government Finances (4)	Only Output (5)	Only Inequality (6)	Only Labor Market (7)	Only Commodity Shocks (8)	Only Weather Shocks (9)	Only Commodity Volatility (10)	Only Weather Means and Volatility (11)
Any Incident (AUC)	0.608	0.682	0.756	0.709	0.738	0.652	0.613	0.506	0.505	0.639	0.723
≥ 5 Incidents (AUC)	0.761	0.841	0.905	0.837	0.862	0.771	0.716	0.650	0.539	0.793	0.859

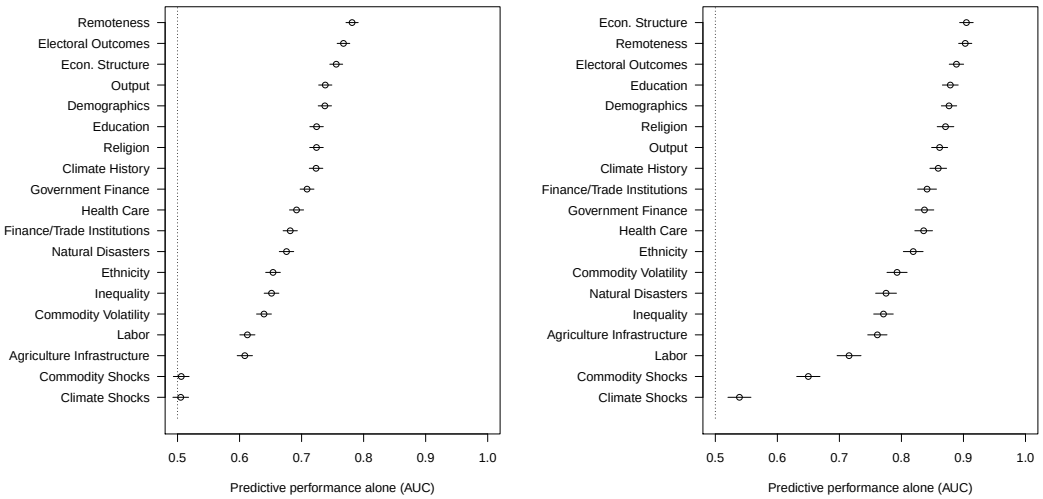
(a) Economic Variables

	Only Religious Predictors (1)	Only Ethnic Predictors (2)	Only Education Predictors (3)	Only Remoteness Predictors (4)	Only Demographic Predictors (5)	Only Health Infrastructure Predictors (6)	Only Natural Disaster Predictors (7)	Only Election Predictors (8)
Any Incident (AUC)	0.724	0.654	0.724	0.781	0.737	0.692	0.676	0.767
≥ 5 Incidents (AUC)	0.871	0.819	0.879	0.903	0.877	0.836	0.775	0.889

(b) Noneconomic Variables

Notes: For each specification, we pool the 2008-2011 out-of-sample estimates and report the area under the ROC curve (AUC) for the pooled sample. P-values are derived from a bootstrapping procedure in which we resample subdistricts to estimate the standard error of the difference in AUC between models.

Figure 5: AUCs improvements from individual predictor groups



(a) Any incident **(b)** Five or more incidents

Among the economic predictors, we find several sets of variables to be relatively good predictors of violence. Variables concerning economic structure, such as sectoral shares of the labor force, are the best performing predictors of violence, with output measures coming in a close second. Among the noneconomic predictors, measures of remoteness, such as road access are important predictors, as are variables concerning religious, ethnic, and political heterogeneity. Commodity price and weather shocks, which receive much attention in the conflict literature, are among the worst performing predictors.

However, we should also note that many of the top performing predictors above are also likely to be the most correlated with population, which we have previously noted to be a good predictor of violence, albeit for uninteresting reasons. Table 6 considers the increase in performance from the addition of each of these predictor groups to a baseline model with population. The most effective economic predictors continue to be measurements of economic structure, and measures of remotenss, ethnicity and religion remain important. However, once we include population, electoral outcomes are shown to be somewhat less important, indicating that voting patterns may closely map to population and density.

Table 6: AUCs from individual predictor groups

	Only Population Measures (1)	Only Agric. Infrastructure (2)	Only Finance and Trade Institutions (3)	Only Economic Stucture (4)	Only Government Finances (5)	Only Output (6)	Only Inequality (7)	Only Labor Market (8)	Only Commodity Shocks (9)	Only Weather Shocks (10)	Only Commodity Volatility (11)	Only Weather Means and Volatility (12)
Any Incident (AUC)	0.763	0.770	0.770	0.797	0.780	0.772	0.753	0.757	0.741	0.749	0.768	0.776
increase from (1)		0.007	0.007	0.033	0.017	0.009	-0.010	-0.007	-0.022	-0.014	0.005	0.013
p-value (bootstrap)		0.000	0.000	0.000	0.000	0.083	0.166	0.396	0.009	0.000	0.000	0.000
p-value (HM test)		0.005	0.010	0.000	0.000	0.002	0.000	0.010	0.000	0.000	0.073	0.000
$\geq$ 5 Incidents (AUC)	0.869	0.894	0.890	0.923	0.901	0.901	0.877	0.874	0.834	0.831	0.887	0.886
increase from (1)		0.025	0.021	0.055	0.033	0.032	0.008	0.006	-0.034	-0.037	0.019	0.018
p-value (bootstrap)		0.000	0.000	0.000	0.000	0.000	0.589	0.637	0.000	0.049	0.000	0.000
p-value (HM test)		0.000	0.000	0.000	0.000	0.000	0.140	0.295	0.000	0.000	0.000	0.000

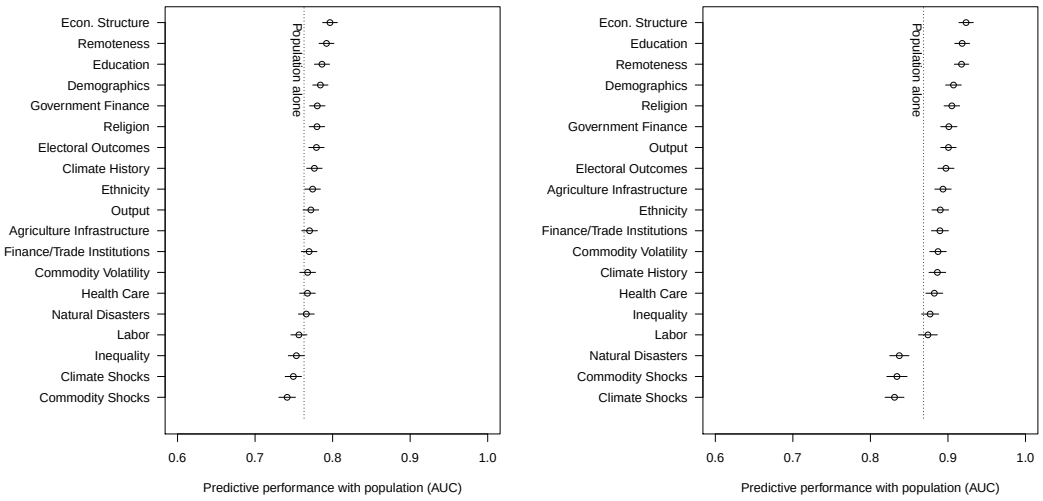
(a) Economic Variables

	Only Population Measures (1)	Only Religious Predictors (2)	Only Ethnic Predictors (3)	Only Education Predictors (4)	Only Remoteness Predictors (5)	Only Demographic Predictors (6)	Only Health Infrastructure Predictors (7)	Only Natural Disaster Predictors (8)	Only Election Predictors (9)
Any Incident (AUC)	0.763	0.780	0.774	0.786	0.792	0.784	0.768	0.766	0.779
increase from (1)		0.017	0.011	0.023	0.029	0.021	0.004	0.003	0.016
p-value (bootstrap)		0.000	0.000	0.000	0.000	0.000	0.022	0.652	0.026
p-value (HM test)		0.000	0.000	0.000	0.000	0.000	0.088	0.303	0.000
$\geq$ 5 Incidents (AUC)	0.869	0.905	0.890	0.918	0.918	0.907	0.882	0.837	0.897
increase from (1)		0.036	0.022	0.050	0.049	0.039	0.014	-0.031	0.029
p-value (bootstrap)		0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000
p-value (HM test)		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

(b) Noneconomic Variables

Notes: For each specification, we pool the 2008-2011 out-of-sample estimates and report the area under the ROC curve (AUC) for the pooled sample. P-values are derived from a bootstrapping procedure in which we resample subdistricts to estimate the standard error of the difference in AUC between models.

Figure 6: AUCs improvements from individual predictor groups



(a) Any incident **(b)** Five or more incidents

To ease reading, we also present these results graphically. Figure 6 shows how the addition of predictor groups to a baseline model that includes only population variables improves performance.

5.3.2 Disentangling correlated variable sets

In the exercises above, we recognize that many of the sets of variables may be highly related to one another. For example, if one ethnic or religious group reside in wealthier areas or specialize in a certain type of economic activity, then measure of that ethnic group's presence and measures of economic structure may be similarly effective predictors of violence, and it is difficult to determine which aspect of the location is the true driver of conflict or a lack of conflict. In some sense, we are interested in what is the unique information contained in each subgroup of variables that is absent from its complement. To measure this, we offer two approaches.

First, we start with the full set of predictors, excluding measures of past violence, and remove a given set of predictors to assess its marginal contribution to prediction. Table 7 reports the reduction in performance as we remove each predictor set. Few predictor groups lead to a statistically (or substantively) significant drop in performance. The marginal cases are variables concerning sectoral structure and education levels. Also, note that the exclusion of economic shocks actually *increases* performance.

