M2 EEE Panel Data

Panel Data Replication Project

Andrew Boomer,

Jacob Pichelmann,

Luca Poll

October 30, 2020

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) though 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[1], that rely on the overall variation in country fixed effect models, exhibit a bias towards predicting conflict onset to where conflict has occurred before. This is partially due to large country fixed effects and slow moving factors like population, ethnic fractionalization, climate, etc. that result in a large between variation. The forecasts are hence 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 to isolate the within from the overall variation and use such to predict the onset of armed conflict and civil war. In order to obtain necessary 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.

[1] They demonstrate their argument by replicating the following papers on conflict prediction:

- ▷ Miguel & Satyanath (2011): Prediction through rainfall growth
- ▷ Besley & Presson (2011): Prediction through proxies for external shocks and political constraints
- ▷ Goldstone et al. (2010): Prediction through political institution dummies, child mortality rates, share of population discriminated against and whether neighboring countries in conflict
- ▷ Ward et al. (2013): Event database on high-intensity and low-intensity conflict events used for analysis
- ▷ Chadefaux (2014): Conflict prediction through analysis of keyword count in newspaper text

2 Sample & Data

The sample for the underlying empirical analysis consists of 700.000 newspaper articles from three internationally-reporting newspapers between 1975 and 2015: the Economist[2], the New York Times[3] and the Washington Post[4]. The newspapers cover in total 185 countries and the average yearly coverage amounts to 120 articles per country (with a range from 1 to 5.500). The authors use an unsupervised learning algorithm to break these articles into 15 distinct topic groups.

The dependent variables on the other hand are constructed through battle-related deaths from the Uppsala Conflict Data Program (UCDP/PRIO). Following their definition, armed conflict (dep. var. 1) is defined as a contested incompatibility that concerns government and/or territory over which the use of armed force between two parties, of which at least one is the government of a state, has resulted in at least 25 battle-related deaths in one calendar year. Civil conflict (dep. var. 2) follows the same definition but requires at least 1.000 battle-related deaths in on calendar year.

The panel summary statistics for these variables are given in Figure 1. (Provide futher explanation about the data)

2.1 Data Preparation for Model

The authors clean and prepare their data before estimation. Some of these techniques we agree with, and others we have some theoretical issues with. The pros and cons of their methods will be discussed in further detail after the initial replication section.

- ▷ Observations with missing values in the topic shares are filled forward. If θ_{it} is missing, and θ_{it-1} is not missing, then $\theta_{it} < -\theta_{it-1}$.

[2] 174.450 articles from 1975 onward
[3] 363.275 articles from 1980 onward
[4] 185.523 articles from 1977 onward

- ▷ 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.
  - Inicidence of Conflict is defined as $Conflict_t = 1$ and $Conflict_{t+1} = 1$. After creating this incidence variable, missing conflict observations are removed.
  - In our replication, we will narrow our focus to only the onset of conflict as the authors define it.
- ▷ Observations where the average population over the entire sample is less than 1000, and where population data is missing are removed.
- ▷ Observations where there are zero words written, or where this data is missing, are removed.
- ▷ As a robustness check, the authors provide the option to restrict the sample to only countries who have experienced conflict at least once in the entire sample.

3 Model

The aim of the model is to create forecasts for an armed conflict/ civil war outbreak in period $T+1$ at period $T \in \{1995, ..., 2013\}$. To create this forecast, the full information set up to period T is included into the forecast. Therefore, the respective country-year topic shares $\theta_{n,i,T}$ are calculated for every newspaper sub-sample available up to period T[5] 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

[5] 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, the every topic characteristic and every topic distribution will vary at every T

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:

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

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

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

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

1) The fitted values are transformed into binary variables depending on cutoff value c

2) Compare forecast (binary variable) to realizations of armed conflict and civil war

3) Assess performance of overall and within model by considering forecasting performance for any given value c through ROC curves

4 Replication Estimations

5 Extensions

After replicating the core findings of the paper, we plan to extend it in several dimensions, following a set of four steps.

