M2 EEE Panel Data

Panel Data Replication Project

Andrew Boomer,

Jacob Pichelmann,

Luca Poll

November 27, 2020

Abstract

This paper discusses the merits of using within variation in newspaper text to both predict the timing of conflict and shed light on the underlying drivers of conflict outbreak. We build on the analysis and data construction by Mueller & Rauh (2018) and replicate their key findings, indeed confirming that relying on within variation allows to mitigate the bias of predicting conflict in countries where it occurred before. We exploit the depth of the data to build a dynamic panel model that elucidates causal relationships related to conflict. We argue that variation in newspaper text can serve as a proxy for variation in the latent set of events that trigger conflict. Indeed, we find that the relative number of articles covering a certain topic is successful in capturing shifts in latent events related to conflict, such as global governance.

Table of Contents

1 Introduction

After the Arab spring and the related outbreak of unforeseen violence, conflict forecasting models were largely criticized, and it was argued that forecasting new civil wars might have reached a limit. Mueller and Rauh (2018), however, show in their paper "Reading between the lines: Prediction of political violence", that this might not be entirely true. Their main argument is structured as follows: Conventional conflict forecasting models, which rely on the overall variation in country fixed effect models, exhibit a bias towards predicting conflict onset where it has occurred before. This is partially due to impactful country fixed effects and slow moving factors like population, ethnic fractionalization, climate, etc. that result in a large between variation. The forecasts are therefore dominated by structural time-invariant (or slow moving) factors, neglecting valuable within variation. As a result these models are relatively good at predicting (biasedly) where conflict will happen, but not when it will happen. In order to improve the forecasting of the timing of conflict, and generate an unbiased forecast, Mueller & Rauh (2018) propose isolating the within from the overall variation, using this to predict the onset of armed conflict and civil war. To obtain this within variation, they propose using topic modeling on newspaper text to create variables of the average distribution of topic shares observed in a country during a given year. Taking the derived topic shares as a starting point, this paper exploits the possibility of using the topic shares as regressors for causal inference. They argue that the topic shares capture variation in the latent, high-dimensional set of events that cause conflict, and can be used as proxies for the latter. By doing so, a dynamic panel data model is defined and transformed into first differences which is ultimately estimated by the GMM estimator following the approach by Blundell & Bond (1998).

2 Sample and Data

2.1 Topic Modeling

The pillar this analysis rests on is the news data used to explain and predict conflict. By constructing topics and calculating the share of articles corresponding to each topic for each country and year, the authors succeed in compiling vast amounts of news data in a format that can be used in the subsequent regressions. The exact procedure of topic construction is described in appendix A which also presents an example of topic compositions. Notably the resulting topics each constitute a probability distribution over thousands of words, meaning they have a certain level of depth that might increase their explanatory power, though being hard to intuitively assess. The initial data from which the topic shares are derived are 700.000 newspaper articles from three internationally-reporting newspapers between 1975 and 2015: the Economist, the New York Times and the Washington Post.

2.2 Measuring Conflict

The dependent variables on the other hand are constructed from counts of battle-related deaths obtained from the Uppsala Conflict Data Program (UCDP/PRIO). Following their definition, armed conflict (first dependent variable) is defined as a contested incompatibility that concerns a government and/or territory, over which the use of armed force is between two parties, one being the government of a state, and has resulted in at least 25 battle-related deaths in one calendar year. Civil conflict (second dependent variable) follows the same definition but requires at least 1.000 battle-related deaths in one calendar year.

Panel summary statistics of these variables are shown in appendix B. Notably, the variation in the dependent variable is not homogeneous across countries. In fact, many countries in our sample have not experienced any conflict in the years between 1975 and 2013 and are not likely to do so in the upcoming years. This lack of variation is visual-

ized in figure 2. This poses a challenge in identifying the true coefficients insofar as the estimates are likely to suffer from attenuation bias. The issue and a mitigation strategy using interaction terms are discussed in greater detail in section 4 and appendix D.

2.3 Data Preperation

In the initial analysis the authors change the data in multiple ways prior to estimating the model:

- ▷ Observations with missing values in the topic shares are filled forward. If θ_{it} is missing, and θ_{it-1} is not missing, then $\theta_{it} < -\theta_{it-1}$.
- ▷ The chosen conflict variable itself is not used as the dependent variable. The authors specifically look at two scenarios, either the onset or the incidence of conflict.
  - Onset of conflict is defined as $Conflict_t = 0$ and $Conflict_{t+1} = 1$. After creating this onset variable, all observations where $Conflict_t = 1$ are removed.
  - Incidence of Conflict is defined as $Conflict_t = 1$ and $Conflict_{t+1} = 1$. After creating this incidence variable, missing conflict observations are removed.
- ▷ Observations where the average population over the entire sample is less than 1000, and where population data is missing are removed.

It is important to note that their way of constructing the dependent variable *onset* results in a highly unbalanced panel. We argue that this artificial sample has its merits and legitimacy when used to build a forecasting model, but it should be acknowledged that the removal of observations is clearly deterministic, resulting in non-randomly missing data. Estimating a model on this data will likely results in biased coefficient estimates (see e.g. Wooldridge (2010) p. 581 for a discussion). In our extension of the analysis we hence refrain from replicating this approach and instead use armed conflict itself (see definition above) as the dependent variable.

3 Within Variation for Conflict Prediction

Mueller & Rauh (2018) exploit within variation in topic shares to predict the outbreak of conflict. Relying solely on within variation enables them to mitigate the bias towards countries that experienced conflict before. In each period $T \in \{1995, ..., 2013\}$ they calculate forecasts for an armed conflict/civil war outbreak in period $T + 1$. Importantly the authors focus on the *onset* of conflict, as described in the preceding section. Each forecast uses the full information set up to period T. Therefore, the respective country-year topic shares $\theta_{n,i,T}$ are calculated for every newspaper sub-sample available up to period T[1] for each country i and topic n. As a consequence, the following two steps are repeated at every T:

Step 1: Estimate model and obtain fitted values

From the model $y_{i,T+1} = \alpha + \beta_i + \theta_{i,T}\beta^{topics}$ the fitted values from the estimation based on the overall variation are obtained:

$$\hat{y}^{overall}_{i,T+1} = \hat{\alpha} + \hat{\beta}_i + \theta_{i,T}\hat{\beta}^{topics} \tag{1}$$

From these fitted values that rely on the overall variation, the estimated fixed effects are subtracted in order to obtain the fitted within model:

$$\hat{y}^{within}_{i,T+1} = \hat{\alpha} + \theta_{i,T}\hat{\beta}^{topics} \tag{2}$$

We provide a replication of the model estimation results in table 3 and 4. Additional to the authors' approach, we also report results from pooled OLS estimation for the sake of comparison to the fixed effects approach.[2]

Step 2: Produce forecast based on fitted values for period T+1

The fitted values are transformed into a binary variable indicating either an outbreak of

[1] As the amount of available articles/ words expands in T, the basis for defining a topic through characteristic words in T does also expand. Hence, every topic composition and every topic distribution will vary at every T.