Table 7: AUCs removing predictors

	Full Predictors Excluding Violence (1)	Excl. Agric. Infrastructure (2)	Excl. Finance and Trade Institutions (3)	Excl. Economic Stucture (4)	Excl. Government Finances (5)	Excl. Output (6)	Excl. Inequality (7)	Excl. Labor Market (8)	Excl. Commodity Shocks (9)	Excl. Weather Shocks (10)	Excl. Commodity Volatility (11)	Excl. Weather Means and Volatility (12)
Any Incident (AUC)	0.784	0.783	0.787	0.781	0.784	0.786	0.788	0.786	0.792	0.784	0.784	0.783
decrease from (1)		0.001	-0.003	0.003	0.000	-0.003	-0.004	-0.002	-0.009	0.000	-0.001	0.001
p-value (bootstrap)		0.000	0.085	0.062	0.912	0.032	0.036	0.074	0.046	0.985	0.029	0.679
p-value (HM test)		0.666	0.216	0.310	0.951	0.309	0.109	0.373	0.001	0.986	0.749	0.759
$\geq$ 5 Incidents (AUC)	0.920	0.920	0.919	0.920	0.919	0.920	0.922	0.922	0.921	0.922	0.919	0.918
decrease from (1)		-0.001	0.001	0.000	0.000	0.000	-0.002	-0.002	-0.002	-0.002	0.000	0.002
p-value (bootstrap)		0.480	0.033	0.693	0.829	0.804	0.175	0.086	0.282	0.198	0.041	0.241
p-value (HM test)		0.820	0.767	0.902	0.935	0.951	0.565	0.484	0.616	0.525	0.892	0.605

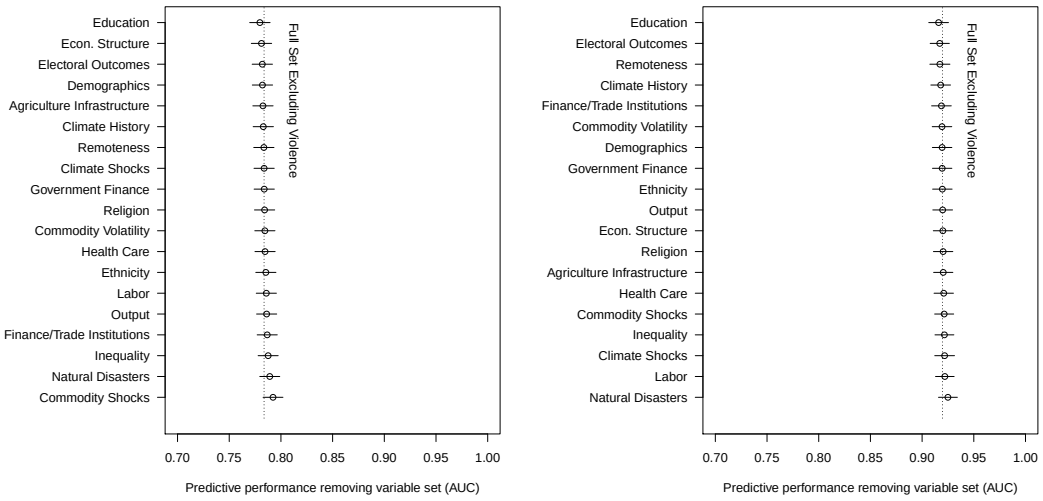
(a) Economic Variables

	Full Predictors Excluding Violence (1)	Excl. Religious Predictors (2)	Excl. Ethnic Predictors (3)	Excl. Education Predictors (4)	Excl. Remoteness Predictors (5)	Excl. Demographic Predictors (6)	Excl. Health Infrastructure Predictors (7)	Excl. Natural Disaster Predictors (8)	Excl. Election Predictors (9)
Any Incident (AUC)	0.784	0.784	0.785	0.780	0.783	0.782	0.785	0.789	0.782
decrease from (1)		-0.001	-0.002	0.004	0.000	0.002	-0.001	-0.006	0.002
p-value (bootstrap)		0.706	0.312	0.002	0.865	0.015	0.175	0.001	0.167
p-value (HM test)		0.835	0.480	0.103	0.935	0.529	0.700	0.025	0.493
$\geq$ 5 Incidents (AUC)	0.920	0.920	0.920	0.916	0.917	0.919	0.921	0.925	0.917
decrease from (1)		-0.001	0.000	0.004	0.002	0.000	-0.001	-0.005	0.003
p-value (bootstrap)		0.373	0.902	0.000	0.082	0.498	0.000	0.105	0.004
p-value (HM test)		0.845	0.977	0.266	0.455	0.920	0.708	0.105	0.432

(b) Noneconomic Variables

Notes: For each specification, we pool the 2008-2011 out-of-sample estimates and report the area under the ROC curve (AUC) for the pooled sample. P-values are derived from a bootstrapping procedure in which we resample subdistricts to estimate the standard error of the difference in AUC between models.

Figure 7: AUCs improvements from individual predictor groups



(a) Any incident **(b)** Five or more incidents

Again, we present these results graphically in figure 7.

Finally, we propose an exercise that asks, "How much information remains when we remove all information contained in a given subset and purge the remaining variables of all correlation with the excluded variable?" More formally, if we want to exclude the predictor set X, we regress each z_i in $(U - X)$ on the predictors in X, set z_i^* equal to the residual, and use the set of z_i^*'s as predictors. Thus, the remaining predictors are uncorrelated with X. The resulting drop in model performance is the most liberal estimate of the importance of the excluded variables. Note that this is analogous to asimple partialling out of regressors in an OLS context.

Table 8: AUCs partialling out predictors

	Full Predictors Excluding Violence (1)	Partial out Agric. Infrastructure (2)	Partial out Finance and Trade Institutions (3)	Partial out Economic Stucture (4)	Partial out Government Finances (5)	Partial out Output (6)	Partial out Inequality (7)	Partial out Labor Market (8)	Partial out Commodity Shocks (9)	Partial out Weather Shocks (10)	Partial out Commodity Volatility (11)	Partial out Weather Means and Volatility (12)
Any Incident (AUC)	0.784	0.783	0.776	0.765	0.761	0.763	0.754	0.755	0.678	0.759	0.773	0.784
decrease from (1)		0.000	0.007	0.019	0.023	0.021	0.029	0.029	0.105	0.024	0.011	-0.001
p-value (bootstrap)		0.956	0.001	0.000	0.001	0.000	0.001	0.000	0.032	0.000	0.000	0.795
p-value (HM test)		0.935	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.780
$\geq$ 5 Incidents (AUC)	0.920	0.919	0.911	0.890	0.895	0.881	0.881	0.894	0.754	0.895	0.911	0.918
decrease from (1)		0.001	0.009	0.030	0.025	0.039	0.038	0.025	0.165	0.025	0.008	0.001
p-value (bootstrap)		0.553	0.001	0.000	0.000	0.000	0.000	0.001	0.103	0.004	0.094	0.625
p-value (HM test)		0.737	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.013	0.718

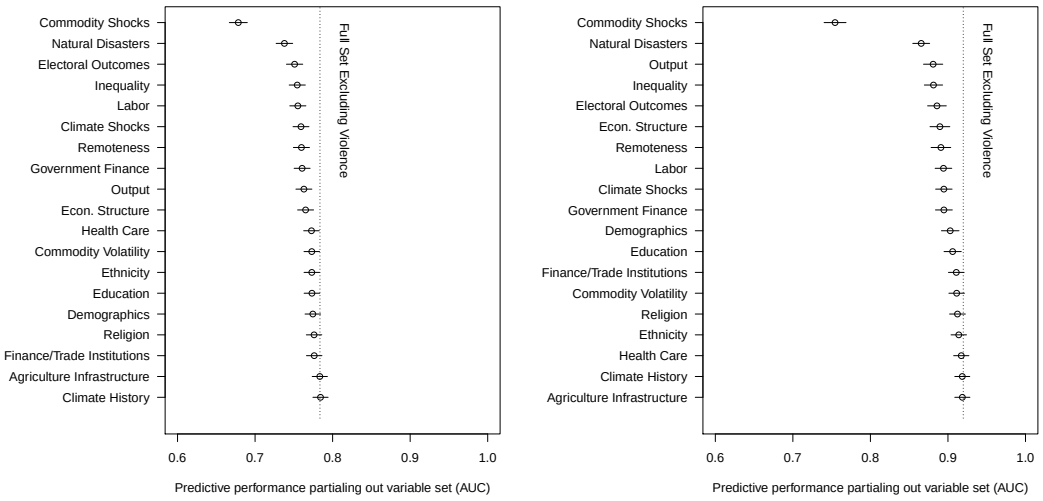
(a) Economic Variables

	Full Predictors Excluding Violence (1)	Partial out Religious Predictors (2)	Partial out Ethnic Predictors (3)	Partial out Education Predictors (4)	Partial out Remoteness Predictors (5)	Partial out Demographic Predictors (6)	Partial out Health Infrastructure Predictors (7)	Partial out Natural Disaster Predictors (8)	Partial out Election Predictors (9)
Any Incident (AUC)	0.784	0.776	0.773	0.773	0.760	0.774	0.773	0.738	0.751
decrease from (1)		0.008	0.011	0.010	0.024	0.009	0.011	0.046	0.033
p-value (bootstrap)		0.000	0.000	0.035	0.000	0.000	0.001	0.000	0.000
p-value (HM test)		0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\geq$ 5 Incidents (AUC)	0.920	0.912	0.914	0.906	0.891	0.903	0.917	0.866	0.886
decrease from (1)		0.007	0.006	0.014	0.029	0.017	0.003	0.054	0.034
p-value (bootstrap)		0.250	0.095	0.000	0.000	0.000	0.505	0.047	0.000
p-value (HM test)		0.029	0.088	0.000	0.000	0.000	0.454	0.000	0.000

(b) Noneconomic Variables

Notes: For each specification, we pool the 2008-2011 out-of-sample estimates and report the area under the ROC curve (AUC) for the pooled sample. P-values are derived from a bootstrapping procedure in which we resample subdistricts to estimate the standard error of the difference in AUC between models.

