Reading Between the Lines: Prediction of Political Violence Using Newspaper Text*

Hannes Mueller (IAE(CSIC), Barcelona GSE)
Christopher Rauh (University of Cambridge, INET Institute)

January 29, 2016

Abstract

This article provides a new methodology to summarize newspaper text using topic models. Using this method we predict the outbreak of violent conflict one and two years before it occurs. In our prediction we distinguish between predicting differences in the likelihood of conflict between countries and the timing of conflict within each country. Our analysis shows that our generated news data has a comparative advantage in predicting the timing of conflict, i.e. the within country variation. We argue that this makes news data a particularly useful addition to the most common conflict predictors used in the literature.

*We thank Stephen Hansen and the participants of the ENCoRe Barcelona meeting and the internal Political Economy Workshop Cambridge for valuable feedback. We are grateful to Alex Angelini, Lavinia Piemontese and Bruno Conte Leite for excellent research assistance. We thank the Barcelona GSE under the Severo Ochoa Programme for financial assistance. All errors are ours.

1 Introduction

The conflict literature has made significant progress in understanding which countries are more at risk of suffering from political violence (e.g., Blattman and Miguel 2010, Goldstone et al. 2010, Fearon and Laitin 2003, Esteban, Mayoral and Ray 2012, Besley and Persson 2011). However, many factors that have been identified as leading to increased risk, like mountainous terrain or ethnic polarization, are time-invariant or very slow-moving and therefore not useful in predicting the timing of conflict. Others, like GDP levels or political institutions, still show much larger variation across countries than within countries over time. So the question *when* a country is at risk remains.

The problem in forecasting the timing of violent conflict is that it is rare and at the same time relatively concentrated in some countries. It is straightforward to illustrate why this is a potential problem. Imagine that at the end of the second world war one had attributed a prediction of conflict to all years in some countries and the prediction of no conflict to all years of the remaining countries. This forecast would have been able to reach a striking level of accuracy of over 90 percent.[1] However, such a model would be completely time invariant and, hence, tell us nothing about the timing of conflict. This is a problem because it implies that the between variation can dominate the analysis unless the between and within variation are separated explicitly. Empirical models that are overall quite accurate can therefore be nonetheless of little use on the time dimension. We show, using a simple panel regression model, that many variables commonly used in the literature indeed face this problem. Yet, policymakers and academics alike might be interested in a measure of conflict risk that is meaningful on the time dimension.

This is one of the reasons why the forecasting literature has embraced the use of data generated from news sources.[2] News content has strong country-specific elements and is available in real time - both of these elements give it a comparative advantage in predicting the time-dimension of conflict. To this end we propose a new, fully automated, method to quantify the content of news using the latent Dirichlet allocation (LDA) model (see Blei, Ng and Jordan 2003), which we apply to over 600,000 newspaper articles from English-speaking newspapers. First, we show that it is possible to summarize news text in a meaningful way in a topic model. More surprisingly, perhaps, the topics we estimate seem to provide meaningful summaries of text across countries and decades. We find that changes in topics that predict political instability in some countries also predict it in other countries out-of-sample. Our news data adds relatively little to the identification of conflict in standard cross-sectional logit models but improves the prediction of the timing of conflict dramatically. It can be implemented with only minimal personal judgement and appears to be a useful addition to other news models.

Our empirical methodology proceeds in three steps. We first download news-

[1] Ward et al. (2013) make a similar argument for using a hierarchical model.

[2] See, for example, Chadefaux (2014) and Ward et al. (2013) which we discuss in detail below.

paper articles from LexisNexis and collect words and series of words, referred to as tokens, in one vector for each article. Newspaper articles offer several advantages for the analysis of conflict. They stretch several decades and report on events in all countries, which means that even rare events are sufficiently common to be analyzed with quantitative methods. Also, newspapers provide a high degree of consistency regarding the density of news per year that can be reported. This makes token counts at least somewhat comparable across years and even decades. We downloaded all articles on 185 countries from the New York Times, the Washington Post, and the Economist for all available years since the 1970s. This gives us a basis of 800,000 newspaper articles with about a million unique word combinations, even after excluding stop words, rare words and stemming.

As a second step, we develop a topic model tailored for the purpose of summarizing the content of news reports in a country and year. We use the LDA model to generate quantitative summaries of the articles. In this way, the high dimensionality of token vectors ($\sim$ 1 Million) can be decreased to as many topics as we choose. The main advantage of this methodology is that we do not need to impose any prior on which part of the text is important when predicting conflict - we can let the data speak.

As the final step, we use the emergence and disappearance of topics on the country level to predict conflict out-of-sample. For this step, we calculate the share of words written on each topic in every country and year. We then use these topic shares to predict conflict in the following one or two years. Topic models have one particular advantage in the context of predicting conflicts out-of-sample. Since topics are a collection of words that co-occur, the model is valid across countries and time. Insurgencies, for example, will trigger certain keywords that are shared across all countries and times even if the specific context differs.

We proceed as follows. We first discuss related literature in Section 2. In Section 3 we argue that linear fixed effects regressions can be quite useful to distinguish between location and timing in forecasting. In Section 4 we present our methodology of aggregating news text into topics and the forecasting method in which this information is used. Section 5 presents results.

2 Related Literature

The academic literature has made large strides towards understanding the triggers of civil conflict. The economics literature has typically focused on establishing links to specific factors like ethnic cleavages (Reynal-Querol and Montalvo 2005, Esteban, Mayoral and Ray 2012), climate (Miguel, Satyanath and Sergenti 2004, Dell, Jones and Olken 2012) or natural resources (Brückner and Ciccone 2010, Bazzi and Blattman 2014). Forecasting in this literature is not a priority and out-of-sample tests with new data on climate or commodity prices revealed that some pre-existing findings are not robust to changes in the sample

or method.[3]

Models that contain several factors at once are more common in political science.[4] This has also led to a much bigger openness towards forecasting and a quickly growing literature. The articles most closely related to our work are Chadefaux (2014), Ward et al. (2013), Goldstone et al. (2010) and Rost, Schneider and Kleibl (2009). Rost, Schneider and Kleibl (2009) use cross-sectional logit regressions on economic and political variables as well as proxies for violations of human rights to predict conflict onset within a 5-year window. They find substantial predictive power of their model within this time-frame. Goldstone et al. (2010) provide predictions of political instability at the country level within a two-year horizon. Their statistical method compares country/years before instability to country/years in the same region that were not followed by onset. Their main finding is that the best predictors of instability are slow-moving variables like political institutions or infant mortality. Ward et al. (2013) use a combination of event data and more standard variables to make monthly predictions up to six months ahead. Their model has a striking degree of accuracy in predicting several kinds of conflict incidence and performs well out-of-sample. Chadefaux (2014) relies on keyword counts of a list of predetermined words to construct an index of tension on a weekly basis for the period 1902 to 2001. He uses the constructed tension data to predict onset of conflict weeks before they occur and shows that news data can contribute significantly to a standard model.

We add to this research in three ways. First, we attempt to predict conflict with news data up to two years before conflict occurs. The longer time period compared to other studies which use news sources reduces our sample size and makes out-of-sample prediction particularly hard. Second, the data we use is generated without any prior assumption on the words that could predict conflict, i.e. our predictions use the entire newspaper text written on each country. Finally, an important conceptual contribution of this project is that our attempt to build forecasting models is geared towards within country variation. In other words, we explicitly focus on predicting the timing of the occurrence of violence. This is an important distinction to existing studies, which do not focus on the time dimension.

Another part of the literature has tried to predict conflict locations within ongoing conflicts. Here the problem of predicting timing is alleviated as it is already clear that conflict risk in the baseline is fairly high. Research can therefore focus on distinguishing the determinants of risk in the cross-section. Blair, Blattman and Hartman (2014) use 56 potential risk factors to predict locations of conflict within Liberia. They find that especially ethnic diversity and polarization, two slow-moving variables, consistently predict the location of violence over time. Similarly, Schutte (2014) predicts location of conflict using structural factors like population or the distance to the capital. An interesting exception are studies which predict the timing of local violence using violence

[3] Findings regarding ethnic cleavages cannot be used to predict the timing of conflict as the key variables are too slow moving.

[4] The seminal paper by Fearon and Laitin (2003) is a case in point.

in neighbouring geographic units. Weidmann and Ward (2010), for example, show that a reliable predictor of violence in a given period is violence in near regions in the previous period.

The LDA has been used by Hansen, McMahon and Prat (2014) to quantify discussions in the central bank committee of the Bank of England. The approach has the big advantage of requiring no human input except for the choice of parameters of a distribution. We contribute to this literature by applying the same methodology as Hansen, McMahon and Prat (2014) to news text and by using the estimated topic shares in cross-country panels. The, perhaps surprising, finding is that topics can be used in a meaningful way across countries and years. We also find that fewer topics generally perform better in out-of-sample predictions. However, within our method, the approach we use is the simplest possible - it should be seen as a first step in the conflict literature. It would, for example, be possible to use a structural topic model as suggested by Roberts et al. (2013) in which covariates are embedded in the topic generation. The appeal of the method we choose here is that it is fully automatic except for the choice of number of topics and two additional parameters.

News sources have previously been used to gather data on expectations and perceptions. Kuziemko and Werker (2006) use the frequency of the United Nations and the Security Council being mentioned in the New York Times as a proxy for its political importance. Ramey (2011) shows that increases in military spending can be predicted through news reports several quarters before they occur. Brückner and Pappa (2015) show that news on the Olympic Games, not the Games themselves, drive investment in countries that host them. The approach chosen in these studies is possible because there is a clear prior regarding which news reports should capture perceptions. An exception in this literature is Gentzkow and Shapiro (2010), who develop a measure of political bias of newspaper outlets in the US. To do this, they generate a list of expressions that indicate Republican or Democratic slant. They generate this list by looking at what expressions distinguish political speeches by Republicans and Democrats. Our methodology follows this basic idea but instead tries to understand how news text changes in the years before political instability compared to other years.

3 Forecasting the Timing of Conflict

In this section, we show that the main difficulty of forecasting conflict is in forecasting the timing. We show this in two steps. First, we argue that separating the variation between countries from the variation within a country is essential to understanding the timing of conflict. We then use a linear fixed effects model to give an example in which a set of variables is quite good at forecasting the between-country variation of conflict, but fails completely to forecast the within-country variation of conflict.

3.1 The Time Dimension of Conflict

Our aim is to train our model to forecast by comparing observations in country i and year t that were followed by conflict within one or two years (treatment) to observations which did not experience conflict later (control). One way to do this is by regressing a dummy y_{it} that indicates two years before conflict on a set of country characteristics x_{it}. The most standard way to do this is the logit model, which has the formal representation

$$\Pr(y_{it} = 1) = F(\alpha + x_{it}\beta) \tag{1}$$

where $\Pr(y_{it} = 1)$ is the probability of observing conflict within two years and F is the cumulative logistic distribution. This is conceptually equivalent to a linear probability model of the form

$$y_{it} = \alpha + x_{it}\beta + \varepsilon_{it}. \tag{2}$$

Such models exploit the deviation of countries from the average propensity of conflict, $y_{it} - \bar{y}$, and contrasts this with their deviation from the average value for the explanatory variables $x_{it} - \bar{x}$. In fact, this encompasses learning about pre-cursors to conflict in two ways: from the differences between countries and from what happens across time within a country. Formally we can write the estimated coefficients β in terms of the two sums of squares which capture the two sources of variation

$$\begin{aligned} \beta \quad = \quad & \left[\sum_i \sum_t \left(x_{it} - \bar{x}_i\right) \left(y_{it} - \bar{y}_i\right)' + \sum_i T \left(\bar{x}_i - \bar{x}\right) \left(\bar{y}_i - \bar{y}\right)' \right] \\ & * \left[\sum_i \sum_t \left(x_{it} - \bar{x}_i\right) \left(x_{it} - \bar{x}_i\right)' + \sum_i T \left(\bar{x}_i - \bar{x}\right) \left(\bar{x}_i - \bar{x}\right)' \right]^{-1} \end{aligned} \tag{3}$$

where $\bar{x}_i$ and $\bar{y}_i$ are the country means of x_{it} and y_{it} respectively. It will be useful to inspect the elements of equation (3) separately. The first term in the nominator, the within sum of squares,

$$SS_{xy}^{within} = \sum_i \sum_t \left(x_{it} - \bar{x}_i\right) \left(y_{it} - \bar{y}_i\right)' \tag{4}$$

describes the association of x_{it} and conflict y_{it} across time within a country, i.e. the *within variation*. The second term in the nominator

$$SS_{xy}^{between} = \sum_i T\left(\bar{x}_i - \bar{x}\right)\left(\bar{y}_i - \bar{y}\right)' \tag{5}$$

describes the association of the average $\bar{x}_i$ and average conflict $\bar{y}_i$ between countries, i.e. the *between variation*. Variation in x_{it} affects both the within and between sum of squares in equations (4) and (5). A change in a given year first immediatly affects x_{it} in (4). If x_{it} changes permanently in a country this implies a change in $\bar{x}_i$.

The split in equations (4) and (5) can be used to illustrate why the vector x_{it} can become a predictor of conflict. The first option is that x_{it} deviates from its mean $\bar{x}_i$ exactly when conflict becomes imminent, i.e. when $y_{it} = 1$. An economic crisis might, for example, trigger conflict which would imply that a deviation of x_{it} from $\bar{x}_i$ coincides with a change of y_{it} from 0 to 1. The second option is that countries whose average characteristics $\bar{x}_i$ deviate from the overall average $\bar{x}$ are more (or less) likely to enter conflict than the average country. Political institutions, which differ vastly between countries but change rarely, have often been found to predict conflict in the cross-section.

Learning from $SS_{xy}^{between}$ exclusively means that the timing of instability cannot be predicted. The between variation allows us to rank countries according to their likelihood of destabilisation but these rankings will not change from one year to the next. This also means that countries which have been free of violence in the past are likely to be "off the radar". It is therefore questionable whether a model that relies on large values of $SS_{xy}^{between}$ alone can be used effectively to forecast conflict in the long run.

In many applications a policymaker might be interested in whether a particular country is more likely to enter a crisis than a year before. This is especially important if sudden developments change established dynamics in a country. Learning about the timing of instability is only possible from the within variation, SS_{xy}^{within}, in equation (4). From the equation we can see that countries which remain in peace or only experience episodes with $y_{it} = 1$ throughout the sample cannot contribute to learning about timing as $y_{it} = \bar{y}_i$ in all years. This might be a subtle point but it turns out to be of crucial importance in the actual application because it prevents the use of the fixed effects logit framework to predict conflict in previously peaceful countries.

Our discussion should make clear that, apart from being of interest for policymakers, distinguishing between within and between variation is important for an academic audience interested in learning about the drivers of conflict. The problem in establishing causality from cross-country panel regressions is well known. This is particularly true, however, if a variable predicts conflict in the between variation but fails to do so with its within variation. If within and between variation point in different directions, it is unlikely that we have found a causal driver of conflict. If a variable predicts conflict between and within countries and is also able to forecast conflict out-of-sample then the variable is more likely linked to a causal mechanism.

It is important to note that, while it is easier to illustrate these points in a linear probability model, the same logic applies to non-linear models like the logit model. Also here, identification comes from differences across countries and changes across time within countries.