[2] We refrain from interpreting the regression output since the model's purpose lies solely in predicting conflict outbreak.

conflict or no outbreak of conflict. The transformation is based on a range of cutoff values c where for each c the authors calculate the true positive rate and the false positive rate by comparing predictions to actual values. The resulting rates can then be visualized through ROC curves that are used to evaluate the predictive power of the model. We show an exact replication of their results in figure 4. The authors' acknowledge that removing between variation reduces the predictive quality of the model. Still, they succeed in mitigating the bias towards countries where conflict occurred before, while most of the model's predictive power is maintained, as seen in figure 4. In turn their model is more likely to succeed in predicting conflict in formerly peaceful countries - a development other models would most likely be unable to forecast (Mueller & Rauh, 2018).

4 Topic Shares for Causal Inference

Acknowledging the novel approach by Mueller & Rauh (2018), we show that the extracted topic shares can not only be used to predict onset or continuation of conflict, but instead can be used as explanatory variables for conflict. A large body of literature has dealt with the identification of causal relationships related to conflict[3]. Generally, factors like ethnic cleavages[4], climate[5], natural resources[6] or a mix of political and economic indicators[7] have widely been agreed upon. All of these identified factors rely, however, on structural differences between countries to explain conflict. Instead of focusing on these structural parameters that allow conflict to happen, we will exploit the fact that conflicts, regardless of the country, are triggered by a set of (latent) events.

This high-dimensional set of mostly unobserved events is recorded only in small parts by commonly used country level data, while newspapers are more likely to capture a larger proportion of events from this set (say scandals, unrest, resentment and election inconveniences among others). Given the explained depth of the extracted topic shares,

[3]For an overview see Blattman and Miguel (2010)
[4]Reynal-Querol & Montalvo (2005); Esteban, Mayoral & Ray (2012); Caselli & Coleman (2013)
[5]Miguel, Satyanath & Sergenti (2004); Dell, Jones & Olken (2012); Buhaug et al. (2014)
[6]BrÃ¼ckner & Ciccone (2010); Bazzi & Blattman (2014)
[7]Fearon & Laitin (2003); Collier & Hoeffler (2004); Collier et al. (2009); Gleditsch & Ruggeri (2010); Besley & Persson (2011)

the topic shares can be considered as proxy variables capturing variation in the latent set of events across fifteen dimensions. Using the topic shares for causal analysis, however, requires some caution. Most prominently, attempting to explain conflict in period t through the topic shares in period t can result in simultaneity bias. This problem can easily be circumvented by using the lagged topic shares, which by construction cannot be caused by the current conflict outcome, but nevertheless preserve possible explanatory power. Secondly, since the events in period t are very likely to be correlated with the conflict outcomes in prior periods, the topic shares cannot be considered strictly exogenous but only weakly exogenous (predetermined). Additionally, since conflict occurs mostly in countries where it has happened before, we expect a high true serial correlation in conflict outcomes. This motivates the use of the first lag of the dependent variable as a regressor. Finally, besides assuming individual-specific effects that measure unobserved heterogeneity which is correlated to the topic shares (and captures most of the structural differences), a regression of the conflict outcome on the topic shares would suffer from an attenuation bias[8]. This is due to the fact that the occurrence of certain events, and hence the higher observed topic shares resulting from related articles, might have different effects depending on the country individual 'threshold for conflict'. While certain events might immediately lead to conflict in some countries, other countries would not experience an offset for the same events. This difference cannot be identified as a fixed effect, as it is time varying. Since it is not possible to select the countries according to their 'threshold for conflict' without inducing a selection bias, interaction terms can be included in order to distinguish the marginal effects of the fifteen different dimensions of events on the conflict outcome for the respective 'thresholds for conflict'.

[8] A more extensive explanation is given in appendix D.

This dynamic panel data model can be expressed as follows:

$$y_{it} = \gamma\, y_{i,t-1} \;+\; \boldsymbol{\theta}'_{i,t-1}\,\boldsymbol{\beta} \;+\; \boldsymbol{\psi}'_{i,t-1}\,\boldsymbol{\delta} \;+\; \alpha_i \;+\; \lambda_t \;+\; \varepsilon_{it} \tag{3}$$

$$\text{for } i = 1, \cdots, N;\; t = 4, \cdots, T$$

where y_{it} represents the conflict outcome of country i at period t, $y_{i,t-1}$ the conflict realization of country i at period $t-1$ with the autocorrelation coefficient γ, $\boldsymbol{\theta}_{i,t-1}$ a (15×1) vector containing the predetermined topic shares and $\boldsymbol{\beta}$ represents a (15×1) vector of parameters that will be estimated. $\boldsymbol{\psi}_{i,t-1}$ is likewise a (15×1) vector, containing the (also weakly exogenous) interaction terms constructed by the topic shares and varying interaction variables[9] while the (15×1) vector $\boldsymbol{\delta}$ contains the respective coefficients that will be estimated. α_i represents the country specific unobserved fixed effects, λ_t describes the time fixed effects and ε_{it} the serially uncorrelated error terms[10].

5 Estimation of the Model

The presumption that we have serial correlation in our dependent variable can easily be tested by running a simple Pooled OLS estimation of y_t on the lagged realization y_{t-1}. Indeed, as expected, the coefficient of 0.8067 indicates a rather large autocorrelation significant at all conventional significance levels. This autocorrelation can be due to true state dependence induced by the persistence of conflict or spurious state dependence that can be attributed to the unobserved time invariant heterogeneity across the countries. The latter can be expressed as country individual fixed effects which reflect inter alia the populational attitude towards violence or the cultural imprint among many other factors. These factors are most certainly correlated with the observed topic shares, as they are very likely to directly affect the set of events happening in a respective country. Given the correlated country fixed effects, a pooled OLS estimator or random effects estimators would result in inconsistent estimations. Due to the persistence in y_t, the within estimator

[9] A discussion of the selected interaction variables can be found in appendix E.
[10] This assumption is necessary in order to include later on exogenous instruments for $y_{i,t-1}$

would also be inconsistent as it relies on strict exogeneity. In order to remove the country fixed effects as well as the bias, a First Differences model as proposed by Anderson & Hsiao (1981) can be specified. Since in this specification $\varepsilon_{i,t-1}$ is correlated with $y_{i,t-1}$, the OLS estimator would be biased and inconsistent. To bypass this problem, $y_{i,t-1}$ has to be instrumented through internal or external instruments. We follow the approach of Blundell & Bond (1998), which builds on the model proposed by Arellano & Bond (1991)[11] by introducing extra moments through an equation in levels. Due to our assumption of weak exogeneity of the topic shares, we instrument these as well. We define a maximum lag depth of three for the instruments in order to mitigate any possible weak relevance, and to prevent issues of over-identification. This specification results in a system of equations that can be estimated through the Generalized Method of Moments estimator[12].