Figure 8: AUCs improvements from individual predictor groups



(a) Any incident **(b)** Five or more incidents

Figure 8 presents these results as well.

The partialling out of almost all variable categories leads to a statistically significant drop in performance, indicating that there is much shared information between each of these variables. The largest reductions come from partialling out inequality and labor market variables, which were not important variable groups in the earlier exercises. Among the noneconomic variables, remoteness and natural disasters contain the most information and lead to the largers reductions in performance when the other variables are orthogonalized to them.

Partialing out economic shocks also leads to large reductions in performance. That said, we regard these economic shock results with caution for two reasons. First, we worry that these shocks are reflecting fixed effects, rather than true exogenous shocks. Several groups of non-economic variables perform relatively well alone (see tables 5 and 6), and most of these variables should not be altered by economic shocks, at least not in the short timespan that we are observing here. Furthermore, many of these variables come from the 2000 census which should not be affected by exogenous price shocks that come years later. Instead, it is likely that subdistricts who are least diversified in their agricultural production are most prone to large weighted price shocks, and therefore the large shocks actually reflect other fixed factors that lead to both a lack of agricultural diversification and other demographic and economic trends. A second concern is that our panel is short, and these shocks are likely to correlated across time and space. Appendix section A.7 offers a placebo test that demonstrates that we should not take this result seriously.

5.4 Summary Indices of Importance

The tables above assess the total and marginal (unique) information contributions of each set of predictors. However, the changes in performance are on different scales for each of the

Table 9: Importance Indices

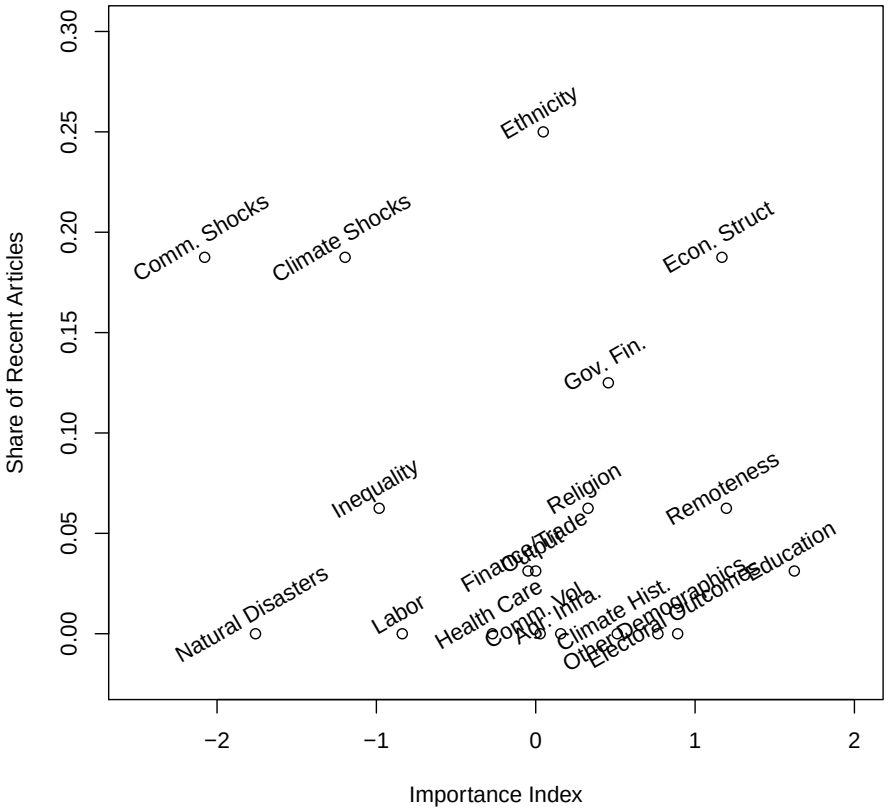
	Any incident			Five or more incidents			
	Excl. from full minus violence	Alone plus pop.	Sub Index	Excl. from full minus violence	Alone plus pop.	Sub Index	Composite Index
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Education	1.69	1.02	**1.45**	1.89	1.16	1.65	**1.62**
Remoteness	0.39	1.42	**0.97**	1.30	1.13	1.31	**1.20**
Economic Structure	1.17	1.74	**1.56**	-0.10	1.34	0.67	**1.17**
Elections	0.89	0.52	**0.76**	1.36	0.38	0.94	**0.89**
Demographics	0.85	0.87	**0.92**	0.26	0.74	0.54	**0.77**
Climate History	0.58	0.33	**0.49**	0.93	-0.03	0.49	**0.51**
Government Finance	0.27	0.59	**0.47**	0.23	0.52	0.40	**0.46**
Religion	0.15	0.56	**0.38**	-0.21	0.66	0.24	**0.33**
Agriculture Infrastructure	0.68	-0.10	**0.31**	-0.27	0.24	-0.01	**0.16**
Ethnicity	-0.26	0.17	**-0.05**	0.14	0.11	0.14	**0.05**
Commodity Shock Volatility	0.06	-0.28	**-0.12**	0.31	0.01	0.17	**0.03**
Output	-0.52	0.02	**-0.27**	0.00	0.50	0.27	**0.00**
Finance and Trade Institutions	-0.70	-0.14	**-0.45**	0.57	0.10	0.36	**-0.05**
Health Infrastructure	0.00	-0.29	**-0.16**	-0.50	-0.17	-0.36	**-0.27**
Labor Markets	-0.42	-1.06	**-0.79**	-1.01	-0.48	-0.81	**-0.84**
Inequality	-1.01	-1.29	**-1.23**	-0.82	-0.37	-0.64	**-0.98**
Climate Shocks	0.31	-1.56	**-0.67**	-0.91	-2.07	-1.61	**-1.20**
Disasters	-1.54	-0.39	**-1.03**	-2.46	-1.84	-2.32	**-1.76**
Commodity Shocks	-2.60	-2.12	**-2.53**	-0.70	-1.95	-1.43	**-2.08**

measures. Here we standardize these scores, in an attempt to calculate a single measure of variable importance.

Table 9 reports these importance measures for predictions of any incident and an indicator of five or more incidents. For each dependent variable, we compute the average of the indices, and we also compute the average over all eight specifications. Predictor sets are ranked by importance. Demographics such as population, gender, and age are the most important predictors by most measures. Fixed economic variables like GDP and output share by industry are a close second, but we should note that most of their importance comes from columns (3), (4), (8) and (9), indicating that most of the information contained in these variables is shared with other predictors. Economic shocks are clearly the least important predictors, with little marginal or shared information.

Finally, we use the composite indices above to assess the degree to which the emphasis in the conflict literature reflects the importance of mechanisms of interest for conflict prediction. To measure importance, we counted the number of articles in top economics and political science journals that investigated each set of predictors as a primary cause or correlate of conflict. We restricted our search to the past five years of the *American Economic Review*, *American Economic Journal*, *Quarterly Journal of Economics*, *Journal of Political Economy*, *Econometrica*, *Review of Economic Studies*, *Review of Economics and Statistics*, *Journal of Development Economics*, *Journal of*

Figure 9: Literature Emphasis versus Predictive Importance



Economic Literature, *Journal of Economic Perspectives*, *American Political Science Review*, *American Journal of Political Science*, and *Political Analysis*. The articles are reported in Appendix table A.5.

Obviously, predictive importance is not necessarily the most important determinant of what makes a mechanism interesting for study. Researchers may be more interested in the mechanisms that can be easily affected by policy or in mechanisms which lend themselves to causal identification. However, it is striking the relationship between journal emphasis is not only not positve, it is actually somewhat negative. Therefore we think prediction exercises such as ours can be a useful tool to identify mechanisms of future interest.

6 Discussion

6.1 Future work

This project is relatively young, and we plan to extend our analysis in many directions.

Predicting violence in 2012-2014

We have held aside about half of our dataset in order to fine-tune our methods. We do this to guard against data-snooping and to increase confidence in results.

Disaggregating Violence

In our current dependent variable, we group several forms of violence that social scientists consider conceptually distinct. We are beginning to disaggregate these categories and believe that more careful consideration of specific types of violence will increase the relevance of this project to contemporary theories of conflict.

Predicting violence at the village level

Although the SNPK data is richer in terms of frequency and conflict classification, Podes has the advantage that it is measured at the village level. We plan to use this data to see how well we can predict village level events, and see if the best predictors of conflict at the village level are similar to those that perform best at the subdistrict level.

6.2 Concluding Remarks

This paper sets out to generate year ahead, sub-national predictions of violence that perform well. We have successfully improved upon earlier attempts in the conflict literature. Furthermore, the comparison of five machine learning methods and an aggregate method offers a useful guide for social scientists moving forward. Finally, we have put forth a principled method of assessing variable importance across vastly different prediction methods, and our analysis demonstrates the importance of considering explanatory power when identifying areas of further research.

References

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.

Barron, P., K. Kaiser, and M. Pradhan, "Understanding variations in local conflict: Evidence and implications from Indonesia," *World Development*, 2009, *37* (3), 698–713.

_ , S. Jaffrey, and A. Varshney, "How Large Conflicts Subside: Evidence From Indonesia," *Indonesia Social Development Paper*, 2014, (18).

Bazzi, S. and C. Blattman, "Economic shocks and conflict: Evidence from commodity prices," *American Economic Journal: Macroeconomics*, 2014, *6* (4), 1–38.

_ and M. Gudgeon, "Local Government Proliferation, Diversity and Conflict," *Unpublished Manuscript*, 2016.

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

Blair, R., C. Blattman, and A. Hartman, "Predicting Local Violence," *Unpublished Manuscript*, 2015.

Burke, Marshall, Solomon M Hsiang, and Edward Miguel, "Climate and Conflict," *Annu. Rev. Econ*, 2015, *7*, 577–617.