3.2 An Illustration Using Standard Variables

The previous discussion made clear that regressions like in equation (2) mix the between and within variation to estimate the parameters β. A straightforward way to separate out the within from the between variation is the fixed effects model

$$y_{it} = \beta_i + x_{it}\beta^{FE} + \varepsilon_{it}. \tag{6}$$

We will stick to the linear probability model here because it has the advantage of producing estimates for the parameters β in countries which have only $y_{it} = 1$ or $y_{it} = 0$. If we want to estimate the probability of future conflict in a country that has remained stable this is a particularly important property. Without assuming a linear probability model forecasting would need to be restricted to countries which already descended into conflict once. The parameter β^{FE} vector in this model is estimated as

$$\begin{aligned} \beta^{FE} &= \left[\sum_i \sum_t \left(x_{it} - \bar{x}_i\right)\left(y_{it} - \bar{y}_i\right)'\right] \\ &\quad * \left[\sum_i \sum_t \left(x_{it} - \bar{x}_i\right)\left(x_{it} - \bar{x}_i\right)'\right]^{-1}. \end{aligned}$$

In other words, the model relies entirely on the within variation. This prevents us from making any conjectures regarding the transition to conflict from either countries which are always in conflict or peace. Instead, the learning process about what constitutes a risk to stability is gathered exclusively from countries that have destabilized. In this model, the between variation is captured by a set of country-dummies. It is important to stress that the variation captured by the estimated fixed effects β_i can be regarded as a measure of our ignorance or lack of data in a forecasting model. If the vector x_{it} captures both the within and the across variation in the danger of conflict then little variation is left for the fixed effects to explain.

We now show that data commonly used in standard forecast models generate surprisingly little useful within variation for forecasting conflict. For variables like historic ethnic polarization or geography this is obviously true as they are time invariant or extremely slow moving. Other variables like GDP per capita or political institutions change across time and are therefore potentially useful. To illustrate our approach, we use a model which uses a set of standard variables in the conflict literature: four political regime dummies from polity IV, infant mortality, the share of the population that is discriminated against and a dummy that captures whether more than three neighbouring countries had an armed

conflict.[5] We combine this data with data from UCDP/PRIO on battle-related deaths. For now, we define conflict as the incidence of more than 1000 battle-related deaths linked to an internal conflict.[6] We have data on all these variables for 141 countries in the years 1975 to 2010.

For illustrative purposes, we use these variables as x_{it} in equation (6) and use the fitted model to forecast conflict out-of-sample. To do so we proceed in two steps. First, we run the regression in equation (6) on the full sample of countries but only for the years 1975-1995. On the left hand side we have a dummy variable that indicates a year preceding conflict onset, excluding all conflict years. Second, we calculate fitted values

$$\hat{y}_{it} = \hat{\beta}_i + x_{it}\hat{\beta}^{FE}$$

out-of-sample, i.e. for the period 1996-2010 and use them to forecast the indicator of a year before onset, y_{it}, in the sample 1996-2010.

Figure 1 demonstrates the overall performance of the forecast. The graph orders all observations in the sample 1996-2010 according to their value of the forecasted probability $\hat{y}_{it}$, which is also displayed as a thick red line. The fitted values are then contrasted with the actual observation of $y_{it} = 1$ which is displayed as vertical thin black bars. The model performs quite well, i.e. most of the black bars are towards the right hand side of the Figure and become more and more concentrated as $\hat{y}_{it}$ increases. High forecast values of $\hat{y}_{it}$ are clearly associated with conflict years. It is important to stress that this is out of sample, i.e. forecasts after 1996 are based on data before that year.

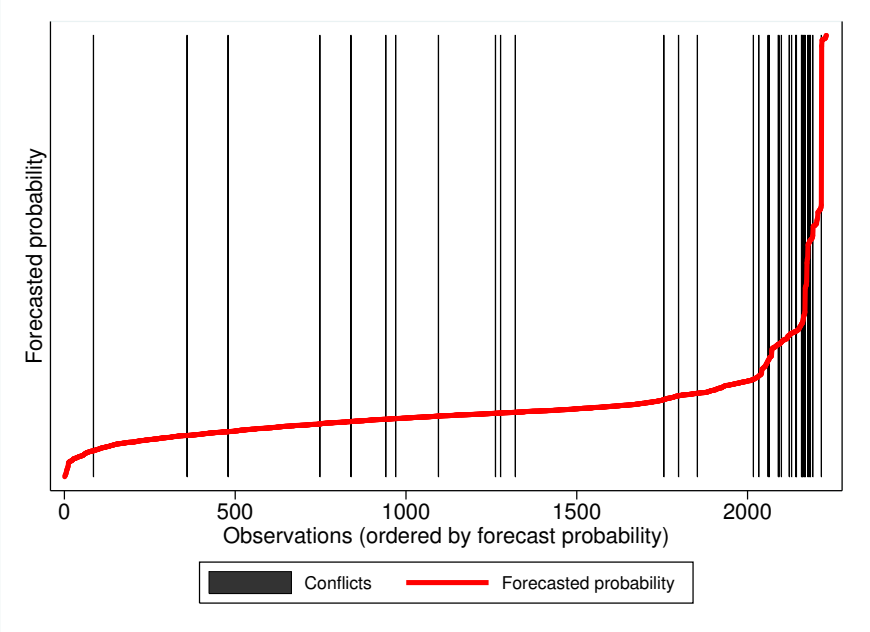
The fixed effects regression we have run here allows us to dissect the forecasted probability $\hat{y}_{it}$ into its between and within components. More specifically, the estimated fixed effects $\hat{\beta}_i$ give an idea of how much variation in the forecast is actually time-invariant and not picked up by the time variation in the explanatory variables, x_{it}. Forecasting with $\hat{\beta}_i$ relies entirely on the propensity of different countries to descend into conflict, i.e. the between variation. For example, forecasting in this model consist of attributing a relatively high constant probability of conflict to Lebanon and relying on the fact that conflict in Lebanon is indeed relatively likely. If, instead, the fitted values $x_{it}\hat{\beta}^{FE}$ explain most variation in conflict, we expect the estimated fixed effects $\hat{\beta}_i$ to be less important for forecasting purposes.

Figures 2 and 3 order observations by the estimated values of $\hat{\beta}_i$ and $x_{it}\hat{\beta}^{FE}$ respectively. The red line in Figure 2 displays the estimated values of $\hat{\beta}_i$ and contrasts them with observations of conflict in the period 1996-2010. The change from Figure 1 to Figure 2 is relatively subtle. Conflict is concentrated in countries with relatively high values of $\hat{\beta}_i$, i.e. the model predicts conflict through

[5] This is roughly the model suggested by Goldstone et al. (2010). But the selected model should be regarded as an illustration more than anything else. As far as we can tell all our observations are robust to a lot of different sets of variables such as the polity2 score, log population, GDP per capita in logs and GDP per capita growth, the share of population excluded from political power. The only exception is a polynomial of the time since the last armed conflict which we discuss in more detail below.

[6] We include conflicts of type 3 and 4. Our results are robust to changes in this definition, i.e. they are robust to including more or less types of conflict and different thresholds.

Figure 1: Forecasting Conflict Out-of-Sample Using Standard Variables



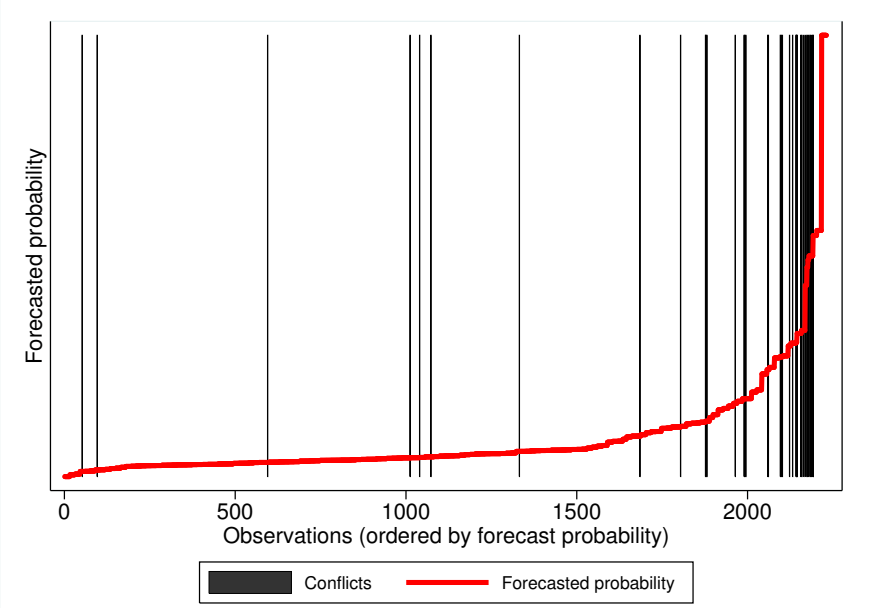
Note: Forecast probability is the predicted value from a fixed effects regression in the years 1975-1995. Observations are from the years 1996-2010 and are ordered by the fitted value (red line). Black lines indicate a year that actually preceded violence by one year. Violence is defined as more than 1000 battle related deaths.

factors that are fixed in time. This already hints at the fact that our forecast in Figure 1 was a forecast that relied on the between variation.

Accordingly, the picture changes dramatically if we use only the within variation contained in $x_{it}\hat{\beta}^{FE}$ in Figure 3. Years before conflict are now distributed across all values of $x_{it}\hat{\beta}^{FE}$. Overall, the within variation is therefore not very telling regarding the incidence of conflict. The model fails to predict the timing of conflict.

While this could be an extreme example we will show below that there is a general tendency of standard variables to predict conflict through between variation. Forecasting models based on these variables provide a good idea which countries are more likely to enter conflict. But they do much worse in predicting the timing of conflict. It is in this context that the utility of data constructed from news sources should be seen.

Figure 2: Between Variation Using Standard Variables is Useful for Forecasting Conflict



Note: Forecast probability is the fixed effect from a fixed effects regression regression in the years 1975-1995. Observations are from the years 1996-2010 and are ordered by the fitted value (red line). Black lines indicate a year that actually preceded violence by one year. Violence is defined as more than 1000 battle related deaths.

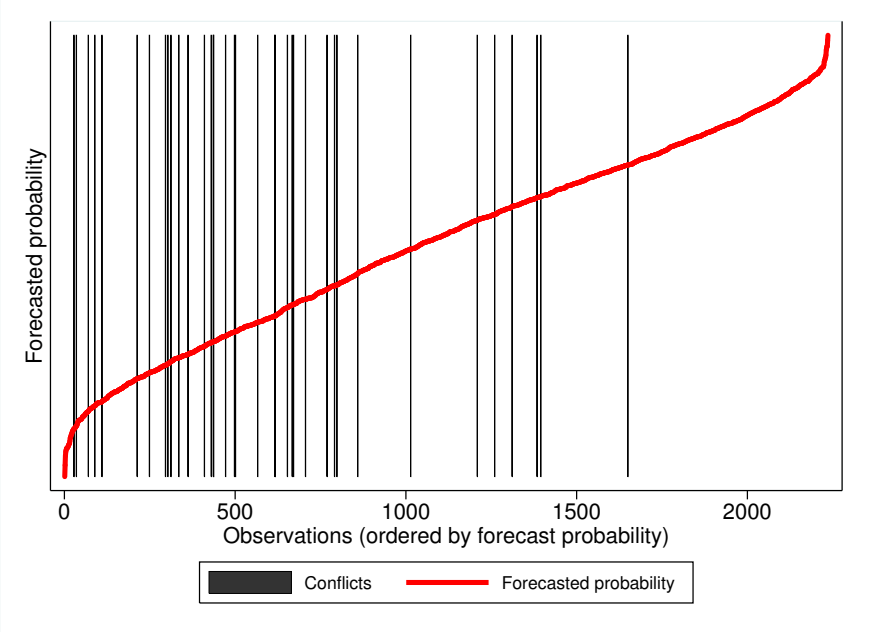
4 A Topic Model of Newspaper Text

News text needs to be transformed into low dimensional data before it can be used effectively in forecasting. Without a strong prior regarding the relevant keywords this task is typically impossible because the researcher faces millions of tokens. We rely on recent developments in text analysis tools to reduce the dimensionality.

4.1 News Text

The first choice we face is the selection of our news sources. Due to their availability over a long time span and international coverage, we focus on three major newspapers published in English, namely the Economist (available from 1975), the New York Times (NYT) (available from 1980), and the Washington Post (WP) (available from 1977). From the database LexisNexis we downloaded all articles dating from 1975 to 2010 containing country names (or slight permuta-

Figure 3: Within Variation Using Standard Variables is Less Useful for Forecasting



Note: Forecast probability is the predicted value from a fixed effects regression, excluding the fixed component, in the years 1975-1995 Observations are from the years 1996-2010 and are ordered by the fitted value (red line). Black lines indicate a year that actually preceded violence by one year. Violence is defined as more than 1000 battle related deaths.

tions thereof) or capital names in the title.[7] In total, we downloaded 633,835 articles, of which 152,102 are from the Economist, 310,673 from the NYT, and 167,932 from the WP.

In the average country/year 126 articles are written on a country. However, the extent of coverage varies drastically with the type of country so that we observe between 1 and more than 5500 articles in a given year. As a general idea, more populous, richer and more democratic countries are covered more. In addition, coverage increases in and before conflict. On average, a conflict year is covered with about 100 articles more while a pre-conflict year is covered with almost 70 articles more than the average year.[8] However, total news articles in our dataset are fairly constant across time as shown in Figure A.1 in the Appendix. This is quite intuitive. Total space for articles is fairly constant

[7] In the case of the Economist we also search in the leading paragraph as the title rarely contains a country or capital name.

[8] The findings regarding conflict come from simple OLS regressions. The findings on conflict are robust to the introduction of country fixed effects. Results available from the authors on request.

and so attention seems to shift towards countries which seem newsworthy. Our methodology accounts for these changes in coverage by using topic shares, i.e. we disregard how much is written on a country and focus instead on what is written on a country. This is important as it facilitates forecasting across countries.

In order to improve the performance of the machine learning algorithm we apply, we process the raw texts of articles of all three newspapers according to standard text mining procedures. First, we remove a library of common words, which in text mining are referred to as stop words, such as "to" or "that".[9] Second, we lemmatize and then stem words using the Snowball algorithm, which is an updated version of the algorithm from Porter (1980).[10] Lemmatizing groups variant forms of the same word into one word, while stemming attempts to harmonize different usages of one word, such that, e.g. "running", "ran", and "run" all become "run". However, unlike the example, the outcome does not necessarily represent an English word. Finally, since for our project we intend to capture general rather than specific content, we remove country names and names of people, identified by a library of names and the usage of titles, such as "Mr" or "Mrs".[11] This leaves us with a total of 5,074,428 unique tokens, which are not only single words (243,946), but also tokens of sequences of two words (3,232,064) and three words (1,598,418), referred to as bigrams and trigrams, respectively. Then as a final step, we remove overly frequent and rare tokens. Dropping rare tokens, in particular, means that we drop a lot of tokens from the list without losing a lot of text. Even after this procedure we are left with over 1 million tokens. This high dimensionality makes it impossible to use the token vectors in standard regressions. Here is where the literature has typically reduced dimensionality by focusing on particular words, i.e. keywords.

4.2 LDA Topic Models

In order to reduce the high dimensionality of our data set, we use the latent Dirichlet allocation (LDA) to model topics, a method introduced by Blei, Ng, and Jordan (2003). The LDA model in text analysis assumes that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. Topics are nothing else but probability distributions over words. They are generated by clustering tokens based on their co-occurrence in articles.