6 Results

The results of the estimation are presented in table 1. The results are structured as follows: the first column presents the estimated coefficients and their robust standard errors for the simple (biased) model without interaction terms. Columns two to five depict the results of the regressions including child mortality, democracy score, standardized GDP and the score for good institutions respectively.[13] One can note at first glance that the initial autocorrelation coefficient (0.8067 in the simple Pooled OLS model) now dropped considerably. Furthermore, we can clearly see a difference in the estimated autocorrelation term between the simple model and the models including the interaction terms, which indicates that the inclusion of the interaction terms filters out additional spurious state dependence. The reported p-values of the Hansen J test for over-identifying restrictions, however, hint at interpreting the results of the simple model with caution as we do reject the exogeneity of the instruments.

[11] The Blundell-Bond estimator is preferred to the Arellano-Bond Estimator to circumvent the weak instrument problem and to improve efficiency.

[12] The Blundell-Bond estimator relies on the rather strong assumption that $y_{i,1}$ is drawn from a steady state distribution and that α_i is uncorrelated with $y_{i,1}$.

[13] The scale of the topic share variables is such that a one unit increase represents a 100% point increase in the observed topic share.

Taking a closer look at the estimated coefficients of the topic shares and their respective interaction terms, we can differentiate between two prominent patterns across the model specifications: for some topic shares, both the estimated coefficient and its respective interaction term are statistically significant. This pattern indicates that the ceteris paribus marginal effect of an increase in the topic share is two-fold. While the coefficient of the topic share represents the same marginal effect for all countries, the significant interaction term indicates differences in the countries' marginal effect depending on the unobserved threshold for conflict. The second pattern, however, specifies the case where the marginal effect can be generalized for all countries since the topic share coefficients are statistically different from zero while the interaction counterpart is not.

A good example for the first scenario is the topic share of the topic 'international relations 1', which can be interpreted as being related to global governance. Four out of the five models indicate that the general marginal effect of a 10 percentage point increase in the topic share lead to an increase in the probability of experiencing armed conflict by a magnitude ranging from 0.03 to 0.045 percentage points (without considering the biased simple model). At the same time, the negative coefficients of the interaction term with the democracy score, GDP or the 'goodness' score indicate that countries with a higher threshold for conflict are less likely to experience conflict for a ceteris paribus increase in the topic share compared to countries with a lower threshold for conflict. The same line of interpretation can be applied to the topic 'business', where one needs to keep in mind that child mortality is negatively correlated with the threshold for conflict. The second scenario can be illustrated though the topic 'conflict 1' which is likely to capture variation in events related to inner security. An increase in the relative presence of such events, and hence in the observed topic share, results in an increase in the probability of experiencing conflict equally for all countries, regardless of the individual threshold for conflict. For other topics like 'conflict 3', 'economics', 'politics' or 'tourism', however, the scenario differs depending on what model is being considered. Countries can have either the same marginal effect, the same general marginal effect with a threshold for conflict specific component, or show no common marginal effect with the marginal effect

entirely defined by the individual threshold for conflict of the country. Since a detailed performance analysis of the different model specifications would overstep the scope of this paper, it is difficult to say which specification appears to be the most reliable.

Table 1: Regression Results

	Dependent Variable = Armed Conflict				
	Initial	**Child**	**Democ**	**GDP**	**Good**
Lag Armed Conflict	0.5640***	0.2737***	0.3203***	0.2657***	0.2810***
	(0.0235)	(0.0145)	(0.0156)	(0.0165)	(0.0158)
Lag Asia	0.2851**	0.2816**	0.2435*	0.3888***	0.3284***
	(0.1111)	(0.1318)	(0.1352)	(0.1282)	(0.1229)
Lag Interacted Asia		0.0003	0.0168	-0.2809**	-0.5192
		(0.0010)	(0.0300)	(0.1115)	(0.3407)
Lag Business	0.0334	0.2343***	-0.2065**	0.1177**	0.0709
	(0.0473)	(0.0512)	(0.0948)	(0.0584)	(0.0501)
Lag Interacted Business		-0.0020***	0.1149***	-0.0372	0.0460
		(0.0006)	(0.0311)	(0.0434)	(0.0819)
Lag CivLife1	0.0511	0.0171	0.1014	0.1506*	0.1104
	(0.0679)	(0.0808)	(0.1366)	(0.0811)	(0.0768)
Lag Interacted CivLife1		0.0011	-0.0171	-0.1058	0.0336
		(0.0009)	(0.0380)	(0.0822)	(0.1246)
Lag CivLife2	-0.0347	0.0505	0.0907	0.1079**	0.1449***
	(0.0506)	(0.0613)	(0.0825)	(0.0483)	(0.0533)
Lag Interacted CivLife2		0.0008	-0.0032	-0.0365	-0.1643**
		(0.0006)	(0.0253)	(0.0494)	(0.0729)
Lag Conflict1	0.2810***	0.2246**	0.0986	0.2163**	0.2859***
	(0.0933)	(0.0975)	(0.0931)	(0.0926)	(0.0963)
Lag Interacted Conflict1		-0.0003	0.0262	-0.1067	-0.3406*
		(0.0009)	(0.0409)	(0.1225)	(0.1778)
Lag Conflict2	-0.1067	0.1559**	0.0314	0.0546	0.1370**
	(0.1190)	(0.0775)	(0.1013)	(0.0835)	(0.0673)
Lag Interacted Conflict2		-0.0001	0.0402	0.0603	0.0545
		(0.0008)	(0.0359)	(0.1253)	(0.1796)
Lag Conflict3	0.2059***	0.0466	0.1152	0.2097***	0.1829***
	(0.0681)	(0.0764)	(0.0779)	(0.0672)	(0.0641)
Lag Interacted Conflict3		0.0013***	0.0268	-0.1752	0.0629
		(0.0005)	(0.0312)	(0.1164)	(0.1365)
Lag Economics	0.1145*	-0.0058	0.1571	0.1749**	0.2036***
	(0.0626)	(0.0540)	(0.1047)	(0.0681)	(0.0637)
Lag Interacted Economics		0.0023***	-0.0210	-0.0960*	-0.2773***
		(0.0006)	(0.0264)	(0.0502)	(0.0807)
Lag IntRel1	0.2839***	0.1376	0.4555***	0.3567***	0.3085***
	(0.0960)	(0.1027)	(0.1537)	(0.0807)	(0.0836)
Lag Interacted IntRel1		0.0005	-0.1156**	-0.2023***	-0.3934***
		(0.0008)	(0.0515)	(0.0607)	(0.1353)
Lag IntRel2	-0.0345	0.0215	-0.0510	0.1062*	0.0376
	(0.0450)	(0.0402)	(0.0803)	(0.0547)	(0.0493)
Lag Interacted IntRel2		0.0001	0.0263	-0.1309**	0.0141
		(0.0006)	(0.0211)	(0.0605)	(0.0626)
Lag Justice	-0.0377	0.0107	-0.2749**	0.0467	-0.0052
	(0.0685)	(0.0832)	(0.1354)	(0.0625)	(0.0770)
Lag Interacted Justice		-0.0001	0.1045**	-0.0733	-0.0503