Dube, O. and J. F. Vargas, "Commodity price shocks and civil conflict: Evidence from Colombia," *The Review of Economic Studies*, 2013, *80* (4), 1384–1421.

Freund, Y. and R. E. Schapire, "A Short Introduction to Boosting," in "In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence" Morgan Kaufmann 1999, pp. 1401–1406.

Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* Springer Series in Statistics, New York, NY, USA: Springer New York Inc., 2001.

Henderson, J. V., A. Storeygard, and D. N. Weil, "Measuring Economic Growth from Outer Space," *American Economic Review*, 2012, *102* (2), 994–1028.

Michalopoulos, S. and E. Papaioannou, "The long-run effects of the scramble for Africa," forthcoming.

Montgomery, J., F. Hollenbach, and M. Ward, "Improving Predictions using Ensemble Bayesian Model Averaging," *Political Analysis*, 2012, *20*, 271–291.

Muchlinski, D., D. Siroky, J. He, and M. Kocher, "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data," *Political Analysis*, 2016, *24* (1), 87–103.

Strobl, C., T. Hothorn, and A. Zeileis, "Party on! A New, Conditional Variable-Importance Measure for Random Forests Available in the party Package," *The R Journal of Animal Ecology*, 2009, *Volume 1/2*, 14–17.

Tibshirani, R., "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 1994, *58*, 267–288.

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.

Yonamine, James, "Predicting future levels of violence in Afghanistan districts using GDELT," *Working paper*, 2014.

Appendix

A.1 Economic Shocks and Conflict

In table A.1 we report the results of regressing our indicator of any non-crime conflict on lagged rainfall and commodity price shocks. All regressions include province fixed effects. In all three regressions, we see that these lagged economic shocks are jointly significantly correlated with future conflict. However, these regressions we also see that the increases in R^2 from the inclusion of the shocks are minimal.

Table A.1: Regressions of Conflict on Lagged Economic Shocks

Shock:	Rainfall		Food Price		Cash Crop Price		Bauxite Price		Coal Price		Copper Price	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Lag:												
Lag 1	0.002	[.028]	-0.087	[.054]	0.037	[.021]*	1.764	[.621]***	-0.778	[.246]***	-0.261	[.528]
Lag 2	0.107	[.041]***	-0.099	[.048]**	-0.012	[.023]	1.094	[.805]	-0.778	[.328]**	0.219	[.5]
Lag 3	-0.016	[.039]	-0.101	[.061]	0.026	[.024]	-0.511	[1.19]	-0.594	[.402]	-0.991	[.527]*
F-test of shocks	0.069		0.158		0.176		0.036		0.015		0.271	
DV Mean	0.532		0.532		0.532		0.532		0.532		0.532	
R2	0.458		0.458		0.458		0.458		0.458		0.458	
R2 from FE only	0.457		0.457		0.457		0.457		0.457		0.457	

Shock:	Gold Price		Iron Ore Price		Nickel Price		Tin Price		Zinc Price		Silver Price	
	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
Lag:												
Lag 1	-0.521	[.643]	0.197	[.333]	0.362	[.432]	-8.865	[7.58]	-0.568	[.632]	-8.630	[6.32]
Lag 2	0.931	[.717]	-0.529	[.409]	-0.813	[.501]	-3.572	[6.11]	-0.166	[.486]	-2.309	[9.23]
Lag 3	0.634	[.800]	0.721	[.695]	-0.811	[.425]*	-15.506	[7.96]*	-0.319	[.519]	-11.244	[8.41]
F-test of shocks	0.461		0.046		0.000		0.110		0.745		0.247	
DV Mean	0.532		0.532		0.532		0.532		0.532		0.532	
R2	0.458		0.458		0.458		0.458		0.458		0.458	
R2 from FE only	0.457		0.457		0.457		0.457		0.457		0.457	

Each column represents a regression of an indicator for any non-crime event on lags of the given shock. All regressions include province fixed effects.

A.2 Data

A.2.1 Administrative Divisions

Subdistrict: 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 some amalgamated) over time. The drivers of this are less studied. There is a requirement of a minimum number of subdistricts for a district to split, so that drove some of it. All results in this paper keep subdistrict boundaries fixed at their 2000 level.

Table A.2 uses various data sources to document how the number of administrative units has changed over time. Discrepancies between early data sources are due to missing data points, but do not change the overall picture. 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.

Table A.2: Administrative Units over Time

	2000 (Census)	2000 (MFD)	2005 (MFD)	2014 (MFD)
Median Villages per Subdistrict	14	14	11	10
Median Villages per District	169.5	172	131	133
Median subdistricts per District	10	10	11	12
Mean Villages per Subdistrict	16.200	16.922	12.855	11.583
Mean Villages per District	190.924	202.394	159.014	158.358
Mean subdistricts per District	11.785	11.961	12.370	13.671
Number of Villages	64914	66790	70125	80921
Number of Subdistricts	4007	3947	5455	6986
Number of Districts	340	330	441	511

A.2.2 Indonesian National Violence Monitoring System (SNPK)

The conflict data comes from the Indonesian National Violence Monitoring System (known by its Indonesian acronym SNPK). The data are reported at the monthly level with 2011 subdistrict level that 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 subdistrictyear. We keep all violence categories with the exception of

domestic violence.

Any Incident: A dummy for whether SNPK recorded any violent incident in the given year, excluding domestic violence and crime.

Active Media: Using data obtained directly from SNPK managers on newspaper availability and usage by province and month, we calculate the number of papers used in any given province-year and include this as a predictor.

A.2.3 2000 Census

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

Employment Shares

We compute the 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**.

Demographic Variables

We compute the share of people in each of the following religions **Muslim, Catholic, Protestant, Hindu, Buddhist, Confucian, and Other**.

We compute the share of the population that is Chinese and the share Arab.

We compute the **share of men** and the **share married**.

In addition we calculate the share of persons under 10 and from 10 to 30 years old.

Education

We also have the share who completed no school and we also compute the average years of schooling.

Rural Population Share

We also have the share of villages within the subdistrict classified as rural.

Ethnic and Religious Diversity

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.
Relgious 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).

A.2.4 Night Lights

Light Intensity: We use 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.

A.2.5 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: We compute 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: We compute 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: We compute the 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.

A.2.6 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.

A.2.7 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.

A.2.8 Crop Shocks

Food Price Shocks: We use the 2003 Agriculture Census to 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: We use the 2003 Agriculture Census to 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.

A.2.9 Mineral Shocks

We have 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.

A.2.10 Disaster Information Management System

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

A.2.11 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.

A.2.12 Database for Policy and Economic Research (DAPOER)

We use 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.

A.2.13 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: We use data on vote share by party and subdistrict in the 1999 district parliamentary (DPRDII) electionsthe first of the post-Suharto erato 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.

A.3 Variable Transformations

For many of these variables, we put them in per capita terms. We take the natural logarithm of variables that are particularly skewed. Finally, for annual variables, we take several lags of the each variable, and to avoid high correlations between our predictors, we convert all but one lag to annual growth rates. Table A.3 reports summary statistics for each variable and reports the number of lags and transformations that we apply.

Table A.3: Predictors

	Mean	Median	Min.	Max.	Data Source	Logged	Changes	Lags	% missing	Imp. R^2
Violence										
Incidents [a]	1.251	0.68	0	74.812	SNPK			2	0%	0.10
Any incident	0.801	1	0	1.000	SNPK			2	0%	0.09
Incidents, no domestic violence [a]	1.119	1	0	65.232	SNPK			2	0%	0.10
Any incident, excluding domestic violence	0.777	1	0	1.000	SNPK			2	0%	0.09
Incidents, excluding crime and domestic violence [a]	0.321	0	0	15.981	SNPK			2	0%	0.15
Any incident, excluding crime and domestic violence	0.532	1	0	1.000	SNPK			2	0%	0.09
Incidents, Resource [a]	0.043	0	0	3.359	SNPK			2	0%	0.05
Incidents, Governance [a]	0.018	0	0	2.067	SNPK			2	0%	0.04
Incidents, Election [a]	0.021	0	0	4.942	SNPK			2	0%	0.03
Incidents, Identity [a]	0.013	0	0	4.463	SNPK			2	0%	0.04
Incidents, Popular Justice [a]	0.141	0	0	9.069	SNPK			2	0%	0.16
Incidents, Law Enforcement [a]	0.067	0	0	5.165	SNPK			2	0%	0.13
Incidents, Crime [a]	0.798	0.41	0	61.127	SNPK			2	0%	0.09
Incidents, Domestic Violence [a]	0.131	0	0	9.580	SNPK			2	0%	0.05
Incidents, Separatist [a]	0.001	0	0	1.284	SNPK			2	0%	0.04
Incidents, Other [a]	0.018	0	0	4.017	SNPK			2	0%	0.04
Deaths [a]	0.160	0	0	9.090	SNPK			2	0%	0.18
Deaths, excluding crime [a]	0.030	0	0	9.090	SNPK			2	0%	0.06
Deaths, Resource [a]	0.008	0	0	2.589	SNPK			2	0%	0.01
Deaths, Governance [a]	0.001	0	0	1.719	SNPK			2	0%	0.00
Deaths, Election [a]	0.000	0	0	0.379	SNPK			2	0%	0.00
Deaths, Identity [a]	0.002	0	0	8.585	SNPK			2	0%	0.01