The exercise consists in splitting each article into topics k. While the number of topics K is pre-specified, the content of the topics is not. One can imagine an article consisting of a topic to be more likely to produce a list of words related to that topic. For instance, on the one hand an article about sports might be

[9] See `http://norm.al/2009/04/14/list-of-english-stop-words/` for the list of stop words.

[10] The Python package for lemmatizing is available at `http://www.nltk.org/_modules/nltk/stem/wordnet.html` and for stemming at `http://snowball.tartarus.org`.

[11] We use the Natural Language Toolkit dictionary of names for males "names.words('male.txt')" and females "names.words('female.txt')".

more likely to contain words such "football", "win", "fans", and "game". On the other hand, an article about a conflict might be more likely to use words such as "violence", "casualties", and "soldier". Through Bayesian learning, the algorithm optimizes the weighted word lists, i.e. the topics, in order to discriminate between articles. For instance, the word "win" might be more of a sports-topic word and will therefore indicate that an article is on sports. The topics are identified on the likelihood of the co-occurrence of tokens. Ultimately, the mixed-membership model represents each document as a set of shares of topics. Coming back to our example, one could imagine that an article is classified as 70% sports and 30% conflict if a particularly violent soccer match took place.

In more technical but simplified terms, the following description is based on Heinrich (2009). LDA generates a stream of observable words $w_{m,n}$, partitioned into documents which are vectors of words $\vec{w}_m$, i.e. the order of words does not matter. For each of these documents, a vector of topic proportions, $\vec{\eta}_m$, is drawn from a Dirichlet distribution $Dir(\vec{\alpha})$. From this, topic-specific words are emitted. That is, for each word, a topic indicator $z_{m,n}$ is sampled according to the document-specific mixture proportion, and then the corresponding topic-specific term distribution, $\vec{\varphi}_{z_{m,n}}$, is used to draw a word. The topics $\vec{\varphi}_k$ are sampled from a Dirichlet distribution $Dir(\vec{\beta})$ once for the entire corpus. The probability that a document m is observed can be written as

$$\begin{aligned} & p\left(\vec{w}_m, \vec{z}_m, \vec{\eta}_m, \vec{\varphi}_1, ..., \vec{\varphi}_K\right) \\ = & \prod_{n=1}^{N_m} p\left(w_{m,n} \mid \vec{\varphi}_{z_{m,n}}\right) p\left(z_{m,n} \mid \vec{\eta}_m\right) \cdot p\left(\vec{\eta}_m \mid \vec{\alpha}\right) \cdot p\left(\vec{\varphi}_1, ..., \vec{\varphi}_K \mid \vec{\beta}\right) \end{aligned} \tag{7}$$

where $p\left(\vec{\eta}_m \mid \vec{\alpha}\right)$ and $p\left(\vec{\varphi}_1, ..., \vec{\varphi}_K \mid \vec{\beta}\right)$ are drawn only once per document. The key in estimating this model is that only the $\vec{w}_m$ are actually observed. Everything else needs to be backed out. Equation (7) allows us to obtain the likelihood of observing a document $\vec{w}_m$, by integrating out the distributions $\vec{\eta}_m$ and $\vec{\varphi}_1, ..., \vec{\varphi}_K$ and summing over all topics K.

Typically, the elements of the vectors $\vec{\alpha}$ and $\vec{\beta}$ are assumed to be the same for all documents and topics, respectively. The LDA model can therefore be described by three parameters α, β and the number of topics K. High α-values mean that each article is likely to consist of a mix of many topics, whereas for low values one could imagine the extreme case of each article only consisting of one topic. Analogously, a high value of β favours a topic to contain a mixture of most words, whereas low values allow topics to consist of a limited number of prominent words. Given that we have no prior preference, we experiment using varying values for both α and β. Similarly, we experiment with the number of topics.

For statistical inference we use a Gibbs sampling technique, which is a Markov chain Monte Carlo method, to implement the LDA and let it run for 1000 iterations.[12] For a detailed and user-friendly description of the usage of LDA for topic modelling, we refer to Hansen, McMahon and Prat (2014) and

[12]The C++ Gibbs Sampler we use is provided by Phan and Nguyen (2007) is available at `http://gibbslda.sourceforge.net`. We use the default values for burn-in and thinning.

Heinrich (2009). Our preferred specifications, which we will be using for all of the baseline results presented in Section 5, is composed of 15 topics and hyperparameters $\alpha = 3.1$ and $\beta = 0.001$. The relatively low value for β forces the model to attribute fewer words to several topics. However, the impact of changing these parameters is fairly modest. The only exception is the number of topics which we discuss in more detail below.

4.3 Topic Estimation Results

As a first check on the usefulness of the LDA model we estimate topics on all text from the entire period 1975 to 2010. Despite the user not providing any prior information on content, the LDA breaks down documents into topics which seem natural and intuitive. When looking at the top 20 word list of topics, it is easy to come up with a title for each topic. Generally, in most specifications topics appear, which we would classify as sports, travel, economy, and culture. Amongst the 15 topics estimated this way one can identify lists which are clearly related to conflict situations. In Figures 4 and 5, we present these three topics as word clouds of the top 20 words of the topic. In these clouds, the size of each word is proportional to its likelihood in the corresponding topic. Notice that "armi" is the stemmed/lemmatized version of army and appears in all three topics. To give a further example, in Figure A.3 we exhibit a word cloud related to finance (a) and travel (b).

Figure 4: Word Cloud of Conflict Topic 1

Note: The cloud contains the 20 most important tokens, where the size is proportional to the importance.

It is important to keep in mind that the tokens shown in Figures 4 and 5 are only the tip of the iceberg. Topics are a probability distribution over more than 1 million tokens, i.e. single words, two-word and three-word combinations. This is important as the full list of tokens associated with the topics in Figures

4 and 5 could include factors that trigger or at least anticipate conflict. An example is the formation of a guerrilla group whose activity is reported. This group will have characteristics which are mentioned in the newspaper article and which makes it similar to other guerrilla groups. At the same time, words describing the group will appear together with words indicating violence once violence has started or is being anticipated. By aggregating tokens into topics, we therefore connect the present to the future and the experience of one country to the experience of other countries.

Figure 5: Word Clouds of Conflict Topics

(a) Conflict Topic 2

(b) Conflict Topic 3

After estimating the topic model, we are in possession of a dataset containing the composition of each article in terms of the k topics. The question remains of how to aggregate these shares of each article to receive a topic distribution in a country-year. We use a simple method that takes into account the prior probability distribution of topics in the assumed Dirichlet distribution. Call the total number of words attributed to topic k in a country/year $words_k$.[13] The share of topic k in that country/year is then

$$\theta_k = (words_k + \alpha) / \left(\sum_{k=1}^{K} words_k + K * \alpha \right) \tag{8}$$

where K is the total number of topics. Note that α enters here as the strength of the prior. If only few words are written on a country then the deviation from this prior will be relatively weak. If a lot of words are written, this indicates a deviation from the prior so as a consequence the posterior topic distribution will deviate strongly from the prior.

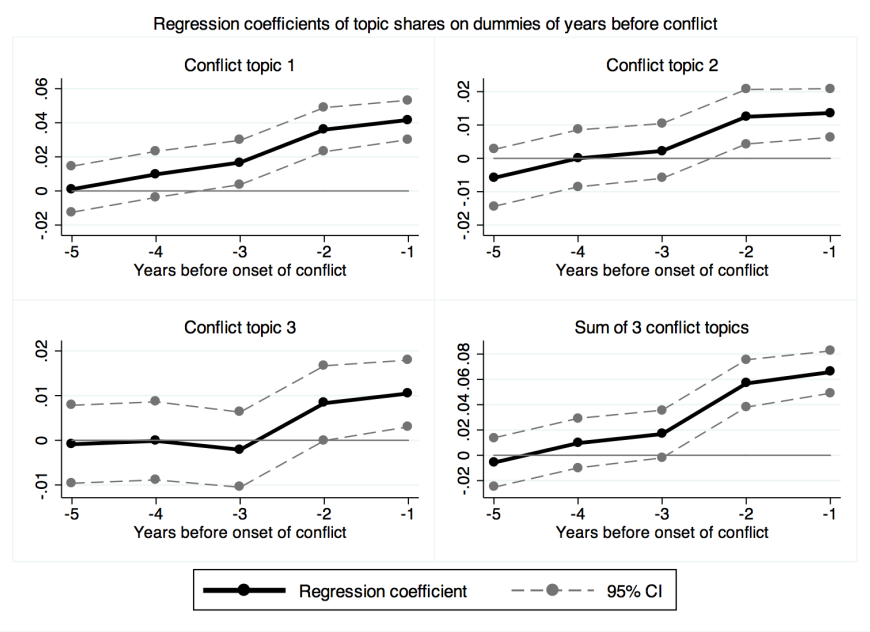
Intriguingly, the shares of topics presented in the word clouds increase in the years leading up to the onset of conflict in a statistically significant manner.[14] We test this idea using a set of country fixed effects panel regressions. In these

[13] We generate these by multiplying the kth element of the topic share vector $\vec{\eta}_m$ with the article length N_m and summing over all articles in a country/year.

[14] In Figure A.2, we show how in four exemplary countries, namely Angola, Azerbaijan, Sierra Leone, and Uganda, the sum of the three conflict topics and the finance topic vary across time. In some instances, such as in the beginning of the 1990's in Azerbaijan and

regressions we use each topic share, θ_k, as well as the sum of the topic shares as dependent variable, and regress it on dummies for the number of years before the outbreak of a conflict with at least 25 battle-related deaths. To illustrate the power of predicting the onset of conflict we set conflict years to missing.[15] As can be seen in Figure 6, two out of three conflict topics and the sum of all three significantly increase their share of the text in a country two years before the onset of conflict.

Figure 6: Conflict Topics and Years Before Conflict



We believe it is an important strength of our approach that we can use aggregates of the whole text to predict topics by simply using all topic shares θ_k when predicting. The topic model allows us to identify topics related to the outbreak of conflict, which one would not originally have in mind. This is especially true for topics that are negatively related with conflict.[16]

Up until now we have used text from all years to estimate topics. We have done so for illustrative purposes only. In what follows, we will only use text available within the training sample. In other words, to predict conflict in years

Uganda, we see spikes in the conflict topic shares preceding conflict (marked by the gray-shaded area). We can also see how in Angola and Sierra Leone, the news covered by conflict topics declines after conflicts have come to an end. Furthermore, we can tell that for these four countries the finance topic does not seem to be covered extensively during most years.

[15] Including conflict years does not alter the findings qualitatively. We later focus on incidence in most cases as the incidence model also predicts onset fairly well.

[16] For instance, in some specifications topics related to traveling seem to be negatively related to the years preceding the outbreak of conflict.

$t+1$ or $t+2$ we will use topics coming from text available until year t. The details of the procedure we use are discussed in the following section.

5 Predicting Conflict with Newspaper Topics

In this section, we use the estimated topic shares in linear fixed effects regression to forecast conflict out-of-sample. We first describe the forecasting method and then present results.

5.1 Empirical Implementation of Forecast

We now turn to the prediction of conflict with newspaper text for which we proceed in the following five steps:

1) We estimate a topic model using text written between year 1975 and year T. Hereby we obtain a vector of 15 topic shares $\theta_{it,T}$ in country i at time t which we calculate as in equation (8).

2) We use the sample from year 1975 until year T as our estimation sample. We estimate two cross-country regressions of the form

$$y_{it} = \beta_i + \theta_{it,T}\beta^{topics} + \varepsilon_{it} \quad (9)$$

$$y_{it} = \beta_i + x_{it}\beta^{controls} + \varepsilon_{it} \quad (10)$$

where y_{it} is a dummy that takes a value of one if conflict occurs one (or two) year(s) later and x_{it} are the variables in our benchmark model. The variables in the benchmark model are four political regime dummies from polity IV, infant mortality, the share of the population that is discriminated against and a dummy that captures whether more than three neighbouring countries had an armed conflict.[17]

3) We then use the respective estimates for $\hat{\beta}_i$, $\hat{\beta}^{topics}$ and $\hat{\beta}^{controls}$ to produce the fitted values

$$\begin{aligned} \hat{y}_{it}^{full,news} &= \hat{\beta}_i + \theta_{it}\hat{\beta}^{topics} \\ \hat{y}_{it}^{full,baseline} &= \hat{\beta}_i + x_{it}\hat{\beta}^{controls} \end{aligned}$$

and

$$\begin{aligned} \hat{y}_{it}^{within,news} &= \theta_{it}\hat{\beta}^{topics} \\ \hat{y}_{it}^{within,baseline} &= x_{it}\hat{\beta}^{controls} \end{aligned}$$

for the same set of countries but year $T+1$. The within fitted values, $\hat{y}_{it}^{within}$, capture the risk of conflict compared to the country's average propensity, i.e. without taking into account whether the country was low or high risk in the past. If we use the estimated $\hat{y}_{it}^{full}$, we use both the within and between variation

[17] Details are discussed in the Appendix. As far as we can tell, changes in this benchmark model do not change results qualitatively.

contained in the model. As we have shown earlier this can be dominated by the between variation.

4) Using a set of varying cutoffs c and our estimates for $\hat{y}_{it}^{within}$ and $\hat{y}_{it}^{full}$, we first calculate the number of true positives (TP_c) (and false positives (FP_c)); the number of years predicted as preceding conflict which were (weren't) preceding conflict. We also calculate the number of true negatives (TN_c) (and false negatives (FN_c)); the number of years predicted as preceding peace that were indeed (or weren't) preceding peace. Finally, we calculate the true positive rate (TPR)

$$TPR_c = \frac{TP_c}{FN_c + TP_c}$$

and the false positive rate (FPR)

$$FPR = \frac{FP_c}{FP_c + TN_c}$$

for each of the cutoffs c. The TPR is the share of all actual conflicts which is correctly identified. A TPR of 0.4 implies that 40 percent of all conflicts in the period 1996 to 2010 are correctly anticipated. The FPR captures the share of stable years which is falsely thought of as preceding conflict. A FPR of 0.4 implies that 40 percent of all years without a conflict in the following year are falsely thought of as preceding conflict.

The fact that we use cutoffs to produce 0 and 1 predictions minimizes the problems brought about by our linear probability model.[18] The transformation of fitted values when predicting 0s and 1s, which are then contrasted with the true values, allows all possible values to be easily converted into TPR and FPR values. In this way, the two dimensions of TPR and FPR provide a common space in which to interpret the fitted values $\hat{y}_{it}^{within}$ and $\hat{y}_{it}^{full}$, despite the fact that $\hat{y}_{it}^{within}$ is centered around zero.

5) This procedure is repeated for all years between $T = 1995$ to $T = 2010$ so that the last predicted instability out-of-sample is in 2011 (2012) when predicting one (two) years ahead. We then use the overall TPR and FPR at different cutoffs to produce receiver operating characteristic (ROC) curves. These depict the TPR and the FPR at various cutoffs and therefore capture the power in a simple and nonetheless meaningful way.