		(0.0006)	(0.0460)	(0.0617)	(0.1131)
Lag Politics	-0.0522	-0.0227	0.2752***	0.1222**	0.0540
	(0.0559)	(0.0592)	(0.0964)	(0.0553)	(0.0515)
Lag Interacted Politics		0.0015**	-0.0742**	-0.0618	0.0650
		(0.0006)	(0.0292)	(0.0655)	(0.0830)
Lag Sports	-0.0433	0.1337***	-0.0380	0.1026	0.0577
	(0.0359)	(0.0395)	(0.1195)	(0.0640)	(0.0646)
Lag Interacted Sports		-0.0012	0.0285	-0.0171	-0.0536
		(0.0008)	(0.0284)	(0.0649)	(0.0748)
Lag Tourism	0.1380**	-0.0736	0.4409**	0.2871***	0.2430***
	(0.0577)	(0.0794)	(0.1943)	(0.0881)	(0.0838)
Lag Interacted Tourism		0.0034***	-0.0850	-0.1090	-0.1168
		(0.0008)	(0.0545)	(0.0667)	(0.1245)
Included Effects:	Time	Time	Time	Time	Time
R-Squared:	0.402	0.361	0.364	0.356	0.357
Observations:	8954	8732	8806	8954	8658
Over-Identification p-Val:	0.031	0.151	0.321	0.514	0.426
AB AR Order 2 p-Val:	0.000	0.000	0.000	0.000	0.000
Iterations:	15	15	15	15	15

Robust Standard Errors are Shown in Parentheses
Max Lag Depth = 3

7 Conclusion and Limitations

Mueller & Rauh (2018) showed that the within variation in topic shares distilled from news data offers high predictive power while simultaneously mitigating the forecasting bias towards countries where conflict occurred before. We build on their analysis by using resulting topic shares as proxies for the underlying drivers of conflict: high-dimensional, latent events. We find that indeed the topic shares can not only be used for forecasting but also provide variation in the latent set of events that can be used for causal inference.

Naturally, this approach has its limitations that have to be kept in mind when evaluating and interpreting the results. First of all, we employ a linear probability model, which implies that the fitted values are not bounded between 0 and 1. One way to circumvent this issue is to employ a random effects probit model with a parameterization of the unobserved effects α_i following the approach of Woolridge (2005). We refrain, however, from employing this model considering that the independent variables in our model cannot be considered strictly exogenous and would hence violate the assumptions of the

dynamic probit model which can lead to inconsistent parameter estimates (Chamberlain, 1984). Blindum (2003) offers a discussion on possible mitigation strategies that could be employed in this setting for future research. Secondly, we fail to reject the presence of serial correlation at lag order two. This might indicate that the dynamic specification of our model is flawed. In case of a misspecification the resulting coefficient estimates might be inconsistent. However, this is highly dependent on the structural composition of the error term.[14] For example, Hujer, Zeiss & Rodrigues (2005) provide an in-depth discussion on errors that follow MA processes, in which case the persistence of the serial correlation would be bounded. With bounded persistence of the errors, further investigation of additional models which include multiple lagged dependent variables would be a feasible approach. Another strategy to mitigate this issue would be to start from deeper lags for the GMM moment conditions depending on the assumption of serial correlation (Hujer, Zeiss, & Rodrigues, 2005). However, we do not believe that there is a strong theoretical foundation for why our dependent variable or the errors would follow an MA process, as correlation to past conflicts is more likely than correlation to past shocks in this context. In testing our models, we found evidence of unbounded serial correlation, lending credence to this hypothesis. Given this hypothesis, the approaches by Hujer, Zeiss & Rodrigues (2005) may not be applicable to our setting.

[14] As an example, if we assume a model of the form $y_t = \beta_0 + \beta_1 y_{t-1} + u_t$ where $u_t = \gamma u_{t-2} + \varepsilon_t$ OLS indeed yields consistent coefficient estimates.

References

Anderson, T. W., & Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American statistical Association*, *76*(375), 598–606.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *Review of Economic Studies*, *58*(2), 277-297.

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

Besley, T., & Persson, T. (2011). The logic of political violence. *The quarterly journal of economics*, *126*(3), 1411–1445.

Blattman, C., & Miguel, E. (2010). Civil war. *Journal of Economic literature*, *48*(1), 3–57.

Blindum, S. W., et al. (2003). *Relaxing the strict exogeneity assumption in a dynamic random probit model* (Tech. Rep.). University of Copenhagen. Department of Economics. Centre for Applied â¦.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, *87*(1), 115-143.

Brückner, M., & Ciccone, A. (2010). International commodity prices, growth and the outbreak of civil war in sub-saharan africa. *The Economic Journal*, *120*(544), 519–534.

Buhaug, H., Nordkvelle, J., Bernauer, T., Böhmelt, T., Brzoska, M., Busby, J. W., ... others (2014). One effect to rule them all? a comment on climate and conflict. *Climatic Change*, *127*(3-4), 391–397.

Caselli, F., & Coleman, W. J. (2013). On the theory of ethnic conflict. *Journal of the European Economic Association*, *11*(suppl_1), 161–192.

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

Chamberlain, G. (1984). Panel data. *Handbook of econometrics*, *2*, 1247–1318.

Collier, P., & Hoeffler, A. (2004). Greed and grievance in civil war. *Oxford economic papers*, *56*(4), 563–595.

Collier, P., Hoeffler, A., & Rohner, D. (2009). Beyond greed and grievance: feasibility and civil war. *oxford Economic papers*, *61*(1), 1–27.

Dell, M., Jones, B. F., & Olken, B. A. (2012). Temperature shocks and economic growth: Evidence from the last half century. *American Economic Journal: Macroeconomics*, *4*(3), 66–95.

Esteban, J., Mayoral, L., & Ray, D. (2012). Ethnicity and conflict: Theory and facts. *science*, *336*(6083), 858–865.

Fearon, J. D., & Laitin, D. D. (2003). Ethnicity, insurgency, and civil war. *American political science review*, 75–90.

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

Hujer, R., Zeiss, C., & Rodrigues, P. J. M. (2005). *Serial correlation in dynamic panel data models with weakly exogenous regressor and fixed effects*. Universitätsbibliothek Johann Christian Senckenberg.

Miguel, E., & Satyanath, S. (2011). Re-examining economic shocks and civil conflict. *American Economic Journal: Applied Economics*, *3*(4), 228–32.

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

Montalvo, J. G., & Reynal-Querol, M. (2005). Ethnic polarization, potential conflict, and civil wars. *American economic review*, *95*(3), 796–816.

Mueller, H., & Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text.

Skrede Gleditsch, K., & Ruggeri, A. (2010). Political opportunity structures, democracy, and civil war. *Journal of Peace Research*, *47*(3), 299–310.