Table A.3: Predictors

	Mean	Median	Min.	Max.	Data Source	Logged	Changes	Lags	% missing	Imp. R^2
Deaths, Popular Justice [a]	0.009	0	0	1.523	SNPK			2	0%	0.04
Deaths, Law Enforcement [a]	0.008	0	0	1.818	SNPK			2	0%	0.08
Deaths, Crime [a]	0.106	0	0	6.236	SNPK			2	0%	0.14
Deaths, Domestic Violence [a]	0.024	0	0	4.136	SNPK			2	0%	0.08
Deaths, Separatist [a]	0.001	0	0	3.494	SNPK			2	0%	0.03
Deaths, Other [a]	0.001	0	0	1.389	SNPK			2	0%	0.01
Injuries [a]	0.887	0	0	54.740	SNPK			2	0%	0.10
Injuries, excluding crime [a]	0.363	0	0	28.185	SNPK			2	0%	0.15
Injuries, Resource [a]	0.041	0	0	15.408	SNPK			2	0%	0.02
Injuries, Governance [a]	0.017	0	0	22.086	SNPK			2	0%	0.01
Injuries, Election [a]	0.014	0	0	13.576	SNPK			2	0%	0.01
Injuries, Identity [a]	0.019	0	0	10.297	SNPK			2	0%	0.02
Injuries, Popular Justice [a]	0.178	0	0	12.182	SNPK			2	0%	0.16
Injuries, Law Enforcement [a]	0.076	0	0	9.620	SNPK			2	0%	0.10
Injuries, Crime [a]	0.451	0	0	45.161	SNPK			2	0%	0.07
Injuries, Domestic Violence [a]	0.073	0	0	7.755	SNPK			2	0%	0.03
Injuries, Separatist [a]	0.001	0	0	2.821	SNPK			2	0%	0.01
Injuries, Other [a]	0.019	0	0	4.815	SNPK			2	0%	0.03
Damaged buildings [a]	0.163	0	0	331.565	SNPK			2	0%	0.01
Damaged buildings, excluding crime [a]	0.104	0	0	331.565	SNPK			2	0%	0.02
Damaged buildings, Resource [a]	0.055	0	0	331.565	SNPK			2	0%	0.01
Damaged buildings, Governance [a]	0.010	0	0	10.484	SNPK			2	0%	0.01
Damaged buildings, Election [a]	0.015	0	0	45.686	SNPK			2	0%	0.01

Table A.3: Predictors

	Mean	Median	Min.	Max.	Data Source	Logged	Changes	Lags	% missing	Imp. R^2
Damaged buildings, Identity [a]	0.010	0	0	11.287	SNPK			2	0%	0.01
Damaged buildings, Popular Justice [a]	0.012	0	0	9.423	SNPK			2	0%	0.01
Damaged buildings, Law Enforcement [a]	0.001	0	0	2.251	SNPK			2	0%	0.00
Damaged buildings, Crime [a]	0.057	0	0	74.617	SNPK			2	0%	0.00
Damaged buildings, Domestic Violence [a]	0.001	0	0	1.406	SNPK			2	0%	0.00
Damaged buildings, Separatist [a]	0.000	0	0	0.408	SNPK			2	0%	0.00
Damaged buildings, Other [a]	0.002	0	0	3.679	SNPK			2	0%	0.00
Destroyed buildings [a]	0.076	0	0	331.565	SNPK			2	0%	0.01
Destroyed buildings, excluding crime [a]	0.055	0	0	331.565	SNPK			2	0%	0.01
Destroyed buildings, Resource [a]	0.044	0	0	331.565	SNPK			2	0%	0.01
Destroyed buildings, Governance [a]	0.002	0	0	10.184	SNPK			2	0%	0.00
Destroyed buildings, Election [a]	0.002	0	0	9.289	SNPK			2	0%	0.00
Destroyed buildings, Identity [a]	0.003	0	0	10.791	SNPK			2	0%	0.00
Destroyed buildings, Popular Justice [a]	0.004	0	0	8.815	SNPK			2	0%	0.01
Destroyed buildings, Law Enforcement [a]	0.000	0	0	1.801	SNPK			2	0%	0.00
Destroyed buildings, Crime [a]	0.021	0	0	22.038	SNPK			2	0%	0.00
Destroyed buildings, Domestic Violence [a]	0.000	0	0	0.800	SNPK			2	0%	0.00
Destroyed buildings, Separatist [a]	0.000	0	0	0.000	SNPK			2	0%	0.00
Destroyed buildings, Other [a]	0.000	0	0	0.276	SNPK			2	0%	0.00
Total active papers [a]	25.3	19.7	0.7	348.6	SNPK			2	0%	0.92
Distance to nearest police post [a]	6.67	4.7	0	98.00	PODES			1	0%	0.00
Number of security posts [a]	14.86	12	0	111	PODES			1	1%	0.25
Number of police posts [a]	2.15	2	0	32	PODES			1	1%	0.34

Table A.3: Predictors

	Mean	Median	Min.	Max.	Data Source	Logged	Changes	Lags	% missing	Imp. R^2
Number of security officers [a]	304.1	254	0	2218	PODES			1	1%	0.15
Theft [a]	0.46	0.43	0	1	PODES			1	1%	0.37
Robbery [a]	0.04	0	0	1	PODES			1	1%	0.18
Looting incident [a]	0.01	0	0	0.76	PODES			1	1%	0.09
Thuggery [a]	0.07	0	0	1	PODES			1	1%	0.01
Arson [a]	0.01	0	0	1	PODES			1	1%	0.05
Murder [a]	0.03	0	0	1	PODES			1	1%	0.05
Fight among citizens [a]	0.02	0	0	1	PODES			1	1%	0.04
Fights with security officers [a]	0.00	0	0	0.20	PODES			1	1%	0.04
Conflict among students [a]	0.00	0	0	0.30	PODES			1	1%	0.01
Conflict between tribes [a]	0.00	0	0	0.32	PODES			1	1%	0.03
Indicator for any conflict [a]	0.03	0	0	1	PODES			1	1%	0.01
Deaths [a]	0.02	0	0	10.3	PODES			1	1%	0.07
Injuries [a]	0.25	0	0	47.2	PODES			1	1%	0.01
Property damage [a]	2300	0.000	0	4363003	PODES			1	1%	0.04
Terrorism (Count), district	0.00	0.00	-0.67	1.78	DIMS			3	0%	
Terrorism (Deaths), district	-0.01	0.00	-6.89	11.11	DIMS			3	0%	
Unrest (Count), district	0.01	0.00	-1.00	5.33	DIMS			3	0%	
Unrest (Deaths), district	-0.01	0.00	-5.33	3.56	DIMS			3	0%	
Population										
Population	10.50	10.60	4.54	13.00	2000 Census	X			2%	0.48
Population	10.64	10.64	7.38	13.25	SUSENAS	X			3%	0.38
Number of villages	18.15	15.00	1.00	132	2000 Census				1%	0.30

Table A.3: Predictors

	Mean	Median	Min.	Max.	Data Source	Logged	Changes Lags	% missing	Imp. R^2
Population	10.67	10.71	7.77	13.05	Podes	X		1%	0.42
Number of villages	19.34	15.00	3.00	155	Podes			1%	0.30
Religion									
Religious Fractionalization	0.144	0.036	0	0.741	2000 Census			2%	0.35
Religious Polarization	0.063	0.018	0	0.250	2000 Census			2%	0.36
Muslim share	0.772	0.961	0	1.000	2000 Census			2%	0.64
Catholic share	0.071	0.005	0	0.998	2000 Census			2%	0.55
Protestant share	0.000	0.000	0	0.000	2000 Census			2%	0.46
Hindu share	0.144	0.036	0	0.741	2000 Census			2%	0.30
Buddhist share	0.063	0.018	0	0.250	2000 Census			2%	0.11
Confucian share	0.772	0.961	0	1.000	2000 Census			2%	0.08
Any Islamic Boarding School [b]	0.005	0	0	0.532	PODES			1%	0.42
Ethnicity									
Ethnic Fractionalization	0.297	0.161	0.000	0.944	2000 Census			2%	0.09
Ethnic Polarization	0.086	0.072	0.0	0.2	2000 Census			2%	0.17
Chinese share	0.009	0.000	0.0	0.7	2000 Census			2%	0.14
Arab share	0.000	0.000	0.0	0.0	2000 Census			2%	0.03
Education									
Mean years of schooling	7.982	7.793	6.2	11.2	2000 Census			2%	0.23
No school share	0.346	0.331	0.1	0.7	2000 Census			2%	0.28
Junior high schools [a]	1.815	1.712	0.0	7.3	2000 Census			4%	0.18
High schools [a]	0.722	0.585	0.0	6.3	2000 Census			4%	0.07

Table A.3: Predictors

	Mean	Median	Min.	Max.	Data Source	Logged	Changes	Lags	% missing	Imp. R^2
Remoteness										
Rural population share	0.761	0.929	0.0	1.0	2000 Census				2%	0.23
Main Road Electrified [c]	0.645	0.769	0.0	1.0	PODES				1%	0.42
Road access year round [b]	1.048	1.000	0.0	2.0	PODES				1%	0.09
Telephone Signal Strength [b]	1.787	1.667	1.0	3.0	PODES				1%	0.37
Distance to province capital	81.4	67.3	0.3	465.5	GIS				0%	0.23
Distance to district capital	42.1	26.1	0.0	715.4	PODES				1%	0.27
Has airport [c]	0.005	0.000	0.0	0.6	2000 Census				4%	0.04
Has bridge [c]	0.558	0.583	0.0	1.0	2000 Census				4%	0.27
Has terminal [c]	0.049	0.000	0.0	0.8	2000 Census				4%	0.04
Has seaport [c]	0.036	0.000	0.0	1.0	2000 Census				4%	0.11
Land area	29545	11110	84.0	726129.0	2000 Census				4%	0.28
Demographics										
Male share	0.5	0.5	0.4	0.6	2000 Census				2%	
Married share	0.4	0.4	0.3	0.7	2000 Census				2%	
Share of population under 10	0.0	0.0	0.0	0.0	2000 Census				2%	
Share of population from 10 to 30	0.0	0.0	0.0	0.1	2000 Census				2%	
Health infrastructure										
Sophisticated health institutions [a]	0.933	0.517	0.0	11.1	PODES				4%	0.17
Mid level health institutions [a]	2.256	1.616	0.0	21.3	PODES				4%	0.45
Natural Disasters										
Number of mudslides [b]	0.076	0.000	0.0	1.0	PODES			1	1%	0.07