5.1 Step I: Extended Data Analysis

Firstly, we will place emphasis on investigating the characteristics of the data at hand, i.e. analyzing the distribution of topic shares over time and their correlation with certain confounders. We control for possible correlations with country specific characteristics such as continents/regions and rainfall and check whether or not there exist outliers (e.g. countries with just one article a year or where coverage only occurred during conflict). We aim to build a sound argument for the resulting estimation method and try to answer if additional measures, such as employing panel robust standard errors, should be taken into consideration. We expect to additionally validate the authors' econometric approach with this exercise.

5.2 Step II: Enhanced Estimation

Based on the findings in step I we advance by employing different models in order to increase the estimates precision. Depending on said findings this will include:

- ▷ Conditional logit - including a discussion of necessary (strict) assumptions (i.e. fixed effects of a country that either always or never experience conflict)
- ▷ Random Effects - the assumption that α_i is uncorrelated with the regressors (i.e. topic shares) is unlikely to hold. However, when interpreting the random effects estimator as a weighted average of within and between estimation its performance in conjunction with the results of the fixed effects estimation might allow for a clearer assessment of the underlying factors at play. Additionally, the random effects model allows for a straightforward out of sample prediction.
- ▷ First Differences - relaxing the strict exogeneity assumption of the fixed effects model

5.3 Step III: Data Changes

The key goal of the paper is to develop a forecasting strategy that allows to predict conflict in formerly peaceful regions, i.e. a regime change. We plan to additionally assess our model performance by building another dependent variable which captures regime change itself and regress on this newly constructed variable.

Moreover, we plan to include further lags of topics in order to capture a possible 'build up' of conflict over the preceding years. However, the topic composition might lack the level of granularity necessary to capture these patterns.

5.4 Step IV: Enhanced discussion

Finally our analysis aims at discovering possible additional sources of bias, to better understand model performance. However, given the data constraints, we do not expect to resolve these issues but rather add to the discussion. For example, we implicitly assume that the newspapers have unbiased reporting. This assumption is likely not to hold. We will investigate options to control for an author specific bias. Additionally, we will further investigate the (changing) importance of topics over the years. This exercise aims at distilling a set of topics that is predictive for each time span.