Ward, M. D., Metternich, N. W., Dorff, C. L., Gallop, M., Hollenbach, F. M., Schultz, A., & Weschle, S. (2013). Learning from the past and stepping into the future: Toward a

new generation of conflict prediction. *International Studies Review*, *15*(4), 473–490.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of applied econometrics*, *20*(1), 39–54.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

8 Appendix

A Construction of Topic Shares

Mueller & Rauh (2018) use an unsupervised learning algorithm to distill topic shares out of the set of 700.000 newspaper articles.[15] They start by processing the articles' contents with standard text mining techniques such as stemming words.[16] This leaves them with roughly 0.9 million tokens, which are then grouped into topics based on the latent Dirichlet allocation (LDA) method. A topic then constitutes a probability distribution over words. The result is intuitive, as one can imagine that an article covering "Sports" might indeed be more likely to contain words such as "score", "win" and "match" whereas an article concerned with "Conflict" could contain the phrases "war", "protest" and "military". An indication of the resulting topic compositions is given in figure 1. The number of topics has to be specified beforehand, while the composition of topics is defined by the algorithm. The authors choose to work with a final set of 15 topics. Notably, each topic is a probability distribution over thousands of words, meaning the resulting topics have a certain level of depth that might increase their explanatory power, although being hard to intuitively assess. A general overview of the evolution of shares over the observed time periods is given by figure 3.

[15]
- The Economist: 174.450 articles from 1975 onward
- The New York Times: 363.275 articles from 1980 onward
- The Washington Post: 185.523 articles from 1977 onward

[16] Stemming refers to the process of finding the common root of a word, i.e. "running", "ran", and "run" all become "run".

Figure 1: Topic content 2013

Topic title	15 most prominent words
Industry	compani, busi, market, firm, year, japanes, new, industri, share, american, sale, million, billion, manag, make
Civic life1	peopl, year, work, famili, say, child, woman, school, home, live, life, like, univers, old, citi
Asia	china, chines, south, korea, offici, year, taiwan, north, vietnam, hong, kong, hong_kong, beij, foreign, govern
Sports	team, game, play, second, year, point, time, world, won, player, win, score, run, final, minut
Justice	offici, report, court, state, case, charg, polic, investig, govern, law, offic, prison, arrest, releas, author
Tourism	citi, like, hotel, street, room, restaur, travel, place, good, food, hour, time, open, new, hous
Politics	parti, elect, polit, govern, vote, democrat, prime, new, power, leader, parliament, opposit, support, year, campaign
Conflict1	govern, peopl, countri, protest, polit, group, leader, islam, militari, polic, state, year, demonstr, support, muslim
Business	oil, year, countri, trade, world, state, import, export, produc, develop, price, govern, product, plant, new
Economics	bank, year, rate, govern, economi, billion, countri, market, econom, price, percent, tax, growth, fund, money
Inter. relations1	state, unit, unit_state, american, offici, administr, washington, nation, nuclear, meet, militari, bush, weapon, secur, talk
Inter. relations2	soviet, european, union, german, germani, europ, west, countri, east, britain, western, new, british, soviet_union, foreign
Conflict3	govern, nation, war, forc, rebel, unit, african, refuge, unit_nation, countri, south, peac, peopl, serb, guerrilla
Civic life2	work, new, like, book, world, time, film, art, life, cultur, year, music, american, centuri, war
Conflict2	forc, attack, militari, kill, offici, arab, armi, american, report, bomb, troop, soldier, war, command, air

Source: Mueller & Rauh (2018), p. 80

B Descriptive Statistics

Variable	Type	Mean	Std. Dev.	Min	Max	Observations
Armed Conflict	**overall**	0.142	0.349	0.000	1.000	7520
	between		0.009	0.106	0.186	188
	within		0.243	-0.833	1.117	40
Civil War	**overall**	0.060	0.237	0.000	1.000	7520
	between		0.011	0.027	0.112	188
	within		0.191	-0.790	1.035	40
ChildMortality	**overall**	66.490	66.912	1.900	368.300	7301
	between		11.112	32.082	115.391	183
	within		30.650	-87.312	260.334	41
RealGDP	**overall**	9795.056	12252.869	160.797	136311.016	6211
	between		863.233	6714.035	13445.277	185
	within		3882.562	-14142.336	82558.156	36
DemocracyIndex	**overall**	2.684	1.551	0.000	5.000	6355
	between		0.210	2.006	3.148	162
	within		0.914	-2.191	5.684	40
AveGoodIndex	**overall**	0.279	0.397	0.000	1.000	4992
	between		0.000	0.279	0.279	156
	within		0.000	0.279	0.279	32
Industry	**overall**	0.053	0.039	0.007	0.560	6639
	between		0.002	0.046	0.063	185
	within		0.028	-0.139	0.552	39
CivLife1	**overall**	0.073	0.041	0.010	0.559	6639
	between		0.005	0.050	0.089	185
	within		0.036	-0.110	0.547	39
Asia	**overall**	0.043	0.049	0.006	0.454	6639
	between		0.002	0.038	0.051	185
	within		0.022	-0.193	0.380	39
Sports	**overall**	0.060	0.068	0.009	0.663	6639
	between		0.006	0.032	0.080	185
	within		0.048	-0.385	0.630	39
Justice	**overall**	0.069	0.045	0.004	0.468	6639
	between		0.004	0.045	0.081	185
	within		0.038	-0.082	0.442	39
Tourism	**overall**	0.063	0.052	0.009	0.765	6639
	between		0.005	0.036	0.081	185
	within		0.043	-0.127	0.715	39
Politics	**overall**	0.074	0.047	0.007	0.514	6639
	between		0.003	0.063	0.086	185
	within		0.043	-0.040	0.492	39
Conflict1	**overall**	0.070	0.052	0.007	0.426	6639
	between		0.003	0.058	0.084	185
	within		0.039	-0.054	0.404	39
Business	**overall**	0.074	0.054	0.010	0.514	6639
	between		0.005	0.058	0.116	185
	within		0.045	-0.102	0.452	39
Economics	**overall**	0.065	0.051	0.007	0.612	6639
	between		0.004	0.053	0.092	185
	within		0.043	-0.043	0.606	39
IntRel1	**overall**	0.063	0.046	0.005	0.407	6639
	between		0.005	0.047	0.082	185
	within		0.035	-0.089	0.340	39

IntRel2	**overall**	0.075	0.069	0.004	0.653	6639
	between		0.008	0.058	0.135	185
	within		0.041	-0.127	0.609	39
Conflict3	**overall**	0.089	0.090	0.008	0.623	6639
	between		0.004	0.070	0.103	185
	within		0.052	-0.164	0.564	39
CivLife2	**overall**	0.067	0.048	0.007	0.582	6639
	between		0.002	0.058	0.076	185
	within		0.040	-0.085	0.517	39
Conflict2	**overall**	0.061	0.055	0.006	0.437	6639
	between		0.003	0.048	0.075	185
	within		0.033	-0.107	0.386	39
Topic Year	**overall**	2013.000	0.000	2013.000	2013.000	7707
	between		0.000	2013.000	2013.000	188
	within		0.000	2013.000	2013.000	41