Table A.3: Predictors

	Mean	Median	Min.	Max.	Data Source	Logged	Changes	Lags	% missing	Imp. R^2
Number of floods [b]	0.208	0.097	0.0	1.0	PODES			1	1%	0.14
Number of earthquakes [b]	0.096	0.000	0.0	1.0	PODES			1	1%	0.41
Number of fires [b]	0.110	0.000	0.0	1.0	PODES			1	1%	0.18
Any other disaster [b]	0.058	0.000	0.0	1.0	PODES			1	1%	0.11
Land abrasions (Count), district	0.026	0.000	-0.9	4.4	DIMS			3	0%	
Land abrasions (Deaths), district	0.001	0.000	-0.3	2.7	DIMS			3	0%	
Disease outbreaks (Count), district	-0.015	0.000	-0.7	3.6	DIMS			3	0%	
Disease outbreaks (Deaths), district	-0.120	0.000	-7.6	7.1	DIMS			3	0%	
Droughts (Count), district	-0.138	-0.111	-3.6	6.2	DIMS			3	0%	
Droughts (Deaths), district	0.000	0.000	0.0	0.0	DIMS			3	0%	
Earthquakes (Count), district	-0.002	0.000	-0.9	3.6	DIMS			3	0%	
Earthquakes (Deaths), district	-28.557	0.000	-8644.9	5.3	DIMS			3	0%	
Floods (Count), district	0.343	-0.111	-5.7	17.3	DIMS			3	0%	
Floods (Deaths), district	0.750	0.000	-93.3	746.7	DIMS			3	0%	
Forest Fires (Count), district	-0.005	0.000	-0.6	1.8	DIMS			3	0%	
Forest Fires (Deaths), district	0.000	0.000	0.0	0.0	DIMS			3	0%	
Industrial Accidents (Count), district	0.010	0.000	-0.4	2.7	DIMS			3	0%	
Industrial Accidents (Deaths), district	0.020	0.000	-2.4	19.6	DIMS			3	0%	
Landslides (Count), district	0.120	0.000	-6.0	13.1	DIMS			3	0%	
Landslides (Deaths), district	0.476	0.000	-93.3	746.7	DIMS			3	0%	
Tornadoes (Count), district	0.116	-0.111	-3.7	13.3	DIMS			3	0%	
Tornadoes (Deaths), district	0.064	0.000	-6.7	53.3	DIMS			3	0%	

Table A.3: Predictors

	Mean	Median	Min.	Max.	Data Source	Logged	Changes	Lags	% missing	Imp. R^2
Transport accidents (Count), district	0.014	0.000	-0.8	6.2	DIMS			3	0%	
Transport accidents (Deaths), district	-0.142	0.000	-33.3	266.7	DIMS			3	0%	
Tsunami (Count), district	0.019	0.000	-0.9	4.4	DIMS			3	0%	
Tsunami (Deaths), district	-28.413	0.000	-8644.9	2.7	DIMS			3	0%	
Volcanoes (Count), district	0.004	0.000	-0.6	2.7	DIMS			3	0%	
Volcanoes (Deaths), district	0.000	0.000	0.0	0.0	DIMS			3	0%	
Deaths from any disaster, district	-55.921	-0.333	-17289.8	1493.3	DIMS			3	0%	
Total number of disasters, district	0.492	-0.222	-11.9	31.0	DIMS			3	0%	
Elections										
Mainstream party (Golkar and PDIP) share, 1999 local elections	0.555	0.579	0.01	0.95	DPRDII				4%	0.43
Islamist Party Share Local Par. Election 99 (kec level)	0.140	0.103	0.00	0.87	DPRDII				4%	0.42
Vote share frac. Local Par. Elec. 99 (kec level)	0.692	0.715	0.12	0.89	DPRDII				4%	0.37
Vote share pol. Local Par. Elec. 99 (kec level)	0.164	0.165	0.06	0.23	DPRDII				4%	0.31
Tot. Votes Local Par. Election 99 (kec level)	0.543	0.530	0.00	1.00	DPRDII				5%	0.13
Tot. Votes Nat. Par. Election (kab level)	0.475	0.511	0.00	1.00	DPRDII			1	1%	0.32
Mainstream party (Golkar and PDIP) share, kab elections	0.334	0.331	0.04	0.71	DPRDII			1	0%	
Islamist Party Share Nat. Par. Election (kab level)	0.164	0.146	0.00	0.49	DPRDII			1	0%	
Vote share frac. Nat. Par. Elec. (kab level)	0.841	0.859	0.55	0.94	DPRDII			1	0%	
Vote share pol. Nat. Par. Elec. (kab level)	0.114	0.111	0.06	0.19	DPRDII			1	0%	

Table A.3: Predictors

	Mean	Median	Min.	Max.	Data Source	Logged	Changes	Lags	% missing	Imp. R^2
Kabutapen election in year	0.352	0.000	0.00	1.00	DPRDII				0%	
Agricultural Infrastructure										
Mean access to irrigation [c]	0.40	0.30	0	1	PODES			1	2%	0.36
Small Rice Mills	0.19	0.07	0	1	PODES			1	1%	0.10
Large Rice Mills	0.07	0.00	0	1	PODES			1	1%	0.07
Finance and Trade Institutions										
Permanent formal market in village [c]	0.19	0.14	0	1	PODES			1	1%	0.23
Number of regular banks [b]	0.25	0.05	0	12	PODES			1	1%	0.49
Total number of rural cooperative banks [b]	0.09	0.00	0	3	PODES			1	1%	0.17
Economic Structure										
Major cash crop output share	0.16	0.04	0.00	1.00	2003 Agriculture Census				1%	0.25
Major crop output share	0.40	0.35	0.00	1.00	2003 Agriculture Census				1%	0.22
Major food crop output share	0.54	0.56	0.00	1.00	2003 Agriculture Census				1%	0.22
Share of HHs with Ag land >0.1 Ha	0.56	0.62	0.00	1.00	2000 Census				10%	0.36
Rice is the primary commodity, [c]	0.41	0.36	0.00	1.00	2000 Census				1%	0.20
Trading share	0.04	0.04	0.00	0.19	2000 Census				2%	0.37
Self-employed share	0.14	0.13	0.01	0.58	2000 Census				2%	0.17
Agriculture output share	0.29	0.31	0.00	0.67	2000 Census				2%	0.27
Services output share	0.07	0.05	0.00	0.29	2000 Census				2%	0.27
Employer share	0.09	0.07	0.00	0.36	2000 Census				2%	0.18
Population share in agriculture	0.21	0.19	0.00	0.63	2000 Census				2%	0.24
Population share in forestry	0.08	0.05	0.00	0.55	2000 Census				2%	0.17

Table A.3: Predictors

	Mean	Median	Min.	Max.	Data Source	Logged	Changes	Lags	% missing	Imp. R^2
Population share in industry	0.02	0.01	0.00	0.36	2000 Census				2%	0.20
Population share in trade	0.04	0.04	0.00	0.19	2000 Census				2%	0.37
Population share in service	0.06	0.05	0.00	0.28	2000 Census				2%	0.27
Population share in transport	0.01	0.01	0.00	0.06	2000 Census				2%	0.14
Share of HHs with any ag land	0.62	0.67	0.00	1.00	SNL				9%	0.32
Distance to nearest Bauxite mine	626.78	561.58	1.60	1692.78	SNL			3	0%	
Distance to nearest Coal mine	146.75	105.68	0.60	635.14	SNL			3	0%	
Distance to nearest Copper mine	97.44	84.19	1.87	362.23	SNL			3	0%	
Distance to nearest Gold mine	57.76	46.53	1.13	291.68	SNL			3	0%	
Distance to nearest IronOre mine	117.55	85.78	1.58	753.71	SNL			3	0%	
Distance to nearest Nickel mine	404.97	319.59	1.36	1233.87	SNL			3	0%	
Distance to nearest Tin mine	508.33	433.09	14.65	1812.89	SNL			3	0%	
Distance to nearest Zinc mine	311.14	302.94	3.84	805.68	SNL			3	0%	
Distance to nearest Silver mine	633.91	541.16	16.49	1539.42	SNL			3	0%	
Government Finances										
Total revenue (000s Rp)	13.05	13.04	8.99	17.48	2000 Census				4%	0.37
Total revenue within village resources	12.08	12.24	5.01	16.69	2000 census				4%	0.31
Revenue from district government	12.01	11.96	6.40	16.74	2000 census				5%	0.13
Routine Expenditures	12.36	12.42	5.01	17.00	2000 census				1%	0.32
Revenue from taxes	9.92	9.96	2.30	15.49	2000 census				0%	0.34
Revenue from social organizations	10.92	10.99	3.91	15.76	2000 census				0%	0.46
Revenue from ROSCAs	10.31	10.39	3.40	15.57	2000 census				0%	0.22
Revenue from other within village	9.47	9.28	3.91	16.65	2000 census				0%	0.31