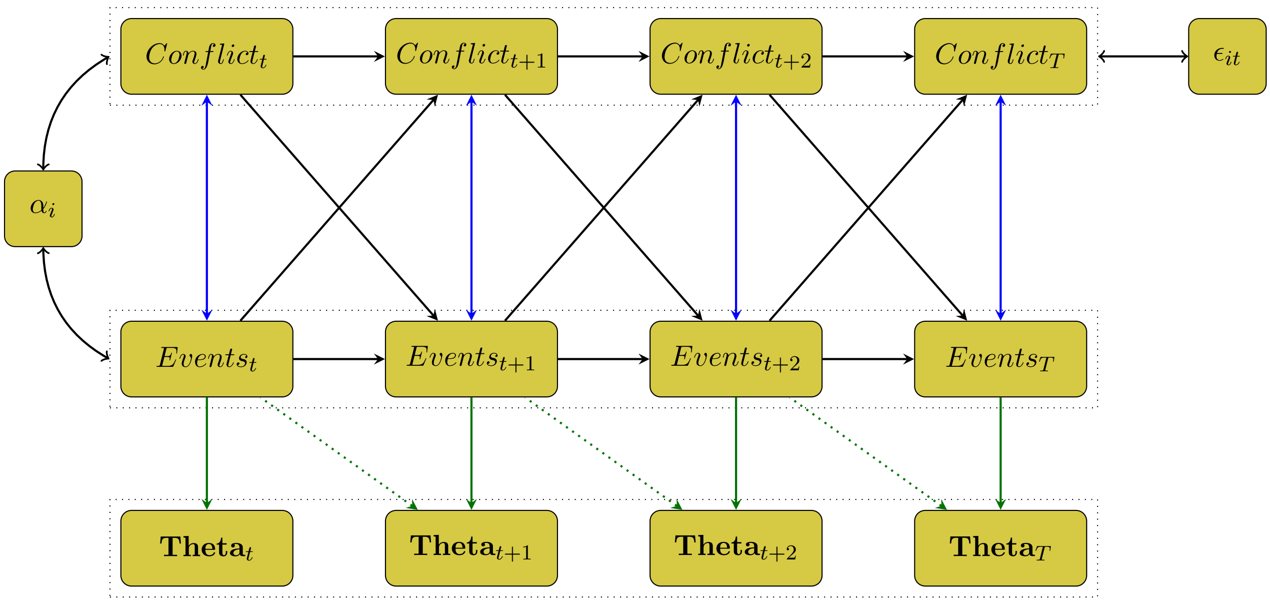
A Figures and Tables

Variable	Type	Mean	Std. Dev.	Min	Max	Observations
Armed Conflict	**overall**	0.142	0.349	0.000	1.000	7520
	between		0.020	0.106	0.186	40
	within		0.349	-0.044	1.036	188
Civil War	**overall**	0.060	0.237	0.000	1.000	7520
	between		0.024	0.027	0.112	40
	within		0.236	-0.052	1.033	188
Topic 1 Share	**overall**	0.053	0.039	0.007	0.560	6639
	between		0.005	0.046	0.063	39
	within		0.038	-0.002	0.561	185
Topic 2 Share	**overall**	0.073	0.041	0.010	0.559	6639
	between		0.010	0.050	0.089	39
	within		0.040	0.004	0.549	185
Topic 3 Share	**overall**	0.043	0.049	0.006	0.454	6639
	between		0.003	0.038	0.051	39
	within		0.049	0.004	0.451	185
Topic 4 Share	**overall**	0.060	0.068	0.009	0.663	6639
	between		0.012	0.032	0.080	39
	within		0.067	-0.006	0.663	185
Topic 5 Share	**overall**	0.069	0.045	0.004	0.468	6639
	between		0.008	0.045	0.081	39
	within		0.045	-0.003	0.476	185
Topic 6 Share	**overall**	0.063	0.052	0.009	0.765	6639
	between		0.011	0.036	0.081	39
	within		0.051	-0.003	0.774	185
Topic 7 Share	**overall**	0.074	0.047	0.007	0.514	6639
	between		0.006	0.063	0.086	39
	within		0.046	-0.005	0.509	185
Topic 8 Share	**overall**	0.070	0.052	0.007	0.426	6639
	between		0.006	0.058	0.084	39
	within		0.051	-0.006	0.420	185
Topic 9 Share	**overall**	0.074	0.054	0.010	0.514	6639
	between		0.012	0.058	0.116	39
	within		0.053	-0.024	0.519	185
Topic 10 Share	**overall**	0.065	0.051	0.007	0.612	6639
	between		0.008	0.053	0.092	39
	within		0.051	-0.009	0.605	185

Topic 11 Share	**overall**	0.063	0.046	0.005	0.407	6639
	between		0.010	0.047	0.082	39
	within		0.044	-0.008	0.410	185
Topic 12 Share	**overall**	0.075	0.069	0.004	0.653	6639
	between		0.017	0.058	0.135	39
	within		0.067	-0.044	0.654	185
Topic 13 Share	**overall**	0.089	0.090	0.008	0.623	6639
	between		0.010	0.070	0.103	39
	within		0.090	-0.001	0.614	185
Topic 14 Share	**overall**	0.067	0.048	0.007	0.582	6639
	between		0.005	0.058	0.076	39
	within		0.048	0.006	0.579	185
Topic 15 Share	**overall**	0.061	0.055	0.006	0.437	6639
	between		0.007	0.048	0.075	39
	within		0.055	-0.006	0.429	185

Table 1: Panel Data Summary

Figure 1: Path Diagram of Model Hypothesis



References

Besley, Timothy and Torsten Persson. 2011. Pillars of prosperity: The political economics of development clusters. Princeton University Press.

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

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.

Miguel, Edward and Shanker Satyanath. 2011. "Re-examining economic shocks and civil conflict." American Economic Journal: Applied Economics 3(4):228-232.

Mueller, H., & Rauh, C. (2018). "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text." American Political Science Review, 112(2), 358-375. doi:10.1017/S0003055417000570

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.