Figure 2: Variation in Armed Conflict across Countries

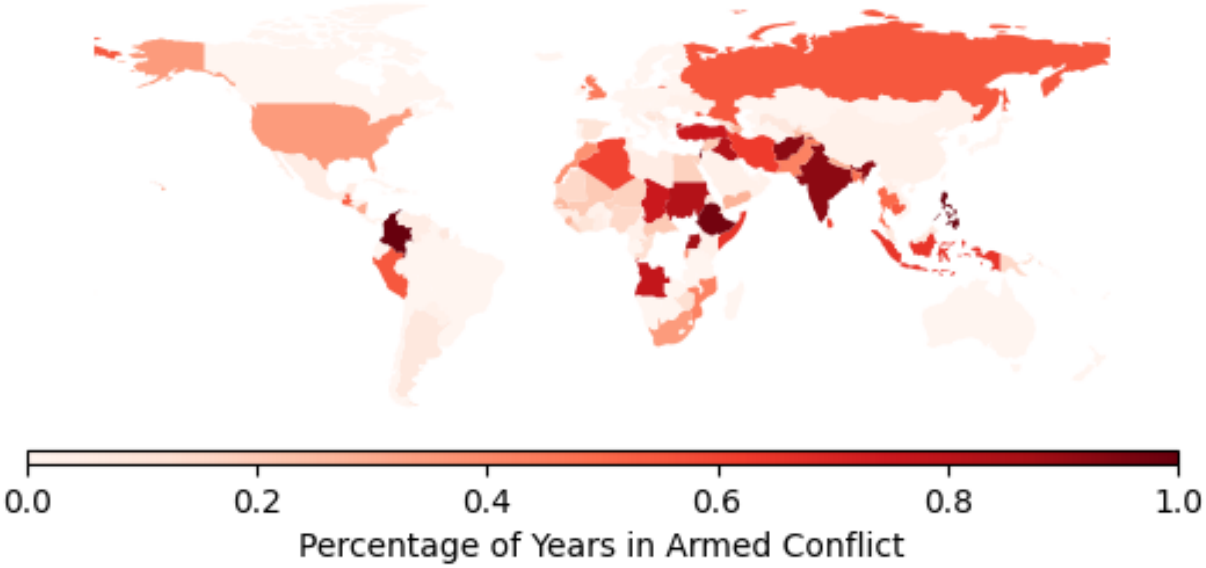
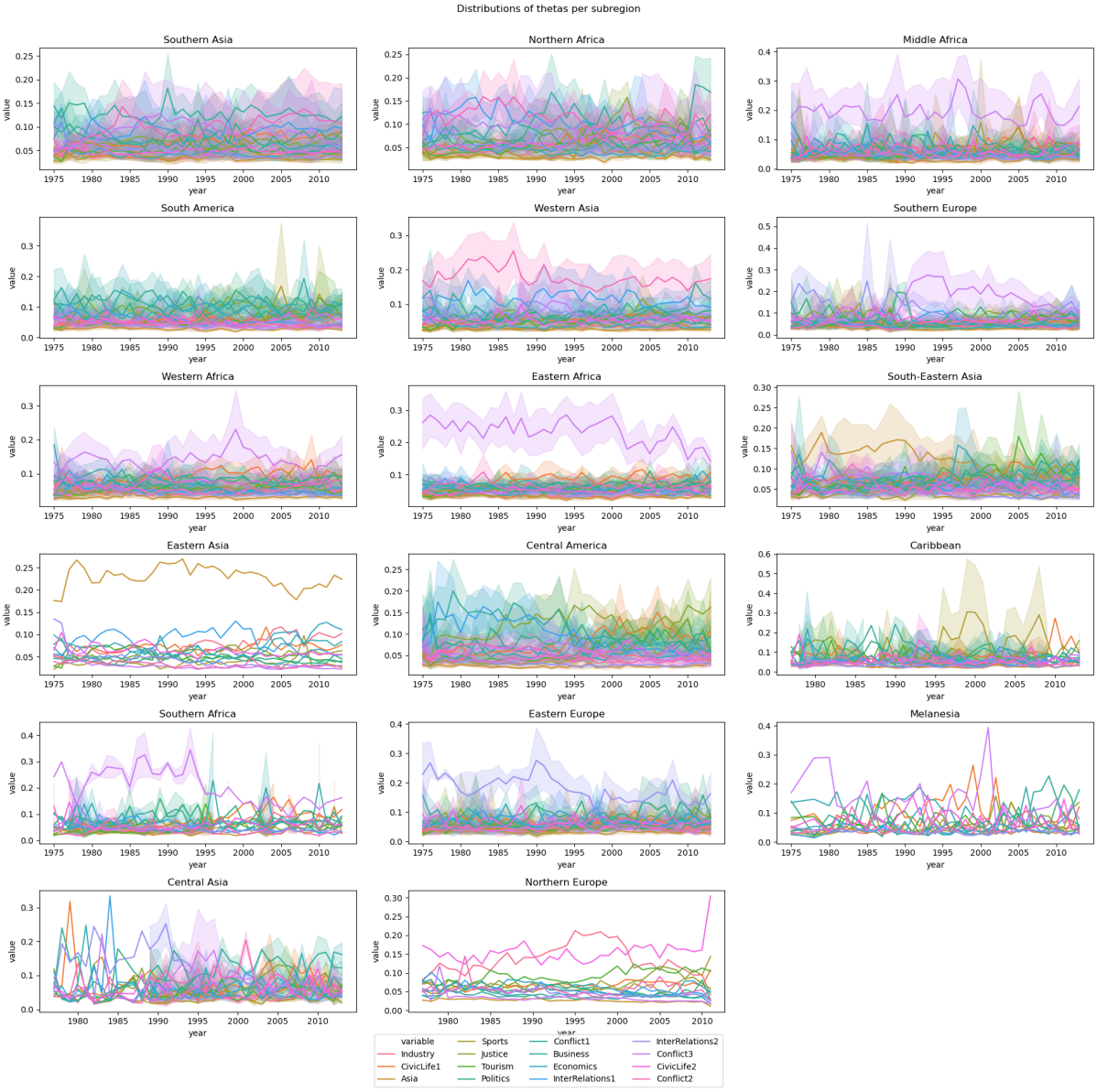
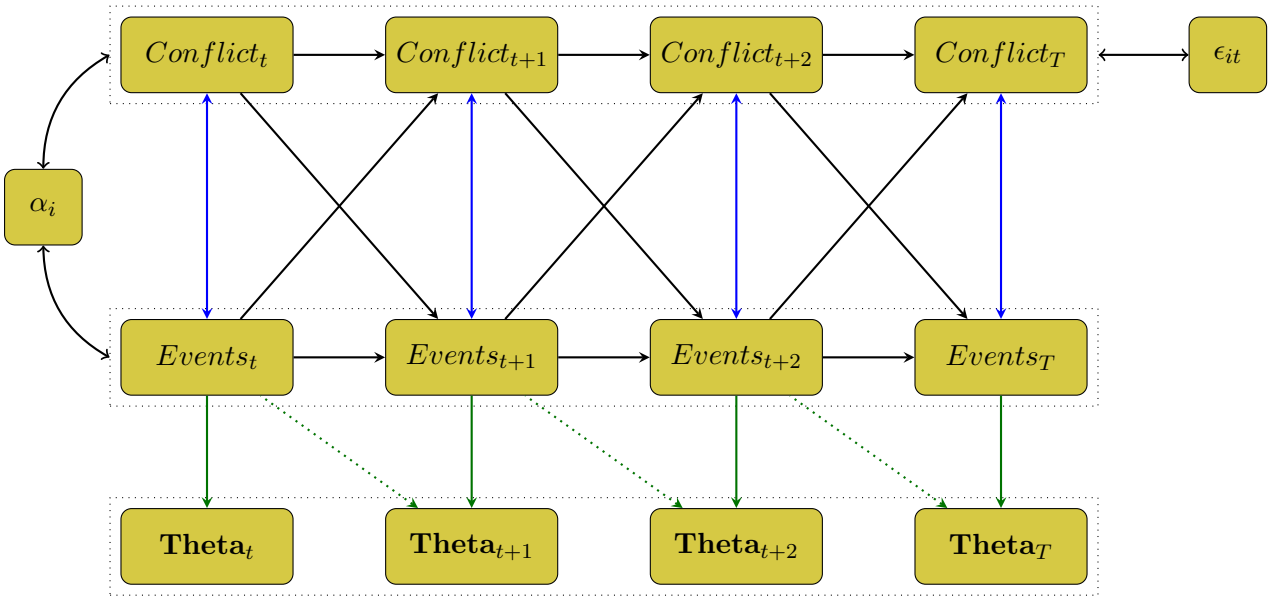