Table A.3: Predictors

	Mean	Median	Min.	Max.	Data Source	Logged	Changes	Lags	% missing	Imp. R^2
Revenue from higher government admin	12.22	12.13	6.91	17.48	2000 census				0%	0.14
Revenue from central government	11.58	11.59	5.70	16.83	2000 census				26%	0.08
Revenue from provincial government	9.31	9.24	3.22	16.78	2000 census				51%	0.28
Revenue from district government	10.84	10.94	3.91	17.22	2000 census				19%	0.08
Total Revenue District [d]	13.95	13.89	11.57	16.38	DAPOER	X		3	10%	0.42
Total Non-Nat Resource Tax Sharing Rev [d]	11.30	11.30	8.94	14.22	DAPOER	X		3	5%	0.32
Self-Generated Kab Rev Rev [d]	11.00	10.96	7.94	13.38	DAPOER	X		3	5%	0.14
Total Nat Resource Tax Sharing Rev [d]	8.93	8.70	4.08	14.83	DAPOER	X		3	10%	0.52
Output										
Ag GDP District [d]	23.41	23.76	13.95	27.34	2003 Ag Census & FAO	X		3	7%	0.19
Total GDP District [d]	2.33	2.27	0.85	5.53	DAPOER	X		3	0%	0.42
Ag GDP District [d]	0.83	1.09	-3.89	2.45	DAPOER	X		3	0%	0.49
Median PCE	12.61	12.58	11.35	14.78	SUSENAS	X		3	3%	0.22
Mean PCE	12.74	12.70	11.42	14.82	DAPOER	X		3	3%	0.21
Nighttime light intensity	9.86	3.43	0.00	63.00		X		3	2%	0.43
Inequality										
GINI	0.24	0.23	0.00	0.77	SUSENAS			3	3%	0.06
PCE Inequality 90-10	1.03	1.01	0.00	2.89	SUSENAS	X		3	3%	0.06
PCE Inequality 80-20	0.66	0.64	0.00	2.13	SUSENAS	X		3	3%	0.06
Ag Wage, 75% - 25%	466816	350000	0	10760000	SUSENAS	X	X	3	8%	0.11
NonAg Wage, 75% - 25%	799873	750000	0	3350000	SUSENAS	X	X	3	4%	0.23
Rural Wage, 75% - 25%	626228	550000	0	3450000	SUSENAS	X	X	3	12%	0.27
Urban Wage, 75% - 25%	891698	870000	0	3960000	SUSENAS	X	X	3	13%	0.13

Table A.3: Predictors

	Mean	Median	Min.	Max.	Data Source	Logged	Changes	Lags	% missing	Imp. R^2
Ag Rural Wage, 75% - 25%	386181	300087	0	6350000	SUSENAS	X	X	3	16%	0.07
NonAg Rural Wage Spread, 75% - 25%	719090	650000	0	3400000	SUSENAS	X	X	3	12%	0.25
Labor Market									0%	0.00
Unemployment rate	0.04	0.03	0.00	0.43	2000 Census		X	3	6%	0.08
Ag Wage, median	13.00	12.96	10.82	15.23	SUSENAS		X	3	8%	0.38
NonAg Wage, median	13.48	13.49	11.70	14.91	SUSENAS		X	3	4%	0.36
Rural Wage, median	13.25	13.30	11.29	14.69	SUSENAS		X	3	12%	0.38
Urban Wage, median	13.55	13.55	11.70	14.91	SUSENAS		X	3	13%	0.31
Ag Rural Wage, median	12.92	12.90	10.57	15.39	SUSENAS		X	3	16%	0.39
NonAg Rural Wage, median	13.39	13.37	11.29	14.91	SUSENAS		X	3	12%	0.34
Commodity Price Shocks									0%	0.00
Food price shock	-0.026	-0.014	-0.553	0.690	2003 Ag Census & FAO			3	1%	0.01
Cash crop price shock	-0.015	0.000	-2.198	0.667	2003 Ag Census & FAO			3	1%	0.02
Major cash crop price shock	0.007	0.001	-0.376	0.258	2003 Ag Census & FAO			3	1%	0.01
Bauxite price shock	0.000	0.000	-0.233	0.173	SNL & GEM			3	0%	0.00
Coal price shock	0.001	0.001	-0.854	0.956	SNL & GEM			3	0%	0.00
Copper price shock	0.001	-0.001	-0.129	0.305	SNL & GEM			3	0%	0.00
Gold price shock	0.003	0.001	-0.017	0.239	SNL & GEM			3	0%	0.02
Iron Ore price shock	0.000	0.001	-0.397	0.342	SNL & GEM			3	0%	0.00
Nickel price shock	0.000	0.000	-0.478	0.344	SNL & GEM			3	0%	0.00
Tin price shock	0.000	0.000	-0.019	0.028	SNL & GEM			3	0%	0.00
Zinc price shock	0.000	0.000	-0.164	0.216	SNL & GEM			3	0%	0.00

Table A.3: Predictors

	Mean	Median	Min.	Max.	Data Source	Logged	Changes	Lags	% missing	Imp. R^2
Silver price shock	0.000	0.000	-0.001	0.025	SNL & GEM			3	0%	0.05
Commodity Price Shock History										
Food price shock	0.063	0.039	0.000	0.413	2003 Ag Census & FAO				1%	0.32
Cash crop price shock	0.054	0.055	0.000	0.138	2003 Ag Census & FAO				1%	0.19
Major cash crop price shock	0.022	0.004	0.000	0.166	2003 Ag Census & FAO				1%	0.25
Climate Shocks									0%	0.00
Rainfall shocks, district level	0.021	0.007	-0.295	0.566	GPCP			3	0%	0.03
Temperature, deviation from subdistrict average	0.001	0.001	-0.037	0.084	Udell			3	5%	0.20
Rainfall, deviation from subdistrict average	0.061	0.028	-0.970	0.723	Udell			3	5%	0.04
Climate History									0%	0.00
Rainfall, 1900-2010 average	182.611	180.324	79.817	327.446	Udell				5%	0.56
Temperature, 1900-2010 average	25.399	25.882	14.761	28.072	Udell				5%	0.21
Rainfall volatility, district level	0.174	0.171	0.088	0.334	GPCP				0%	

[a]per 10,000; [b]per town; [c]percent of towns; [d]per capita

A.4 Predictor Rankings

Below we report the top ten predictors for the 2007 LASSO, Random Forests, and AdaBoost models. For LASSO, coefficients are ranked by their absolute standardized coefficient values. For Random Forests and AdaBoost, variables are randomly permuted and ranked by the consequent decrease in mean squared error. We report importance scores for predictions using all variables and predictions excluding measures of past violence.

Table A.4: Predictor Rankings

LASSO		Random Forests		AdaBoost	
Predictor	β	Predictor	Imp.	Predictor	Imp.
(1)	(2)	(3)	(4)	(5)	(6)
Any Incident					
Population, 2000 Census	0.37	Injuries, non-crime, t	10.85	Population, 2000 Census	8.08
Crime incidents, t	0.32	Injuries, crime, t	9.81	Distance to province capital	4.27
Mean schooling, 2000 Census	0.25	Injuries, total, t	7.61	Crime incidents, t	3.76
Any non-crime event, t	0.20	Non-crime incidents, t	5.91	Injuries, total, t	3.27
Total incidents, t	0.19	Any non-crime event, t	5.21	Injuries, crime, t	3.06
Any non-crime event, t-1	0.17	Injuries, popular justice, t	3.75	Total active newspapers, t-1	2.67
Popular justice incidents, t	0.13	Population, 2000 Census	3.19	Hindu population share, 2000 census	2.36
Number of security posts, previous Podes	0.13	Popular justice incidents, t	3.13	Total crime incidents, t-1	2.29
Mean PCE, t-2	0.11	Transportation population share, 2000 census	2.94	Total incidents, t-1	2.25
Population growth, 2000-2002, Podes	0.11	Night lights, t-3	2.89	Non-crime incidents, t-1	1.70
Five or more incidents					
Five or more non-crime incidents, t	0.50	Five or more non-crime incidents, t	17.43	Five or more non-crime incidents, t	9.73
Mean schooling, 2000 Census	0.33	Five or more non-crime incidents, t-1	8.61	Population, 2000 Census	9.71
Five or more non-crime events, t	0.31	Distance to province capital	3.14	Mean schooling, 2000 Census	4.44
Percent with any farm holdings, 2000 census	-0.15	Mean schooling, 2000 Census	2.73	Distance to province capital	3.71
Services output share, 2000 Census	0.13	Agriculture output share, 2000 Census	2.30	Injuries, popular justice, t	3.51
Population, 2000 Census	0.09	Services output share, 2000 Census	1.22	Total incidents, t	3.00
Total votes in 1999 local election	0.08	Injuries, popular justice, t-1	1.00	Five or more non-crime incidents, t-1	2.75
Injuries, t-1	0.07	Night lights, t-3	0.78	Popular justice incidents, t	2.46
Total incidents, t	0.05	Population, 2000 Census	0.74	Rural population share, 2000 census	2.05
		Popular justice incidents, t	0.65	Injuries involving police, t-1	1.58

(a) Including Past Violence

LASSO		Random Forests		AdaBoost	
Predictor	β	Predictor	Imp.	Predictor	Imp.
(1)	(2)	(3)	(4)	(5)	(6)
Any Incident					
Population, 2000 Census	0.88	Population, 2000 Census	10.98	Population, 2000 Census	20.59
Population growth, 2000-2002, Podes	0.38	Rural population share, 2000 Census	7.16	Distance to province capital	7.30
Mean schooling, 2000 Census	0.31	Transportation pop. share, 2000 Census	5.27	Rainfall shock, t-2	2.68
Rainfall shock, t-2	-0.23	Perc. with any farm holdings, 2000 census	4.14	Perc. with any farm holdings, 2000 census	2.53
Total GDP per capita, t-2	0.17	Night lights, t-3	3.54	Hindu population share, 2000 Census	2.49
Distance to province capital	-0.17	Mean schooling, 2000 Census	2.30	Rural population share, 2000 Census	2.31
Pop. growth, last Podes to previous Podes	0.15	Services output share, 2000 Census	2.08	Night lights, t-3	2.08
Total votes in 1999 national election	0.13	Hindu population share, 2000 Census	1.66	Mean schooling, 2000 Census	2.07
Natural resource sharing income, t-3	0.13	Services population share, 2000 Census	1.65	Transportation pop. share, 2000 Census	1.69
Perc. with any farm holdings, 2000 census	-0.13	Total votes in 1999 local election	1.45	Agricultural GDP, t-3	1.25
Five or more incidents					
Population, 2000 Census	1.417	Distance to province capital	10.03	Population, 2000 Census	16.61
Mean schooling, 2000 Census	0.611	Mean schooling, 2000 Census	7.04	Distance to province capital	11.83
Population growth, 2000-2002, Podes	0.468	Agriculture output share, 2000 Census	6.49	Mean schooling, 2000 Census	7.88
Rainfall shock volatility	0.356	Population, 2000 Census	4.93	Hindu population share, 2000 Census	3.57
Perc. with any farm holdings, 2000 census	-0.321	Services output share, 2000 Census	3.74	Rainfall shock volatility	3.49
Islamist Party share in nat. parl. election, 1999	-0.316	Services population share, 2000 Census	2.20	Total votes in 1999 local election	2.96
Population growth, Susenas, t-3	0.273	Total votes in 1999 local election	2.09	Perc. with any farm holdings, 2000 census	2.39
Distance to province capital	-0.256	No schooling share, 2000 Census	1.61	Total rural cooperative banks, t	1.32
Rainfall shock, t-2	-0.216	Rural population share, 2000 Census	1.51	Gini, t-2	1.25
Catholic population share, 2000 census	-0.197	Perc. with any farm holdings, 2000 census	1.26	Rural population share, 2000 Census	1.25