5.2 Main Results

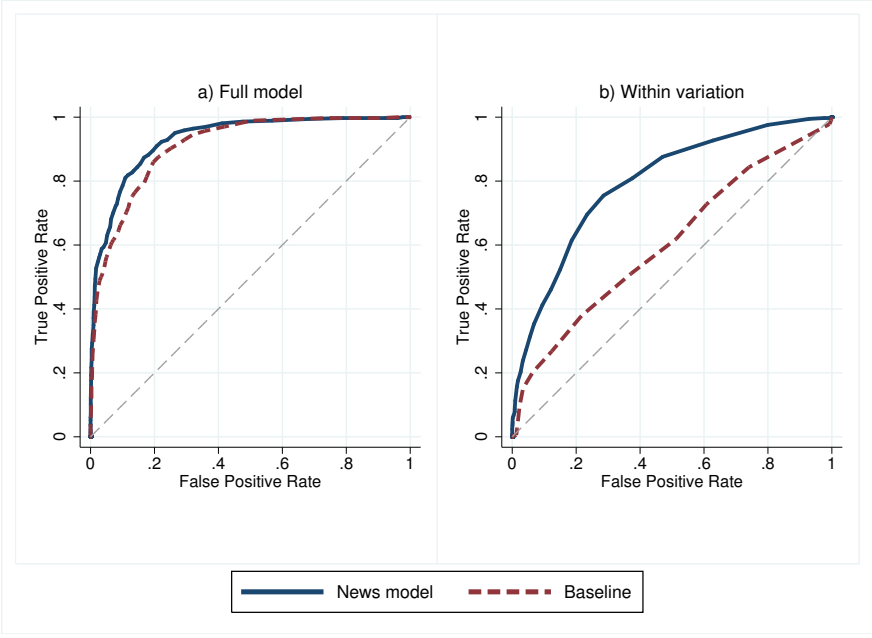
In this section, we present results. In principle, both the prediction of onset and incidence should be of interest. An additional choice is which level of conflict to focus on. In the main results sections we focus on incidence and two definitions of internal conflict: more than 25 battle-related deaths and more than 1000 battle-related deaths. We also show robustness checks using conflict onset as well as other definitions of conflict.

[18] We use the linear model to maintain comparability across models. In a logit model it would be impossible to separate within and across variation without excluding a large number of countries from the dataset.

Our main results is shown in the two graphs in Figure 7 which show receiver operating characteristic (ROC) curves. The blue lines in both graphs show the forecasting performance using the news model solely relying on topic shares. The red lines provide the ROC curve of the benchmark model without the topic shares. Panel a) in Figure 7 shows the outcome when we use the full model $\hat{y}_{it}^{full,news}$ (blue line) and $\hat{y}_{it}^{full,controls}$ (red dashed line). In panel b) we show the within variation given by $\hat{y}_{it}^{within,news}$ (blue line) and $\hat{y}_{it}^{within,controls}$ (red dashed line).

The ROC curves are a good way to summarize forecast power as they present the trade off between true positives and false positives. Optimally, one would want a true positive rate of 1 at a false positive rate of 0. This would mean that all conflicts are predicted without raising any false alarms. The 45 degree line in ROC curves is the benchmark that would be reached by random forecasts - the dart throwing chimp.

Figure 7: ROC Curves of Conflict Defined as Armed Conflict Incidence



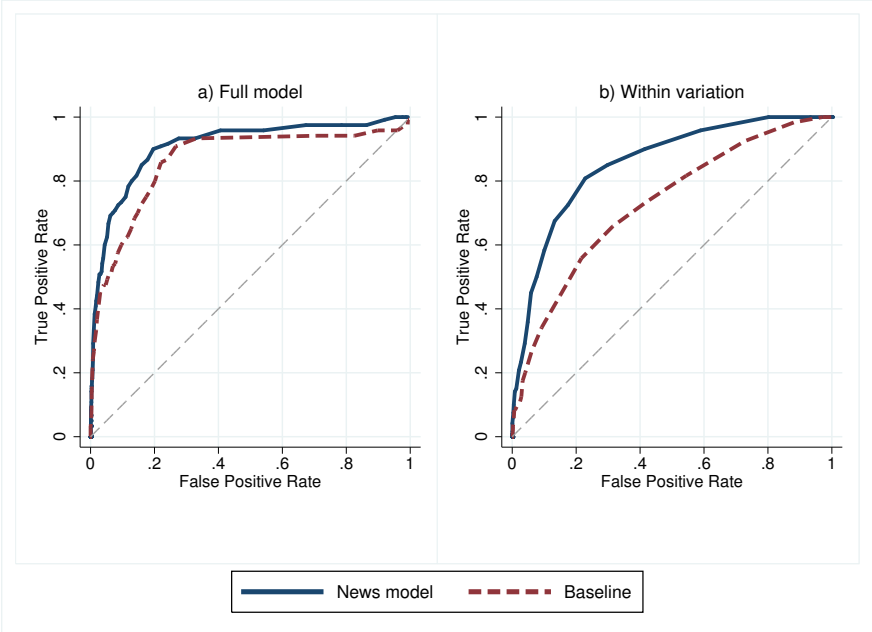
Quite clearly, both models perform extremely well predicting conflict incidence in the full model displayed in a). At a false positive rate (FPR) of 20 percent both reach a true positive rate (TPR) of about 90 percent. At a false positive rate of 50 percent the true positive rate is close to 1. The news model statistically dominates the benchmark model in panel a). While the difference looks small it amounts to over a ten percentage point difference in the FPR at small TPRs.

The picture changes dramatically when we turn towards the within variation

in panel b) in Figure 7. The benchmark model now performs less well. At a FPR of 20 percent the TPR of the benchmark model is now below 40 percent while it is over 60 percent in the news model. This change in the performance reflects the fact that a larger share of the variation contained in the fitted value $\hat{y}_{it}^{full,controls}$ comes from the fixed effect.

Figure 8 and Figure 9 repeat the same exercise for our alternative conflict measure. In Figure 8 we show the attempt of predicting civil wars, i.e. years with more than 1000 battle-related deaths in a year. In the news model, the predictive power of the within model is almost as strong as the predictive power of the full model now. This is quite a striking finding given the difficulty of forecasting the timing of such rare events. The benchmark model now also performs better within with a TPR of about 50 percent at a FPR of 20 percent. The likely reason for this improvement in performance is the higher threshold of violence which means events take place in circumstances that are easier to distinguish from normal times.

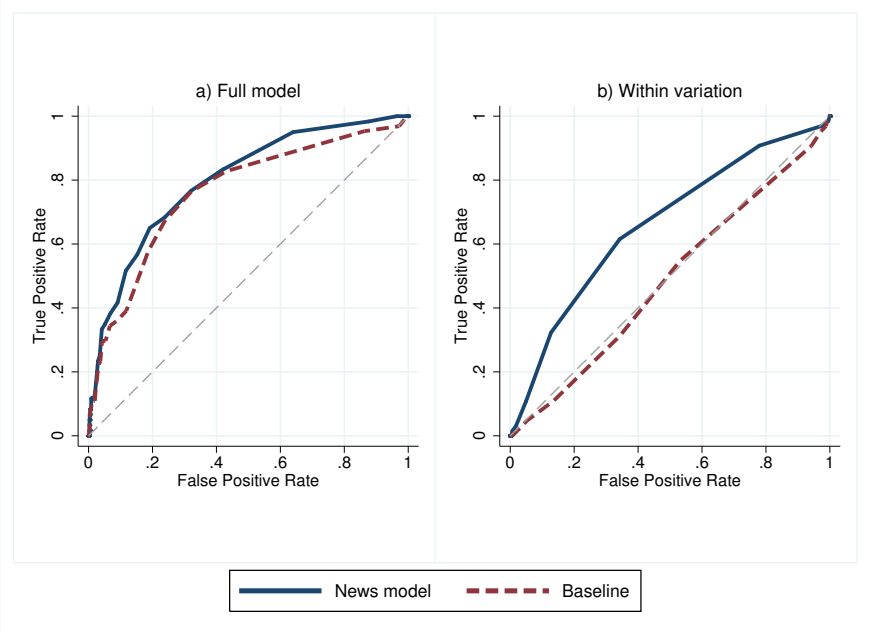
Figure 8: ROC Curves of Conflict Defined as War



A possible concern when using conflict incidence is that all forecasting power comes from situations in which a conflict year follows the next. While this concern is reduced by the use of a fixed effect model we should confirm that news are indeed of use in predicting onset. We therefore change our conflict definition by excluding all years of conflict. This is particularly useful when looking at armed conflict as we now identify the model exclusively from situations without

violence.[19] In Figure 9 we display the result. This reveals, quite clearly, that predicting the much rarer onset is relatively hard. Model performance generally drops. However, the basic message is maintained. The baseline model now becomes literally useless in predicting the timing of conflict. It can be shown that the baseline model performs less poorly when predicting civil war but, again, the basic pattern is always maintained.

Figure 9: ROC Curves of Conflict Defined as Armed Conflict Onset



We now return to the relatively encouraging results in Figure 7.[20] The figure displays a relatively high true positive rate for relatively low values of the false positive rate. However, it should be kept in mind that in rare events like armed conflict, false positives are a problem even at low rates as the majority of country/years are non-conflict years. A way to think about the problem of the number of false positives is what is called *precision* in the forecasting literature. Precision can be calculated by the formula

$$P_c = \frac{TP_c}{FP_c + TP_c}$$

which gives the number of years in which conflict was predicted correctly divided by all conflict predictions. Hence, the difference compared to the TPR_c is that

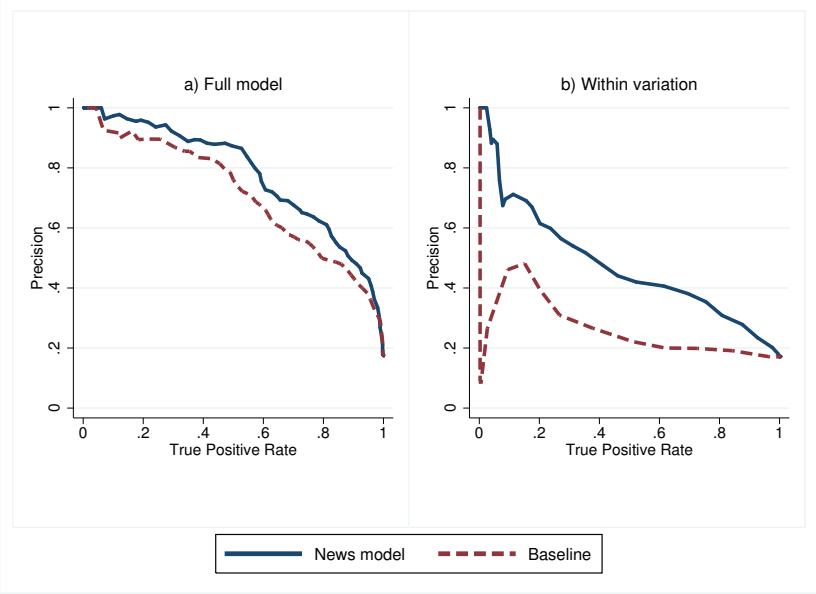
[19] We also checked that our results here are not driven by the conflict definition we use. This is a concern as some conflicts changed from external conflicts to internal conflicts. Our results also hold when using a broader definition of conflict.

[20] We also find that our incidence models (displayed in Figures 7 and 9) are actually quite good in predicting onset.

true positives are not set into a relation to all years of conflict ($FN_c + TP_c$) but to all years in which conflict was predicted ($FP_c + TP_c$).

We then compare this statistics to the TPR_c in a precision-recall curve displayed in Figure 10. Note that this curve can take values of P_c of 0 and 1 for $TPR_c = 0$, depending on whether the first positives are true or false positives. As the TPR_c goes towards 1 precision will converge towards the ratio between conflict and non-conflict observations. For armed conflict this is below 20 percent in our sample.

Figure 10: Precision Recall Curves for Armed Conflict Incidence



Precision in the full model is fairly high both in the baseline model and in the news model. At a true positive rate of 90 percent, precision in the news model is still close to 50 percent. This means that the large majority of armed conflicts are correctly anticipated and at the cost of raising false alarms in only half the cases. The within variation contained in the model is, again, dominating the benchmark.

In Figures 11 and 12 we show the precision recall curves for our other two conflict definitions. Precision is generally smaller and now the news model does not complete dominate the benchmark model in terms of precision. For example, for the very high cutoffs (TPR close to 0) the full benchmark model predicts some civil wars while the news model does not. Still, generally the perception that the time variation is particularly useful in the news model is maintained. In the per capita model of civil war the news model generates a precision of over 20 percent at a true positive rate of 70 percent.

Figure 11: Precision Recall Curves for Armed Conflict Onset

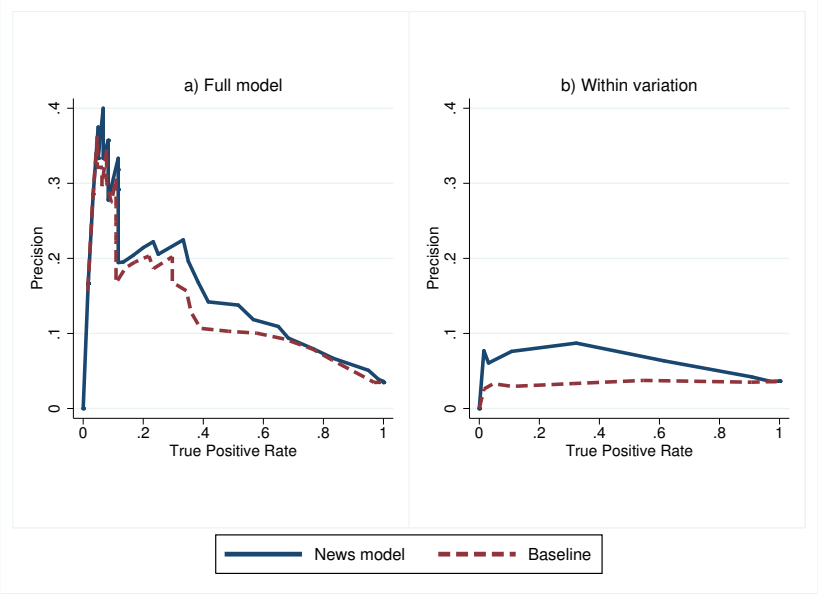
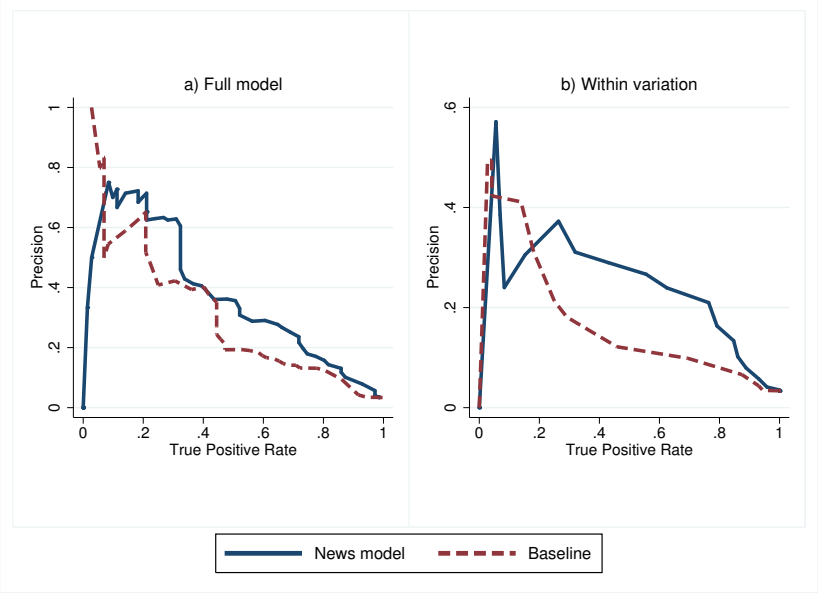


Figure 12: Precision Recall Curves for Conflict Incidence Defined in Per Capita Terms



5.3 Policy View

Imagine the news model would be used by a policymaker who thinks that she is able to prevent or stop violence in a country. As pointed out by Kennedy (2015) the relevant information for a decision problem of how to set up a cut-off, c, in such a scenario combines the results in the previous section with the policymaker's attribution of weights to the four outcomes TP_c, FP_c, TN_c and FN_c. Call the costs for these c_{TP}, c_{FP}, c_{TN} and c_{FN}. The policymaker will then minimize total costs

$$costs_c = TP_c \times c_{TP} + FP_c \times c_{FP} + TN_c \times c_{TN} + FN_c \times c_{FN}. \qquad (11)$$

Potentially this problem could be fed with country-specific cost parameters. One could, for example, give countries with larger populations more weight.