Figure 3: Evolution of Topic Shares over Time by Region




$Conflict_t$
$Conflict_{t+1}$
$Conflict_{t+2}$
$Conflict_T$
ϵ_{it}
α_i
$Events_t$
$Events_{t+1}$
$Events_{t+2}$
$Events_T$
$\mathbf{Theta}_t$
$\mathbf{Theta}_{t+1}$
$\mathbf{Theta}_{t+2}$
$\mathbf{Theta}_T$

C Replication Results

Table 3: Estimating Onset and Incidence of Armed Conflict

	Dependent Variable = Armed Conflict			
	POLSOnset	**FEOnset**	**POLSIncidence**	**FEIncidence**
CivLife1	0.1181	-0.0023	-0.3401	-0.2972
	(0.1126)	(0.1258)	(0.2895)	(0.2094)
IntRel1	0.1297	-0.0299	0.3802	0.0863
	(0.1311)	(0.1354)	(0.3751)	(0.2583)
IntRel2	-0.0154	0.0092	-0.6449**	-0.2180
	(0.0765)	(0.1034)	(0.2675)	(0.1924)
Conflict3	0.2401***	0.1666	0.6237**	0.8491***
	(0.0854)	(0.1164)	(0.2999)	(0.2083)
CivLife2	0.1200	0.0473	-0.2574	-0.1144
	(0.0970)	(0.1056)	(0.2586)	(0.1458)
Conflict2	0.2946***	0.3849***	1.1284**	1.0185***
	(0.1093)	(0.1415)	(0.5458)	(0.2854)
Asia	0.0134	0.1930	-0.3014	-0.0512
	(0.0835)	(0.1560)	(0.4367)	(0.3916)
Sports	-0.0090	0.0131	-0.5305**	-0.1645
	(0.0772)	(0.0991)	(0.2294)	(0.1436)
Justice	-0.0967	-0.0897	-0.1949	-0.3126
	(0.0950)	(0.1393)	(0.5624)	(0.2401)
Tourism	0.0101	0.0920	-0.2812	-0.2310
	(0.0923)	(0.1145)	(0.3172)	(0.1943)
Politics	-0.0199	0.0482	-0.6622**	-0.3630**
	(0.0881)	(0.1160)	(0.2886)	(0.1573)
Conflict1	0.3885***	0.3360***	0.6203*	0.2193
	(0.1035)	(0.1280)	(0.3304)	(0.2258)
Business	0.0946	0.1615	-0.6745**	-0.2840*
	(0.0845)	(0.1184)	(0.2687)	(0.1570)
Economics	0.0220	0.0333	-0.4033	-0.2527*
	(0.1190)	(0.1434)	(0.2878)	(0.1488)
Constant	-0.0499	-0.0497	0.2467	0.1494
	(0.0678)	(0.0941)	(0.2237)	(0.1268)
Included Effects:	Time	Entity, Time	Time	Entity, Time
R-Squared:	0.027	0.010	0.187	0.097
Observations:	4486	4486	5499	5499

Robust Standard Errors are Shown in Parentheses

Table 4: Estimating Onset and Incidence of Civil War

	Dependent Variable = Civil War			
	POLSOnset	**FEOnset**	**POLSIncidence**	**FEIncidence**
CivLife1	-0.0396	-0.0477	-0.2777**	-0.3656**
	(0.0560)	(0.0726)	(0.1334)	(0.1432)
IntRel1	-0.0364	-0.1777*	0.1289	-0.0789
	(0.0737)	(0.0955)	(0.1895)	(0.1943)

IntRel2	-0.0324	-0.0010	-0.2468*	-0.2116**
	(0.0404)	(0.0723)	(0.1403)	(0.0956)
Conflict3	0.1462**	0.1853**	0.4173**	0.3846***
	(0.0614)	(0.0833)	(0.1668)	(0.1396)
CivLife2	-0.0428	-0.0234	-0.2888**	-0.1908*
	(0.0441)	(0.0570)	(0.1380)	(0.1115)
Conflict2	0.2316***	0.3093***	0.5673	0.6557**
	(0.0884)	(0.1040)	(0.3546)	(0.2928)
Asia	-0.0767*	-0.1608*	-0.0834	0.0046
	(0.0419)	(0.0841)	(0.2456)	(0.2787)
Sports	-0.0438	-0.0289	-0.0955	-0.1190
	(0.0390)	(0.0534)	(0.1046)	(0.0970)
Justice	-0.0401	-0.1311	-0.3439*	-0.4580***
	(0.1060)	(0.0804)	(0.1772)	(0.1643)
Tourism	-0.0585	-0.0025	-0.1746	-0.2418**
	(0.0526)	(0.0659)	(0.1220)	(0.1169)
Politics	-0.0790	-0.0620	-0.4303***	-0.3388***
	(0.0544)	(0.0694)	(0.1576)	(0.1051)
Conflict1	0.0377	0.0702	0.2549*	-0.0558
	(0.0449)	(0.0728)	(0.1546)	(0.1491)
Business	-0.0856*	-0.0389	-0.3481***	-0.2548**
	(0.0471)	(0.0683)	(0.1277)	(0.1176)
Economics	-0.0609	-0.0314	-0.1339	-0.1475
	(0.0405)	(0.0520)	(0.1671)	(0.1248)
Constant	0.0296	0.0241	0.1293	0.1566*
	(0.0328)	(0.0490)	(0.1107)	(0.0862)
Included Effects:	Time	Entity, Time	Time	Entity, Time
R-Squared:	0.028	0.018	0.113	0.058
Observations:	5062	5062	5499	5499