(a) Excluding Past Violence

A.5 Forms of violence

For each of the 183,984 events from 2005 to 2014, we have data on the form that the violence took. Table A.5 reports this breakdown. Across violence categories, we see a large degree of heterogeneity in forms of violence.

Table A.5: Categories by form

	Crime	Law Enforcement	Resource Conflict	Governance	Election Violence	Identity Violence	Popular Justice	Separatist Violence	Other	Total
Assault	80,668	9,699	2,342	589	521	587	-	191	27	94,624
Blockade	-	4	55	31	14	4	39	5	12	164
Demonstration	-	5	387	1,169	407	24	65	13	50	2,120
Fights	7,374	907	469	93	94	160	-	38	5	9,140
Group Clash	-	64	884	150	197	2,245	1,531	142	551	5,764
Kidnapping	1,148	3	23	10	14	4	-	14	1	1,217
Lynching	-	118	407	138	147	431	25,276	13	2,048	28,578
Riot	-	-	54	29	35	45	336	1	19	519
Robbery	29,126	-	-	-	-	-	-	-	-	29,126
Sweeping	-	6	13	6	5	20	39	3	5	97
Terror Attack	-	-	58	24	105	92	84	13	361	737
Vandalism	7,447	76	1,507	1,057	1,382	383	-	29	17	11,898
Total	125,763	10,882	6,199	3,296	2,921	3,995	27,370	462	3,096	183,984

A.6 Conflict Literature

Table A.5: Recent articles investigating conflict

Cause/Correlate Investigated	Paper	Year	Author	Journal
Aid/Government Spending (4)	Can hearts and minds be bought? The economics of counterinsurgency in Iraq	2011	Eli Berman, Jacob Shapiro, and Joseph Felter	JPE
	Aid Under Fire: Development projects and civil conflict	2014	Benjamin Crost, Joseph Felter, and Patrick Johnston	AER
	US Food Aid and Civil Conflict	2014	Nathan Nunn and Nancy Qian	AER
	Conditional cash transfers, civil conflict and insurgent influence: Experimental evidence from the Philippines	2016	Benjamin Crost, Joseph Felter, and Patrick Johnston	JDE
Economic Resources (6)	Natural Resources: Curse or Blessing?	2011	Frederick van der Ploeg	JEL
	Ethnic cleansing or resource struggle in Darfur? An empirical analysis*	2013	Ola Olsson and Eyerusalam Siba	JDE
	Well, water, and welfare	2014	Sheetal Sekhri	AEJ: Applied
	Do giant oil fied discoveries fuel internal armed conflicts?*	2014	Yu-Hsiang Lei and Guy Michaels	JDE
	Resource concentration and civil wars*	2015	Massimo Morelli and Dominic Rohner	JDE
	Strategic Mass Killings	2015	Joan Esteban, Massimo Morelli, and Dominic Rohner	JPE
Economic Shocks (12)	Economic Shocks and Civil Conflict: A comment	2011	Antonio Ciccone	AEJ: Applied
	Re-examining Economic Shocks and Civil Conflict	2011	Edward Miguel and Shanker Satyanath	AEJ: Applied
	Oil and Conflict: What does the evidence really show?	2013	Anca Cotet and Kevin Tsui	AEJ: Macro
	Economic shocks and civil conflict: Evidence from foreign interest rate movements	2013	Peter Hull and Masami Imai	JDE
	Commodity Price Shocks and Civil Conflict: Evidence from Colombia	2013	Oeindrila Dube and Juan Vargas	REStud
	Economic Shocks and Conflict: Evidence from Commodity Prices	2014	Samuel Bazzi and Christopher Blattman	AEJ: Macro

Table A.5: Recent articles investigating conflict

Cause/Correlate Investigated	Paper	Year	Author	Journal
	Can cultural norms reduce conflicts? Confucianism and peasant rebellions in Qing China*	2014	James Kai-sing Kung and Chicheng Ma	JDE
	Do giant oil fied discoveries fuel internal armed conflicts?*	2014	Yu-Hsiang Lei and Guy Michaels	JDE
	What Do We Learn from the Weather? The New Climate-Economy Literature?	2014	Melissa Dell, Benjamin Jones, and Benjamin Olken	JEL
	Rainfall and conflict: A cautionary tale	2015	Heather Sarsons	JDE
	Economic shocks, civil war, and ethnicity*	2015	Thorsten Janus and Daniel Riera-Crichton	JDE
	External shocks, internal shots: The geography of civil conflict	2015	Nicolas Berman and Mathieu Couttenier	REStat
Ethnicity (8)	Horizonal Inequalities and Ethnonationalist Civil War: A global comparison*	2011	Lars-Erik Cederman, Nils Weidmann, Kristian Skrede Gleditsch	APSR
	Ethnicity and Conflict: An empirical study	2012	Joan Esteban, Laura Mayoral, and Debraj Ray	AER
	Social Identification and Ethnic Conflict	2013	Nicholas Sambanis and Moses Shayo	APSR
	Ethnic cleansing or resource struggle in Darfur? An empirical analysis*	2013	Ola Olsson and Eyerusalam Siba	JDE
	Ethnicity and the Spread of Civil War	2014	Maarten Bosker and Joppe de Ree	JDE
	Segregation and Conflict: An empirical analysis*	2015	Alejandro Corvalan and Miguel Vargas	JDE
	Resource concentration and civil wars*	2015	Massimo Morelli and Dominic Rohner	JDE
	Economic shocks, civil war, and ethnicity*	2015	Thorsten Janus and Daniel Riera-Crichton	JDE
Income Growth (1)	Implications of Economic Theory of Conflict: Hindu-Muslim Violence in India	2014	Anirban Mitra and Debraj Ray	JPE
Inequality (2)	Horizonal Inequalities and Ethnonationalist Civil War: A global comparison*	2011	Lars-Erik Cederman, Nils Weidmann, Kristian Skrede Gleditsch	APSR
	Income inequality and violent crime: Evidence from Mexico's drug war	2016	Ted Enamorado, Luis Lopez-Calva, Carlos Rodriguez-Castelan, and Hernan Winkler	JDE

Table A.5: Recent articles investigating conflict

Cause/Correlate Investigated	Paper	Year	Author	Journal
Infrastructure/Remoteness (2)	Tehnology and Collective Action: The effect of cell phone coverage on political violence in Africa	2013	Jan Pierskalla and FlorianHollenbach	APSR
	Propoganda and Conflict: Evidence from the Rwandan Genocide	2014	David Yanagizawa-Drott	QJE
Institutions/Trade (1)	Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia	2013	Saumitra Jha	APSR
Religion (2)	Can cultural norms reduce conflicts? Confucianism and peasant rebellions in Qing China*	2014	James Kai-sing Kung and Chicheng Ma	JDE
	Segregation and Conflict: An empirical analysis*	2015	Alejandro Corvalan and Miguel Vargas	JDE
Social Cohesion? (1)	How to Promote Order and Property Rights under Weak Rule of Law? An experiment in changing dispute resolution behavior through community education	2014	Christopher Blattman, Alexandra Hartman, and Robert Blair	APSR

Notes: * indicates that article appears more than once in the list because it examines correlates spanning multiple categories.

Table A.6: Placebo test of economic shocks

	Full set of predictors (1)	Partial out shocks (2)	Partial out permuted shocks (3)
AUC, predicting any event	0.784	0.584	0.663

A.7 Investigating the effect of partialing out economic shocks

While economic shocks alone have little predictive power and removing economic shocks from the set of predictors does not hurt performance, we do observe a large drop in performance when we partial out the economic shocks from the remaining set of predictors. Our concern is that because our panel is short and economic shocks may be serially correlated, this partialling out effect may be a statistical artifact, rather than an indication that much of the information from the economic shocks is reflected in all of these other variables, that that is the source from which they derive their predictive power. Table A.6 reports the predictive performance for the full set of predictors (excluding past violence) predicting any event, as well as predictive performance when we partial economic shocks out of all of the other variables. Column (3) reports a placebo test constructed as follows. We randomly permute the full set of shocks and lagged shocks across subdistricts, but we keep subdistrict groups of shocks and lagged shocks entact. Hence, were eliminate any meaningful correlations between shocks and other subdistrict variables, but we maintain the intershock correlations as well as the serial correlation structure. We then partial out these permuted shocks.

We see that partialling out the permuted shocks leads to a substantial drop in performance, although not as large as when we partial out the actual shocks. Regardless, this casts suspicion on the results in which we partial out commodity price and weather shocks.