For now, we only aim to illustrate the choice of a cutoff implied by this model. To simplify the discussion, assume that the costs of correct predictions are $c_{TP} = c_{TN} = 0$. Typically one would attribute a high cost to c_{FN} because a negative surprise might be particularly costly for the policymaker, $c_{FN} > c_{FP}$. We then look at two scenarios regarding the remaining parameters. First, we assume that false positives are relatively cheap, i.e. that they cost $c_{FP} = \frac{1}{10}c_{FN}$. Second, we assume that the costs of false positives are half the costs of false negatives $c_{FP} = \frac{1}{2}c_{FN}$. To simplify the discussion further we set $c_{FN} = 1$.

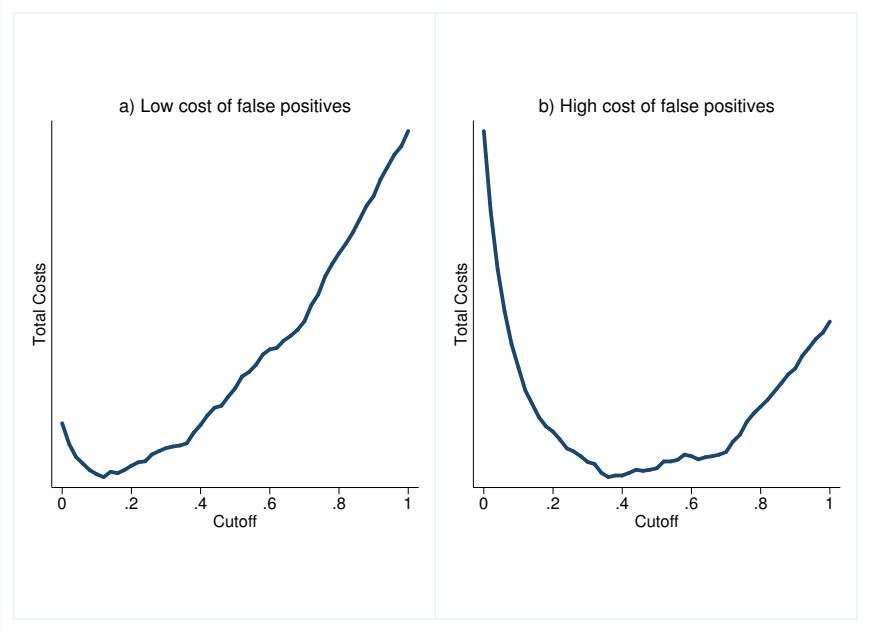
We use these parameter values to evaluate the total costs generated by different cutoffs, c. For every country/year we first generate a dummy that takes a value of 1 if the condition $\hat{y}_{it}^{within,news} \geq c$ is satisfied. We then compare that dummy to the variable y_{it} to generate the set of variables TP_c, FP_c, TN_c and FN_c. Finally, we calculate $costs_c$ from equation (11).

In panel a) of Figure 13 we show the results under the assumption that false positives carry a low costs. The cost function takes a U-shape with a minimum at around $c = 0.15$. At this value the model generates a large TPR of around 90 percent. For lower values than $c = 0.15$ the number of false positives would increase too much compared to the gain in true positives. For higher values of the cutoff the generation of more false negatives leads to an increase in costs. The picture changes significantly in panel b) where we assume a higher cost of false positives. Now the minimum cost is reached just below $c = 0.4$. The higher cutoff reflects the fact that false positives have become more expensive. The basic trade-off remains in place.

Thanks to the power of the fixed effects, this cost model does not perform much worse in overall costs. It also has a similar cost-minimizing cutoff under the different assumptions on costs. This allows us to run a thought experiment. Assuming a cutoff and costs, how much would each country have contributed to total costs? The answer, for the scenario with $c_{FP} = \frac{1}{2}$ and $c = 0.4$, is in Figure 14. The two curves show the kernel density of average costs for the news and the benchmark model for the 53 countries with positive costs.[21] Two features of the graph are particularly remarkable. On the one hand, the news model generates a

[21] These are exactly the same in the two models.

Figure 13: Cost Curves for Armed Conflict Incidence

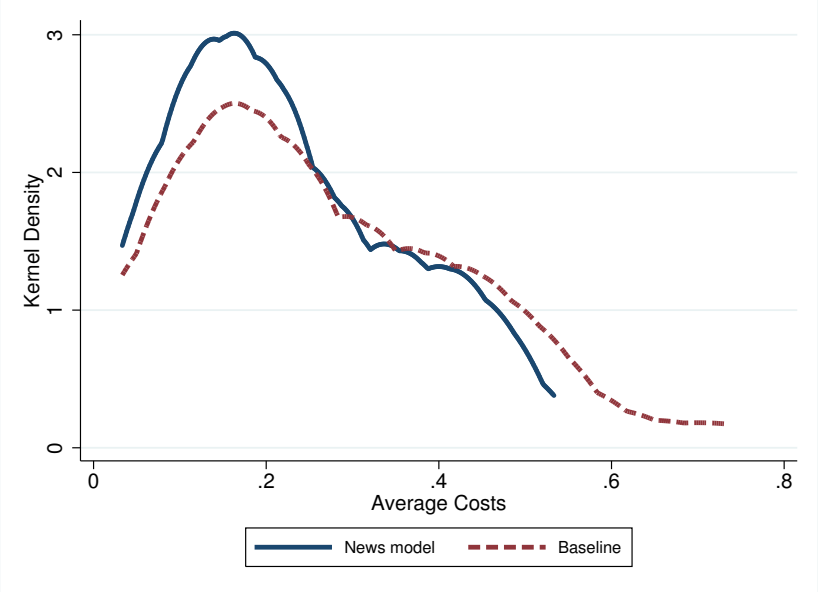


higher density for several countries at low costs. On the other hand, the baseline model generates a higher density at extremely high average costs. The reason is that the baseline model generates a lot of false positives for some countries and a lot of false negatives for others. Burundi or the Democratic Republic of the Congo, for example, enter conflict in the 1990s but were without violence throughout most of the training sample. The baseline model does not pick this up quickly enough and accumulates a lot of false negatives for these countries. The news model reacts to new information and therefore generates more true positives instead of false negatives. The news model raises more alarms in a less concentrated way which leads to higher densities at low costs. This is one of the advantage of having more useful time variation in the news model.

5.4 Robustness

In this section we discuss the robustness of our main findings. The, perhaps, most important question is whether the parameters of the topic model we have chosen has repercussions for its performance in forecasting. The little we have experimented with different parameter settings indicates that the parameters α and β do not affect outcomes systematically. The number of topics, however, does affect performance considerably. Higher numbers of topics increase the fit of the model within sample but fail to provide reliable out-of-sample predictions. The reason is that if we estimate, say, 150 topics these will be specific to one

Figure 14: Cost Curves for Armed Conflict Incidence (Baseline vs News Model)



conflict. Specific topics are then not useful in predicting conflict in a different country. By assuming few topics we force the algorithm to move the description of more situations into the same topic. We believe that this is what allows us to do forecasts. As an illustration, Figures A.4 and A.5 show our main results regarding armed conflict onset and incidence for a model with five topics. Forecasting power remains strikingly similar to the 15 topic model and sometimes even improves. This is also true if we try to forecast civil wars.

We also used our model to forecast conflict one or two years before it happens. Results are in Figure A.6 for armed conflict and Figure A.7 for civil war. When predicting the timing of armed conflict the news model again clearly dominates. The same is true for civil war except that now baseline model performs slightly better than the news model for high true positive rates (low c).

Mueller (2014) suggests a measure of per capita violence intensity motivated by the fact that some outcomes could be driven by violence risk instead of aggregate counts. News could be generated by the same logic - events that trigger conflict in a part of India might receive less attention simply because India has a larger population. Forecasting civil war defined by a per capita cutoff would then be easier. Figure A.8 shows results when we try to forecast conflict defined by more than 0.08 battle-related deaths per 1000 population. Again, the news model predicts the timing of conflict better.

Finally, we use a different benchmark model which relies only on the most readily available data: Penn World Table data, Polity IV data and the conflict data. In this model we use log GDP, log population, the polity2 score and a

level three polynomial of the number of years since the last conflict. Figures A.9 and A.10 show results in the prediction of armed conflict incidence and onset, respectively. Figure A.9 bears a surprise. The timing of conflict is now predicted almost perfectly by the benchmark model - much better than in the news model. However, Figure A.10 reveals that this is entirely due to an excellent prediction of incidence. The baseline model completely fails to predict the timing of onset. The same pattern holds for civil war. Further investigation of these facts reveals that this is due to the polynomial of the number of years since last conflict. Given an onset, the timing of more violence is predicted extremely well by a polynomial. Most conflict years follow other conflict years so that the estimate is drawn into this direction. However, a researcher or policymaker who wants to forecast onset does not benefit from this fact. This highlights the importance of studying both the forecast ability of the model regarding incidence and onset together.

6 Comparison to Other News Models

It is widely accepted that news represent a huge potential for forecasting conflict. However, the literature has typically not framed this potential in terms of between and within variation. Given the focus on monthly or quarterly forecasts this is somewhat surprising as this issue is particularly important when forecasting within small time units. Countries in conflict will generate a large number of consecutive positives.

In this section we compare our model to two important parts of the literature; models coming from the Integrated Conflict Early Warning System (ICEWS) as described in Ward et al. (2013), which use event data generated from news, and a model from Chadefaux (2014), which uses counts of keywords. We see our work as complementary to these two for several reasons. First, our method allows us to use all news content to predict civil war without imposing any priors. The use of negative correlations to elements in the news, for example tourism, is a direct result of this. However, this means we need to rely on sources that provide the entire text of articles. Second, the fact that we try to forecast the onset of rare events one and even two years before they happen implies that we need to rely on news sources that are consistently available for decades. We therefore relied on three newspapers which gives us 630,000 articles. For comparison, the ICEWS uses more than 30 million news stories whereas Chadefaux (2014) searches keywords in over 60 million pages of news text.

Despite these differences, it is interesting to see whether the restriction of news sources we impose leads to large disadvantages. On the other hand, we expect an advantage over word counts derived from the same data. If not, it would not be worthwhile using a topic model.

We first turn towards comparing our results to two models using ICEWS events. We use the two models from Ward et al. (2013) that try to predict ethnic violence and insurgency. The first model contains the proportion of population

excluded, the number of excluded ethnic groups, the number of armed conflicts in nearby countries and an interaction between the proportion and the number of groups. Additionally, the model adds events from ICEWS: the number of high intensity actions from dissidents to government. Since the event data is only available since 1995 we need to change the starting point of our analysis. We now start running out-of-sample regressions in the year 2000. Results regarding armed conflict incidence are in Figure A.11 in the Appendix. Both models have an extremely good fit in the between variation. Perhaps surprisingly, our news model completely dominates on the within dimension as well. The same is true for the prediction of civil war incidence which is displayed in Figure A.12. This does not change if we look at onset. The timing of both armed conflict onset (Figure A.13) and civil war onset (Figure A.14) are forecasted much more precisely through a topic model. The only exception are small ranges of cutoffs for armed conflict.

The second ICEWS model we compare in terms of its forecasting power is a model of insurgency, which contains the number of excluded groups, the number of included groups, the proportion of population excluded, the squared proportion of population excluded, the polity score, the squared polity score and violence in neighboring countries. In addition, the model adds the number of high intensity actions from ethnic groups to government. Our topic model now dominates for armed conflict incidence (Figure A.15) but does not do so any more for civil war (Figure A.16). For false positive rates below 0.18 the insurgency model produces a higher true positive rate. As our analysis in the previous section has shown, the insurgency model would therefore be preferred to forecast civil wars if relatively high costs are attributed to false positives, i.e. with a high optimal cutoff. A similar pattern arises if we look at onset. Again, our topic model almost completely dominates when forecasting the timing of armed conflict onset (Figure A.15) but only provides a better forecast for higher positive rates when forecasting civil war onset (Figure A.17).[22]

Finally, we move towards a comparison to word counts of words also used by Chadefaux (2014). We follow his methodology closely and count the words in our news data for every country/year. We then use the count together with the total number of counts in the same year and an interaction between the counts and the total counts. Finally, we add the polity2 score and a level three polynomial of the years since the last conflict. As we have seen in the robustness section, including the last three variables leads to an extremely good fit when looking at incidence and at the same time an extremely bad fit when looking at onset. We find the same pattern again. The news count model is completely uninformative on the time dimension when forecasting either armed conflict onset (Figure A.18) or civil war onset (Figure **??**) but dominates when forecasting incidence. Given the strength of the polynomial we needed to confirm that the news counts themselves

[22] We also tested the forecasting power of the ICEWS event data by running a model with five different event types (high intensity actions: dissidents → government, ethnic groups → government, conflictual words: any domestic group → government, opposition → government, government → opposition). This model seems to combine the better parts of both ICEWS models but still performs worse than the topic model for most parameter values.

have some forecasting power by running a model only with the three news data variables. Our topic model now completely dominates in all specifications but the count data has some forecasting power when forecasting civil war onset (Figure A.20).

Overall there are two main conclusions we draw from this. First, models that contain news data perform generally well in the prediction of the timing of conflict. We can also confirm that the event data provided by ICEWS leads to quite good forecasting power. Second, the use of a topic model seems to have some advantages for the task at hand. Our model provides better forecasts in most cases despite the fact that we rely on fewer articles.

7 The Advantage of Time-varying Country Risk

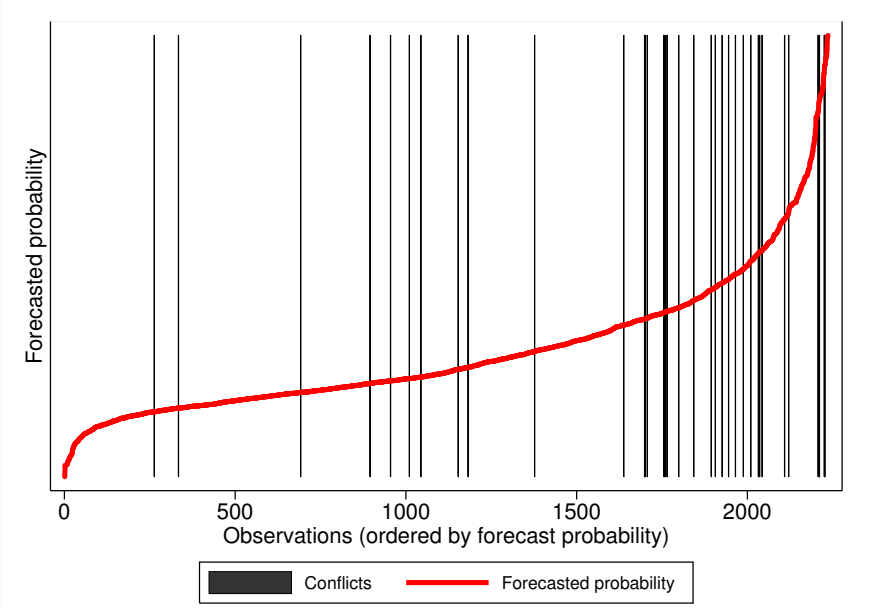
Our results show that differences across models are relatively small when using the full model. It is when looking at the within variation that differences become clear. Take Figure 15, which shows the out-of-sample prediction, $\hat{y}_{it}^{within,news}$, and contrasts it with the actual incidence of armed conflict. The model is clearly able to predict the timing of conflict incidence. Most black bars are towards the right-hand-side of the Figure, which is in stark contrast to the analog Figure 3 obtained from using standard variables. Changes of the predicted probability within a country are useful to identify the incidence of conflict. This has several appeals.