Robust Standard Errors are Shown in Parentheses

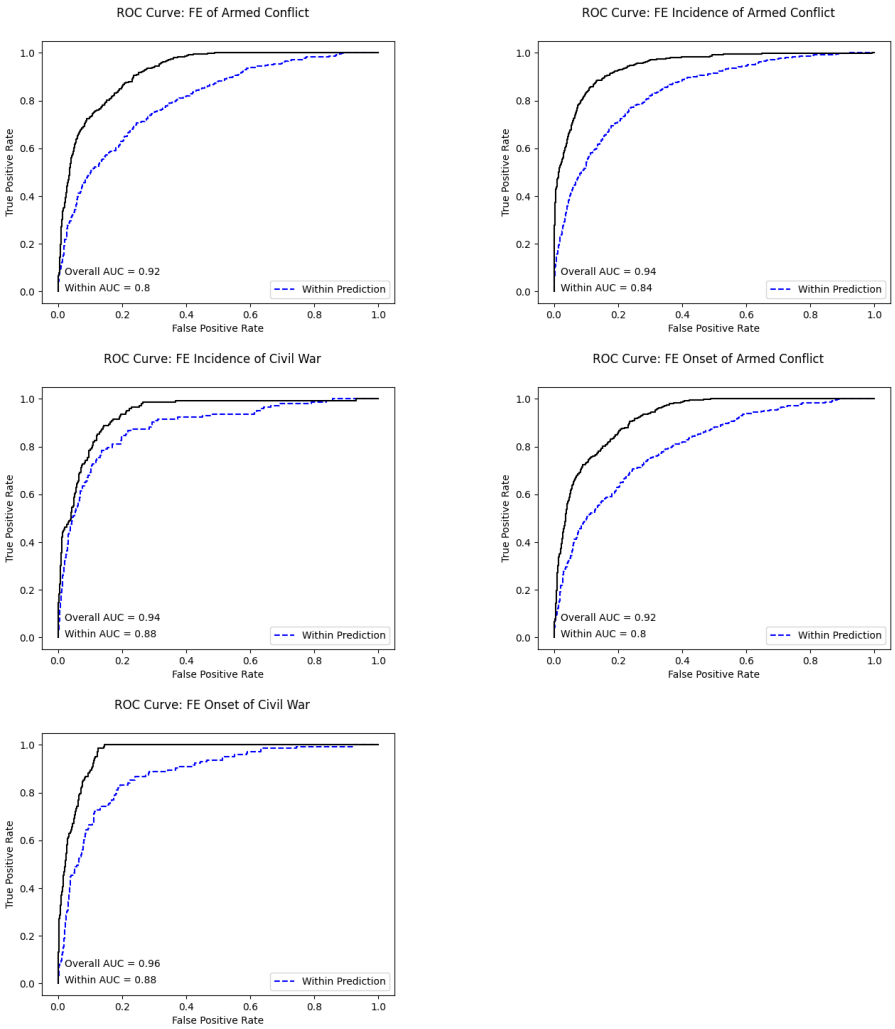
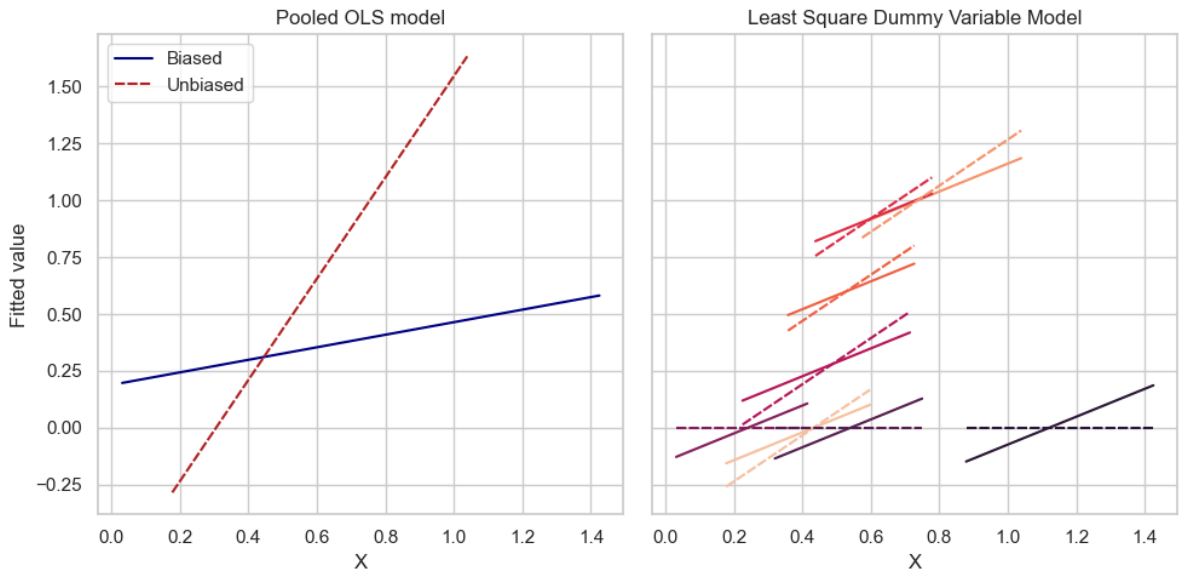


Figure 4: Prediction Evaluation

D Attenuation Bias

As briefly introduced in Section 4, the regression of conflict outcomes on topic shares suffers from an attenuation bias. This is due to the fact that countries might have different thresholds for conflict which, in contrast to the country fixed effects, can vary over time. The intuition behind the bias is that the conflict outcome for countries with a high threshold for conflict (say Norway, Switzerland, Costa Rica) is seemingly uncorrelated with events that likely cause conflict in other countries (say Afghanistan, Philippines, Iraq). The countries with a high threshold can nevertheless not be excluded from the sample since that would introduce a selection bias. The effect of the attenuation bias is shown through a simulation in Figure 5, where the biased estimation is the estimation of the full sample and the unbiased estimation does not consider selected countries with a high threshold for conflict (slope = 0).

Figure 5: Simulation Attenuation Bias



In order to circumvent the described bias, we can interact the topic shares with variables that reflect possible differences in the threshold for conflict and hence allow to separate the effects of specific events on the realized conflict outcome.

E Interaction Terms

To account for the different thresholds of conflict, we propose four different variables that are to be interacted with the topic shares: child mortality[17], democracy index[18], real GDP[19] and "goodness index"[20]. All four variables have in common that they can be seen as measures of how stable or progressive a country can be considered. While a country with strong institutions and a higher level of development is more likely to have low child mortality, a high real GDP, a high score in the democracy index as well as as high score in the "goodness index", the opposite is likely to hold true for countries with a comparably low level of development or weak institutions. Figure 6 gives an impression of the direction of the correlation between conflict and the respective variables.

[17]Source: World Bank Open Data
[18]Source: The Economist
[19]Source: World Bank Open Data
[20]Source: Reconstructed by Mueller Rauh (2018) based on Besley and Persson (2011). This indicator provides a measure of good institutions.

Figure 6: Correlation Interaction Variables and Conflict Outcome