There is a clear conceptual advantage of using within variation. If a model contains variation that can be captured by only fixed effects then the forecasting essentially relies on the fact that conflict will occur where it has occurred before. From an academic perspective this is not particularly satisfactory as it does not help us learn about factors that lead to the outbreak of conflict.

But even purely from a forecasting perspective it makes sense to care about the degree to which the model performs on the time dimension. Forecasting from the between variation generates a problem in two cases; if a country that used to be violent becomes peaceful or if a country that was peaceful for a long time turns violent. The first problem generates false positives. For example, high values of $\hat{y}_{it}^{full,news}$ without imminent conflict are sequential years in countries like Cambodia, Iraq or Guatemala which did suffered a lot of internal conflict after the second world war, but remained stable in the period 1996-2010. In contrast, the highest values of $\hat{y}_{it}^{within,news}$ without a conflict following are all from different countries. The five cases with the highest probability are:

- Kuwait (2002): Faylaka Island attack against United States Marines.
- Liberia (1996): last year of civil war, heavy fighting in Monrovia.
- Lebanon (2007): heavy fighting broke out but ended within the same year.
- Eritrea (2000): external war between Eritrea and Ethiopia.

Figure 15: Within Variation Using News is Useful for Forecasting



Note: Forecast probability is the predicted value from a fixed effects regression, excluding the fixed component, in the years 1975-1995 Observations are from the years 1996-2010 and are ordered by the fitted value (red line). Black lines indicate a year that actually preceded violence by one year. Violence is defined as more than 1000 battle related deaths.

- Kyrgyzstan (2010): massive riots and violence.[23]

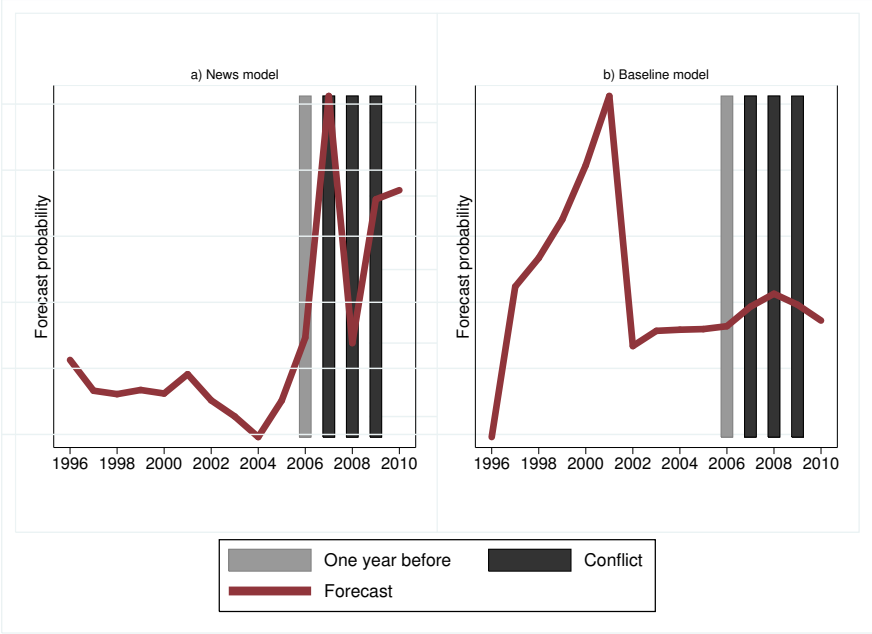
These cases illustrate that, given the information available at a given time, it is not easy to predict whether the respective country/year will be followed by more violence. The model indeed provides a reasonable measure of conflict risk which is varying in time. In this way, the probability estimates could actually be an additional source of information for the study of policies for de-escalation.

The second problem are false negatives, the surprising outbreak of conflict not captured by the fixed effect model. The difference between the full model and the within model here is not as clear-cut. The five biggest failures of the full model, where a very low value of $\hat{y}_{it}^{full,news}$ are followed by conflict incidence in the following year are Eritrea (1996), the Central African Republic (2000), Guinea Bissau (1997), Libya (2010) and Eritrea (2003). In the within model they are China (2008), Guinea Bissau (1997), Georgia (2007), Peru (2010) and Angola (2009).

[23]The case of Kyrgyzstan (2010) is particularly interesting as the model predicts a marked increase of the probability already in 2009. However, the year 2010 is not coded as armed conflict by PRIO/UCDP.

However, the contrast to the benchmark model is quite strong. We know that $\hat{y}_{it}^{within,news}$ leads to better forecasts than $\hat{y}_{it}^{within,baseline}$. An example, Mali, is shown in Figure 16. The panel on the right shows the within fitted values $\hat{y}_{it}^{within,news}$ for Mid-Intensity conflict as a red line. The black bars depict conflict years while the gray bars depict the year before conflict onset, i.e. the event we would want to predict with an onset model. In 2006, Touareg rebels became active in the north of Mali and began raids. The fitted values of our forecast model of incidence displays a sharp move upwards in that year.[24] This is particularly striking as the fitted values from the baseline model $\hat{y}_{it}^{within,baseline}$ show no such anticipation of the imminent conflict. Instead, the prediction of the baseline model remains relatively low in 2006. It is important to remember that these are out-of-sample predictions, i.e. to generate the fitted value in 2006 only the news data until the that year is used.

Figure 16: Prediction of Touareg Rebellion in Mali



[24] We show the forecast of the incidence model to illustrate that this model is able to forecast onset. The onset model unsurprisingly performs equally well in this case.

8 Conclusion

In this article we present a new method of aggregating news text in a meaningful way. Topic models have the ability to diminish the dimensionality of text from counts of more than one million expressions to, for example, fifteen topics. We have argued that, aggregated this way, news text can be used to predict the timing of conflict.

Our findings highlight that models need to be tested for whether their within variation is meaningful. If not, policymakers might rely on meaningless changes of risk across time and this has the potential to lead to large errors. We have shown, for example, that the within variation of a standard model has surprisingly little power when forecasting the timing of armed conflict. This is a finding that should be taken into account when interpreting existing studies that do not distinguish within from between variation.

Ultimately, forecasters might face a trade-off between prediction with maximum accuracy overall or using a less accurate model that generates useful variation across time. At the very least, using a model with useful variation across time should provide a useful addition for forecasters. Examples like Mali provide a strong case for using news to spot dangers that would otherwise stay off the radar. At the same time, we believe that the fixed effects framework offers a promising way forward to understand how events like the Arab Spring might have been predicted. Relying on between variation in building forecast models will not help in this task.

Having useful variation in conflict risk across time might be of value on its own right. There are many cases for which our model reports sudden increases in conflict risk which were not followed by conflict. There are two options regarding this variation. First, it might be due to reporting in the newspapers we use. For example, the invasion of Iraq generated a lot of "conflict" news which might have spilled over to other countries. Second, risk might have actually increased but policies prevented destabilization. False positives, if derived from a model with meaningful within variation, provide an opportunity to study policies that prevent or stop conflict.

Topic models could provide a useful alley for research in political events more generally. We have used the most simple, off-the-shelf, version of the various algorithms available and have used the same text collection for estimating the topic model and calculating topic shares. One could instead train a topic model on specific sub-sets of texts or on a separate set of texts and then use this to spot the generated topics in the main body of newspaper articles. Applications include training a topic model on academic articles or specialized country reports and then using the generated topics to figure out which set of topics lead to better forecasts. Technical extensions or refinements could include using more recently developed topic modelling techniques, such as dynamic topic models (Blei and Lafferty 2006) or a structural topic model (Roberts et al. 2013).

A Appendix

Figure A.1: Total number of news articles per year

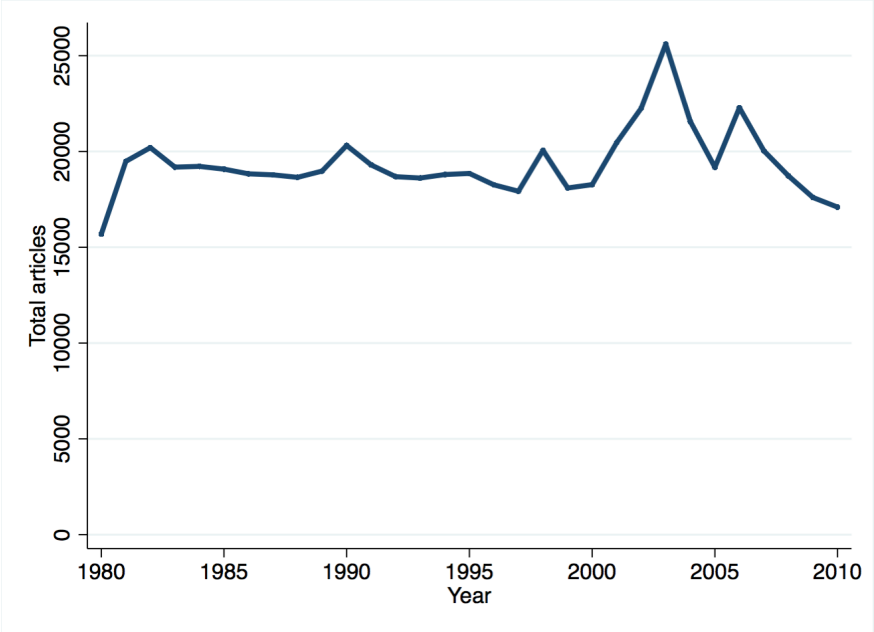


Figure A.2: Conflict Topics and Conflict Across Countries

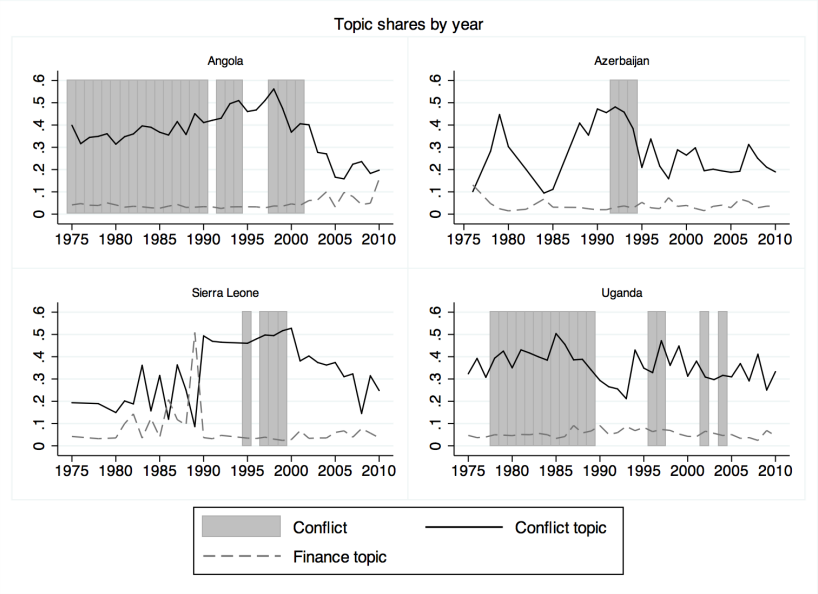
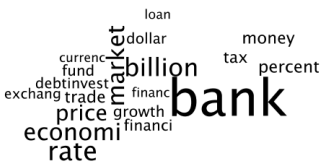


Figure A.3: Word Clouds of Topics



(a) Finance Topic

(b) Travel Topic

Figure A.4: ROC Curves of Conflict Defined as Armed Conflict Incidence (5 Topics)

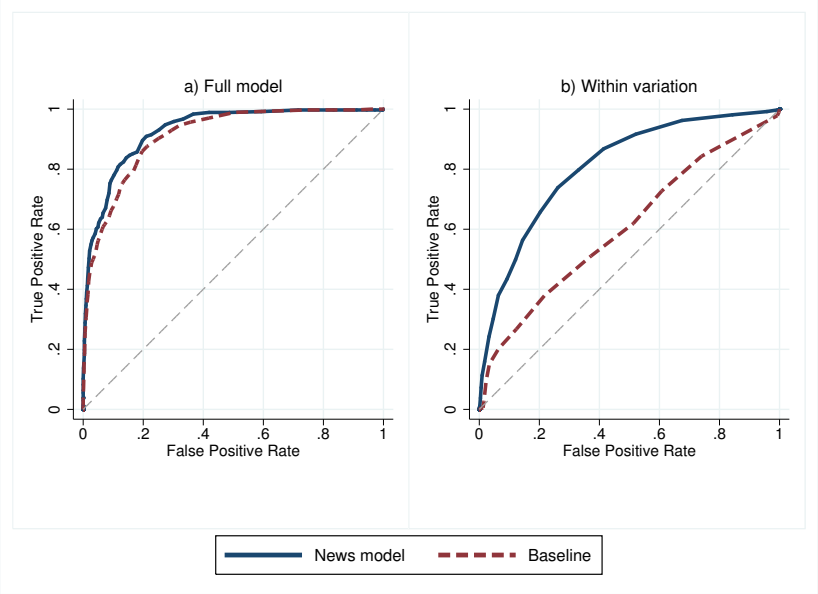


Figure A.5: ROC Curves of Conflict Defined as Armed Conflict Onset (5 Topics)

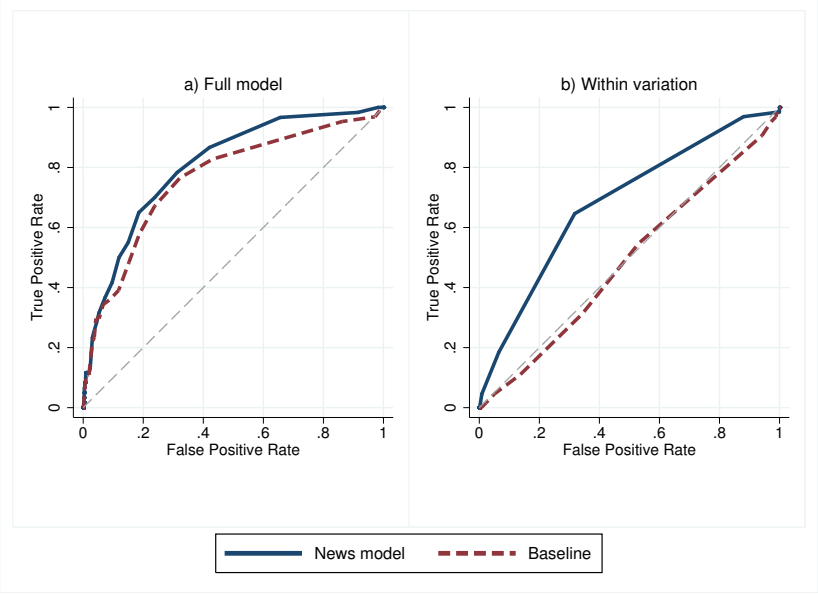


Figure A.6: ROC Curves of Conflict Defined as Armed Conflict Incidence (2 Years Before)

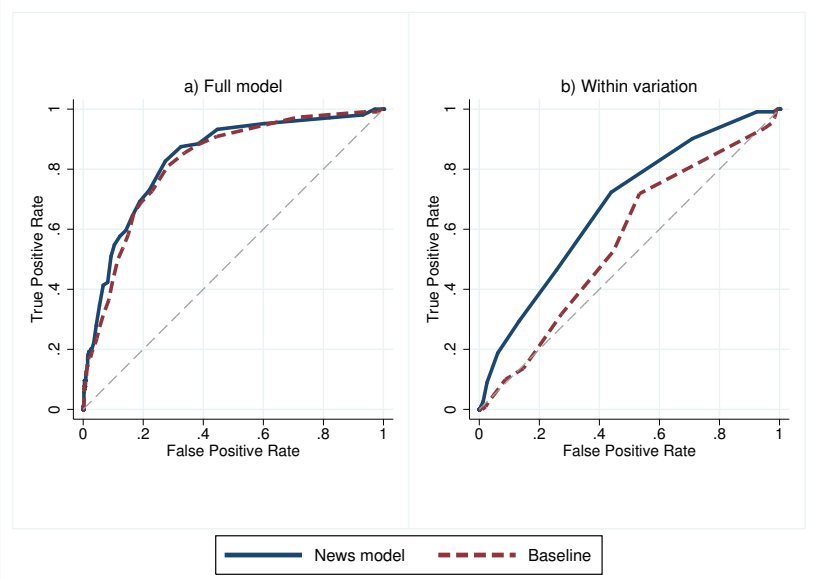


Figure A.7: ROC Curves of Conflict Defined as War Onset (2 Years Before)

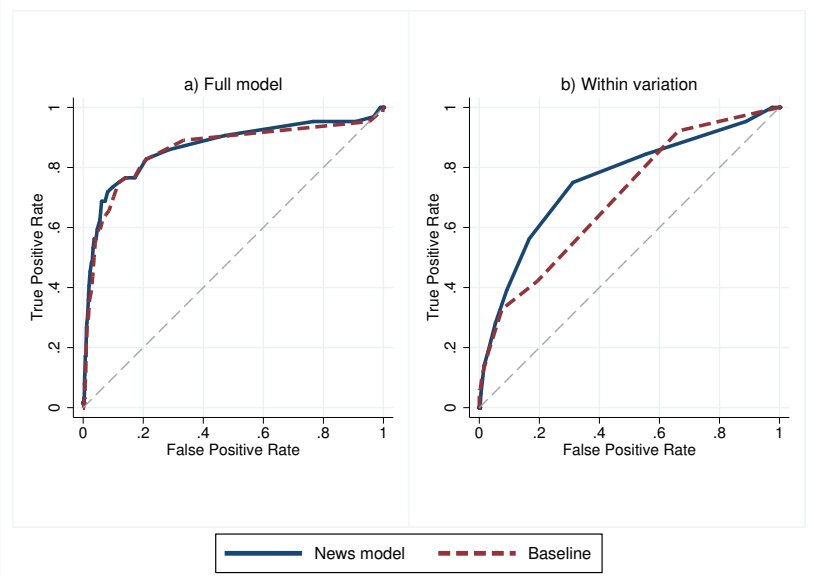


Figure A.8: ROC Curves of Conflict Incidence Defined in Per Capita Terms

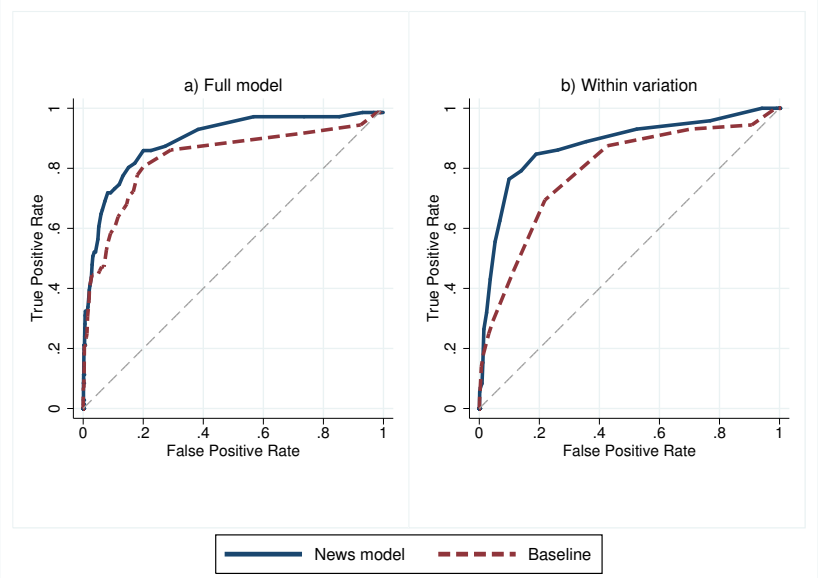


Figure A.9: ROC Curves of Conflict Defined as Armed Conflict Incidence (News vs Simple Model)

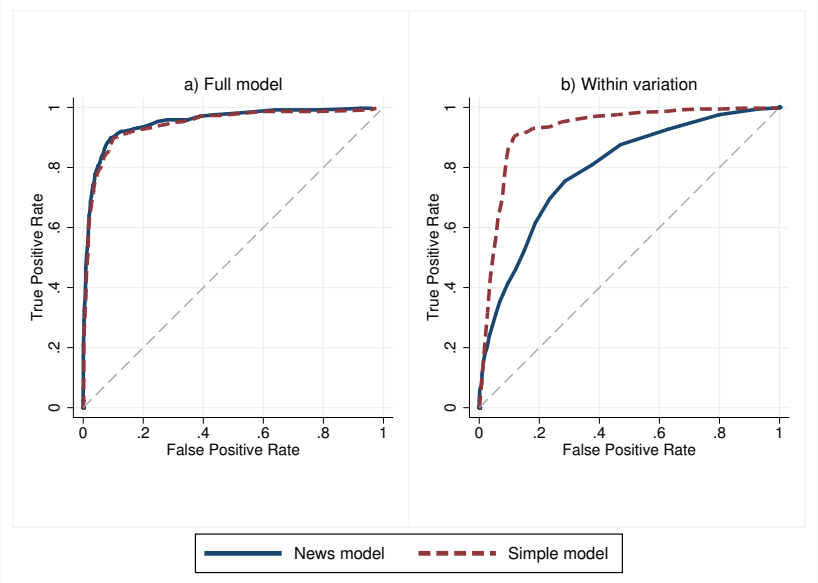


Figure A.10: ROC Curves of Conflict Defined as Armed Conflict Onset (News vs Simple Model)

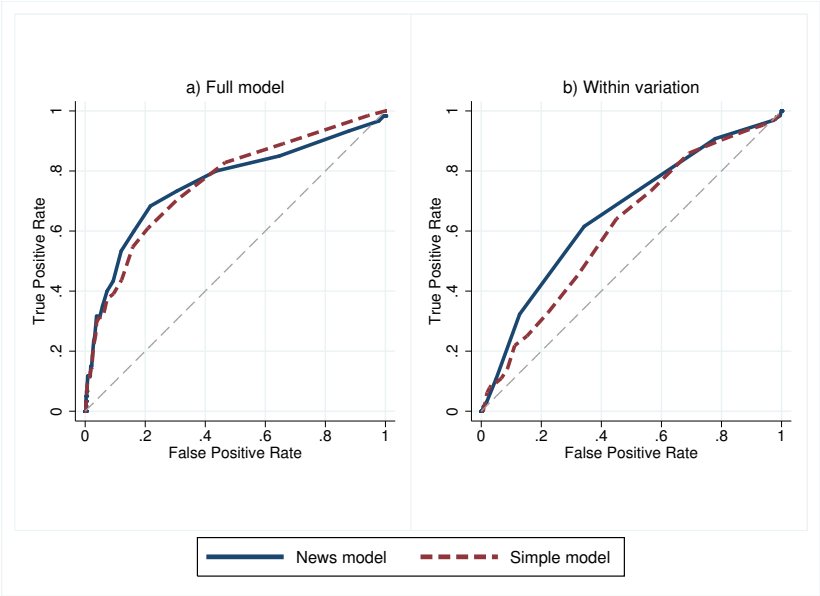


Figure A.11: ROC Curves of Conflict Defined as Armed Conflict Incidence (News vs Ethnic Events Model)

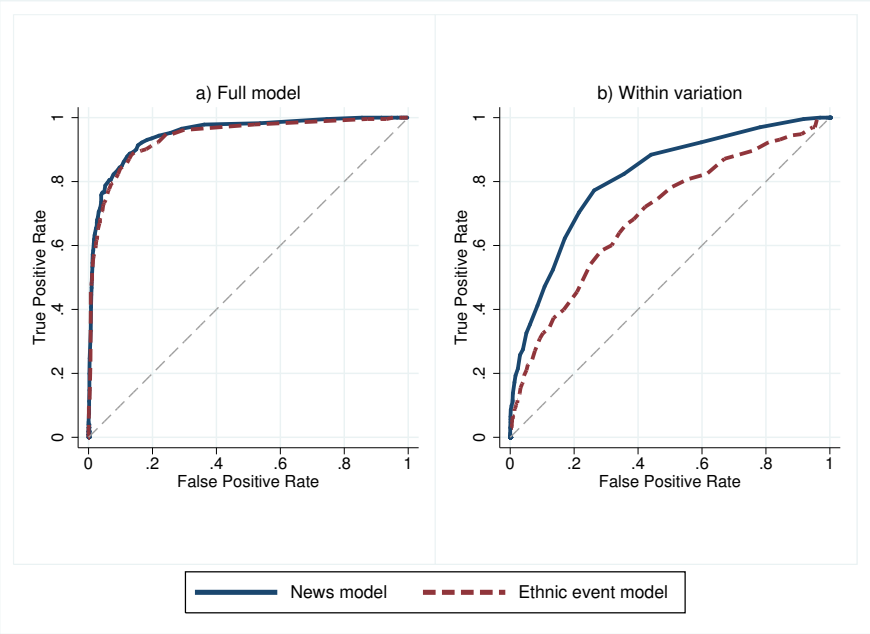


Figure A.12: ROC Curves of Conflict Defined as War Incidence (News vs Ethnic Events Model)

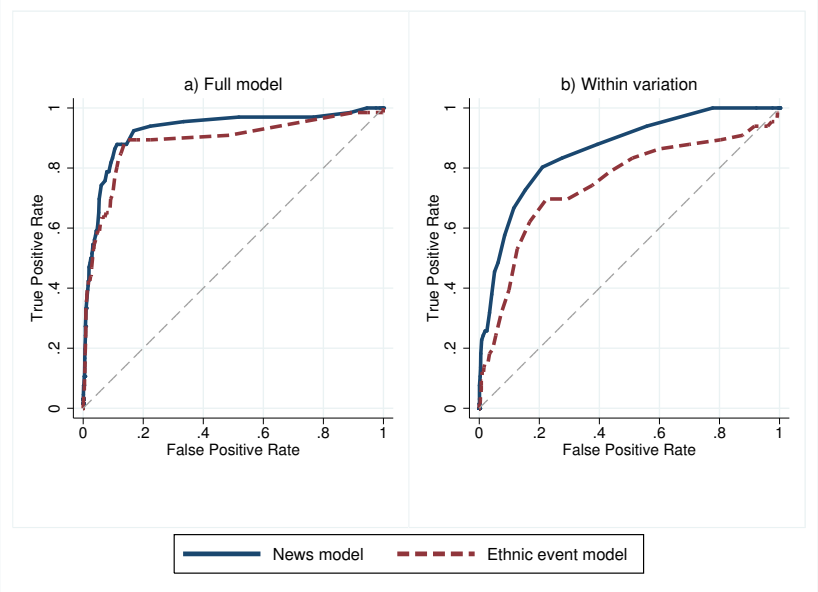


Figure A.13: ROC Curves of Conflict Defined as Armed Conflict Onset (News vs Ethnic Events Model)

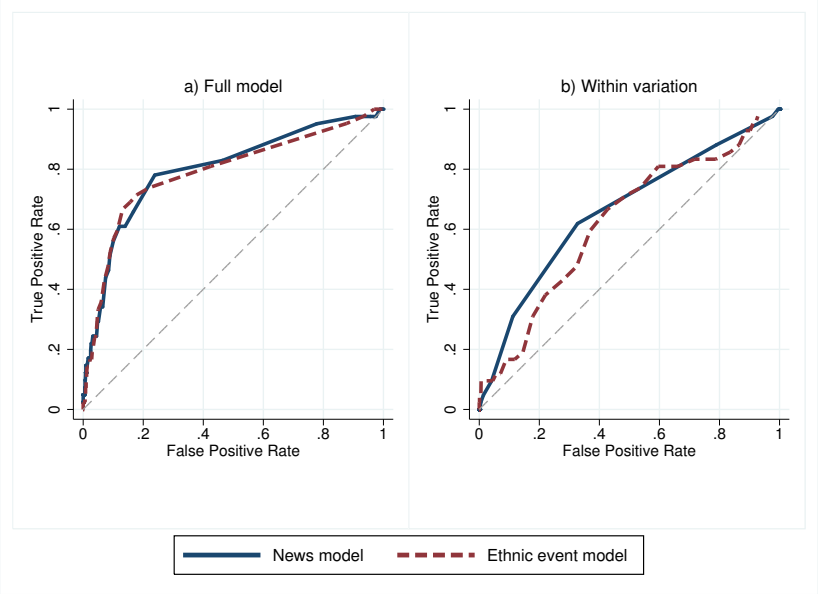


Figure A.14: ROC Curves of Conflict Defined as War Onset (News vs Ethnic Events Model)

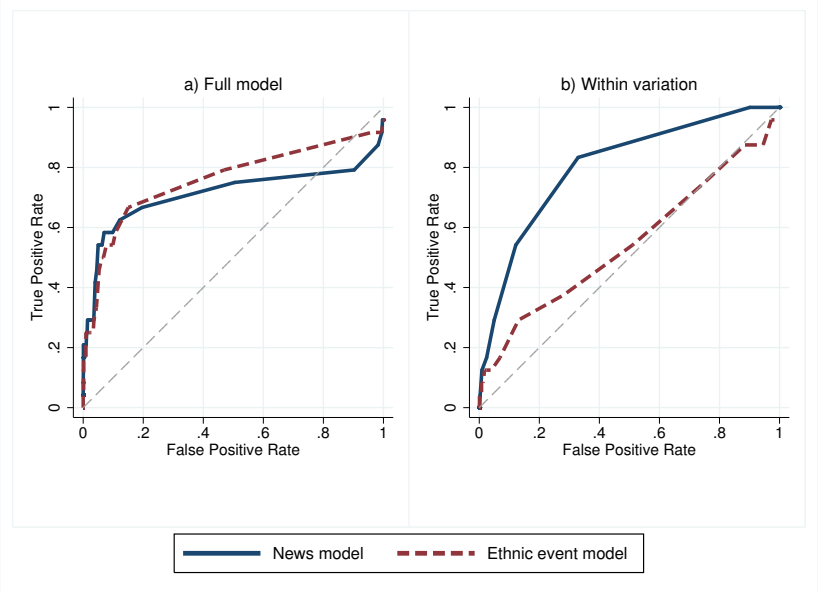


Figure A.15: ROC Curves of Conflict Defined as Armed Conflict Incidence (News vs Insurgency Events Model)

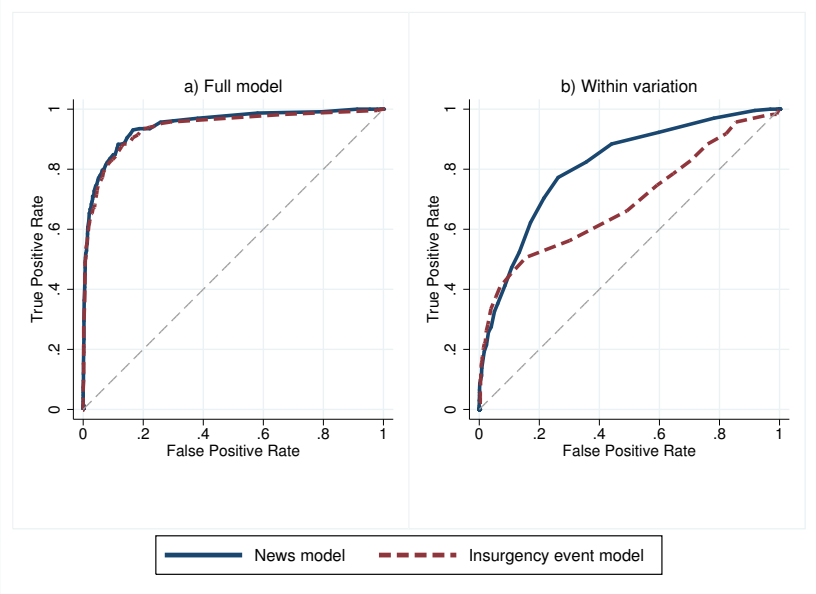


Figure A.16: ROC Curves of Conflict Defined as War Incidence (News vs Insurgency Events Model)

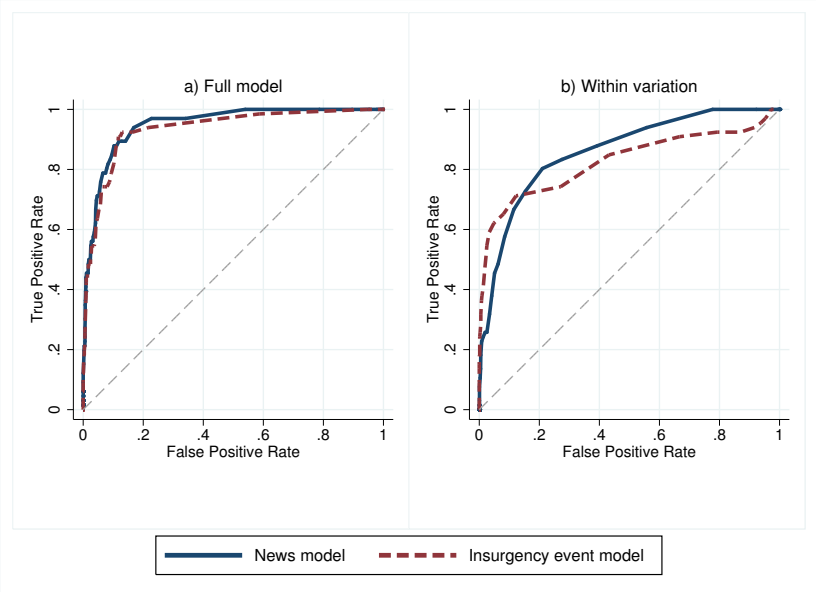


Figure A.17: ROC Curves of Conflict Defined as War Onset (News vs Insurgency Events Model)

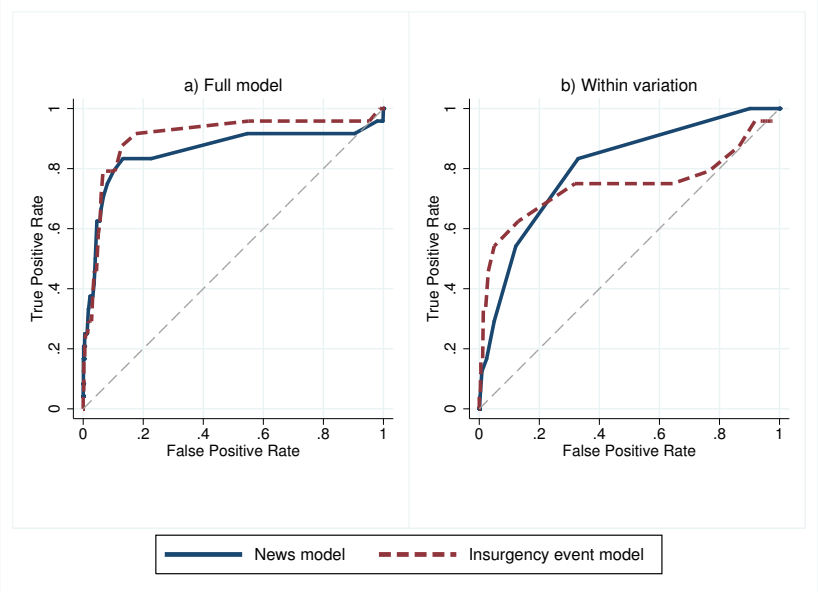


Figure A.18: Conflict Defined as Armed Conflict Incidence (News vs Keyword Model)

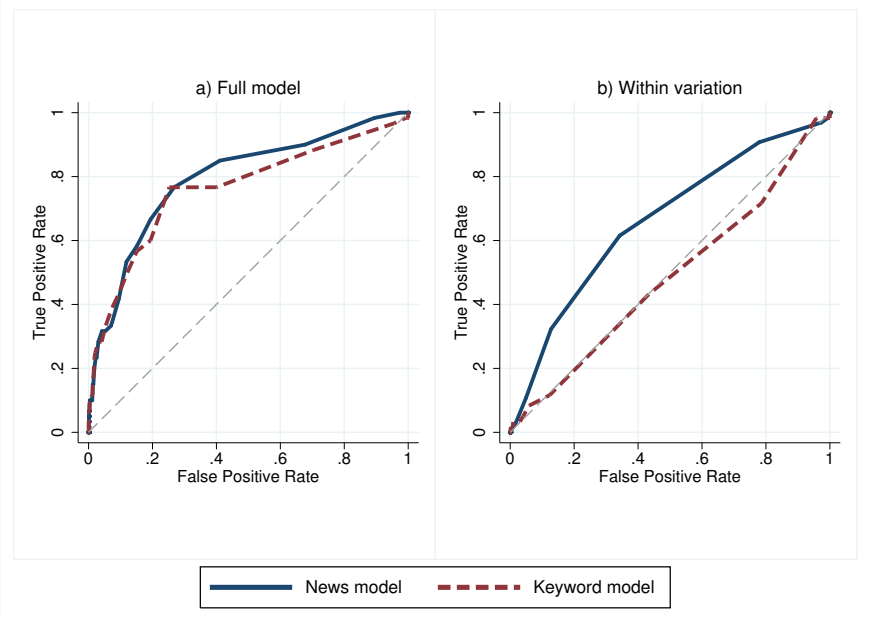


Figure A.19: Conflict Defined as War Onset (News vs Keyword Model)

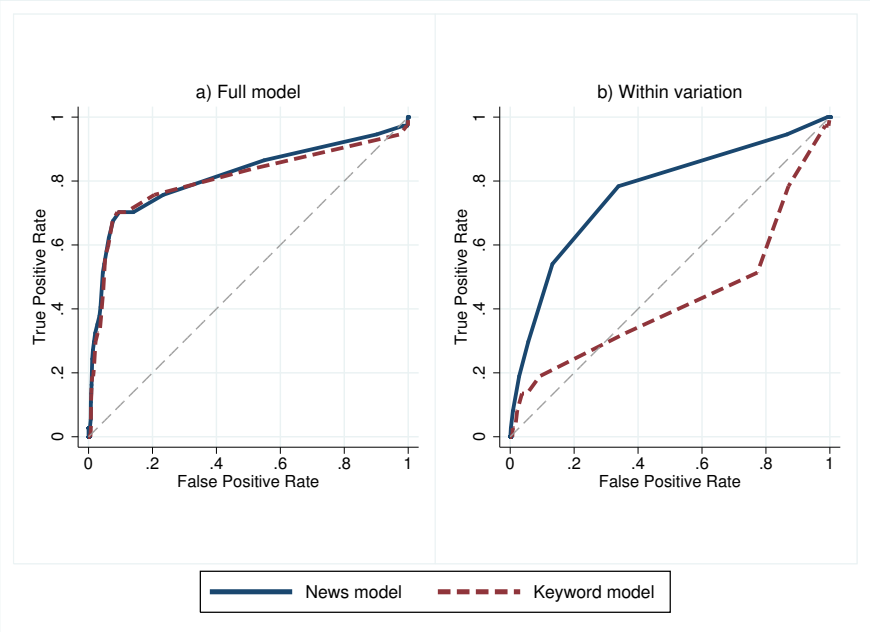
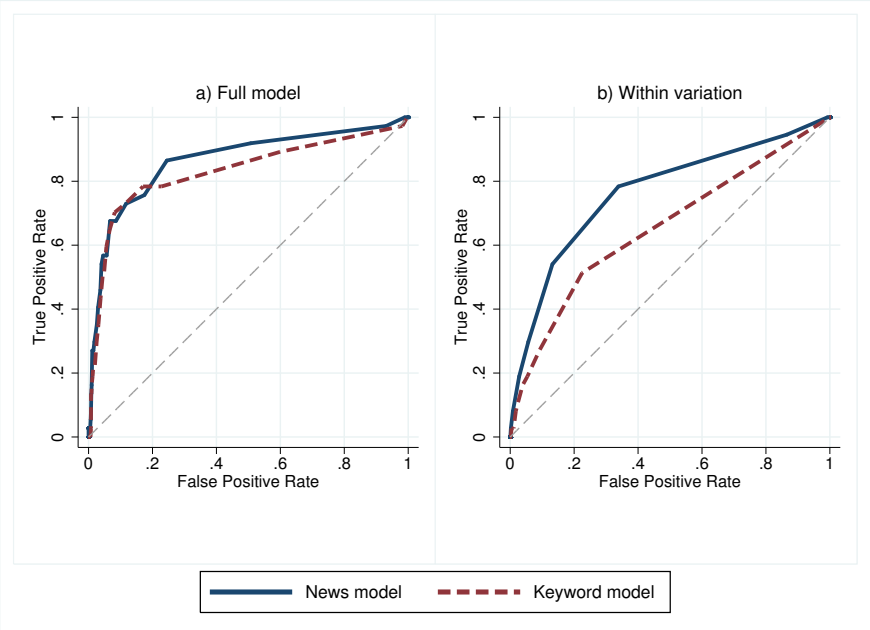


Figure A.20: Conflict Defined as War Onset (News vs Simple Keyword Model)



References

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

Besley, Timothy, and Torsten Persson. 2011. "The Logic of Political Violence." *Quarterly Journal of Economics*, 126(3).

Blair, Robert A, Christopher Blattman, and Alexandra Hartman. 2014. "Predicting Local Violence." *Available at SSRN 2497153*.

Blattman, Christopher, and Edward Miguel. 2010. "Civil war." *Journal of Economic Literature*, 3–57.

Blei, David M, and John D Lafferty. 2006. "Dynamic topic models." 113–120, ACM.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." *the Journal of machine Learning research*, 3: 993–1022.

Brückner, Markus, and Antonio Ciccone. 2010. "International Commodity Prices, Growth and the Outbreak of Civil War in Sub-Saharan Africa*." *The Economic Journal*, 120(544): 519–534.

Brückner, Markus, and Evi Pappa. 2015. "News shocks in the data: Olympic Games and their macroeconomic effects." *Journal of Money Credit and Banking*, forthcoming.

Chadefaux, Thomas. 2014. "Early warning signals for war in the news." *Journal of Peace Research*, 51(1): 5–18.

Dell, Melissa, Benjamin F Jones, and Benjamin A Olken. 2012. "Temperature shocks and economic growth: Evidence from the last half century." *American Economic Journal: Macroeconomics*, 66–95.

Esteban, Joan, Laura Mayoral, and Debraj Ray. 2012. "Ethnicity and conflict: An empirical study." *The American Economic Review*, 1310–1342.

Fearon, James D, and David D Laitin. 2003. "Ethnicity, insurgency, and civil war." *American political science review*, 97(01): 75–90.

Gentzkow, Matthew, and Jesse M Shapiro. 2010. "What drives media slant? Evidence from US daily newspapers." *Econometrica*, 78(1): 35–71.

Goldstone, Jack A, Robert H Bates, David L Epstein, Ted Robert Gurr, Michael B Lustik, Monty G Marshall, Jay Ulfelder, and Mark Woodward. 2010. "A global model for forecasting political instability." *American Journal of Political Science*, 54(1): 190–208.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2014. "Transparency and deliberation within the FOMC: a computational linguistics approach."

Heinrich, Gregor. 2009. "A generic approach to topic models." In *Machine Learning and Knowledge Discovery in Databases*. 517–532. Springer.

Kennedy, Ryan. 2015. "Making useful conflict predictions Methods for addressing skewed classes and implementing cost-sensitive learning in the study of state failure." *Journal of Peace Research*, 52(5): 649–664.

Kuziemko, Ilyana, and Eric Werker. 2006. "How much is a seat on the Security Council worth? Foreign aid and bribery at the United Nations." *Journal of Political Economy*, 114(5): 905–930.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. "Economic shocks and civil conflict: An instrumental variables approach." *Journal of political Economy*, 112(4): 725–753.

Mueller, Hannes. 2014. "Growth and Violence: Argument for a Per Capita Measure of Civil War." Barcelona GSE Working Paper Series Working Paper no 756.

Phan, Xuan-Hieu, and Cam-Tu Nguyen. 2007. "GibbsLDA++: AC/C++ implementation of latent Dirichlet allocation (LDA)."

Porter, Martin F. 1980. "An algorithm for suffix stripping." *Program*, 14(3): 130–137.

Ramey, Valerie A. 2011. "Can government purchases stimulate the economy?" *Journal of Economic Literature*, 49(3): 673–685.

Reynal-Querol, Marta, and Jose G Montalvo. 2005. "Ethnic polarization, potential conflict and civil war." *American Economic Review*, 95(3): 796–816.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. 2013. "The structural topic model and applied social science."

Rost, Nicolas, Gerald Schneider, and Johannes Kleibl. 2009. "A global risk assessment model for civil wars." *Social Science Research*, 38(4): 921–933.

Schutte, Sebastian. 2014. "Regions at Risk Predicting Conflict Zones in African Insurgencies." Mimeo.

Ward, Michael D, Nils W Metternich, Cassy L Dorff, Max Gallop, Florian M Hollenbach, Anna Schultz, and Simon Weschle. 2013. "Learning from the past and stepping into the future: Toward a new generation of conflict prediction." *International Studies Review*, 15(4): 473–490.

Weidmann, Nils B, and Michael D Ward. 2010. "Predicting conflict in space and time." *Journal of Conflict Resolution*, 54(6): 883–901.