M2 EEE Panel Data

Panel Data Replication Project

Andrew Boomer,

Jacob Pichelmann,

Luca Poll

November 22, 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 key pillar of this analysis is the news data that is used to explain and predict conflict. The authors use an unsupervised learning algorithm to distill topic shares out of a set 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]. They start by processing the articles' contents by standard text mining techniques such as stemming words.[5] This leaves the authors with roughly 0.9 million tokens, which are then grouped into topics based on the latent Dirichlet allocation (LDA) method. A topic is then a probability distribution over words. The result is intuitive, as one can imagine that an article covering Sports is might indeed be more likely to contain words such as "score", "win", "match". 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.

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]174.450 articles from 1975 onward
[3]363.275 articles from 1980 onward
[4]185.523 articles from 1977 onward
[5]Stemming refers to the process of finding the common root of a word, i.e. ârunningâ, âranâ, and ârunâ all become ârunâ.

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 $\theta_{it} < -\theta_{it-1}$.
- ▷ 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.
  - 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[6] 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:

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

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

In Table 2 and Table 3 we provide a replication of the models used by the authors. They use a fixed effects model, and we show this compared to both a Pooled OLS model and a FE model where the topic shares are additionally interacted with an autocracy dummy. The interaction coefficients are omitted from the regression output.

We also replicated the ROC curve, comparing the false positive prediction rate to the true positive predictoin rate, in Figure 2. As the authors found in their research, the

[6] 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

predictive quality of the estimation drops when excluding the between variation from the prediction.

5 Improved model

We extend the authors' analysis by shifting the focus from a purely forecast driven evaluation to a more thorough understanding of the model. This change of perspective ultimately aims at developing a model that is both better at truly assessing the underlying relationship between conflict and events/topics (?) and better at enabling an intuitive interpretation of the resulting estimates. We start by easing some of the restrictions the authors placed on the data, yielding a more balanced and complete panel data set. We then continue with an assessment of the suitability of a set of panel data models, namely pooled OLS, fixed effects (the authors' model of choice) and a dynamic panel data model, each in conjunction with the most suitable estimation strategy for this specific setting. Lastly, we provide a thorough discussion of the possible sources of bias and outline mitigation strategies.

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

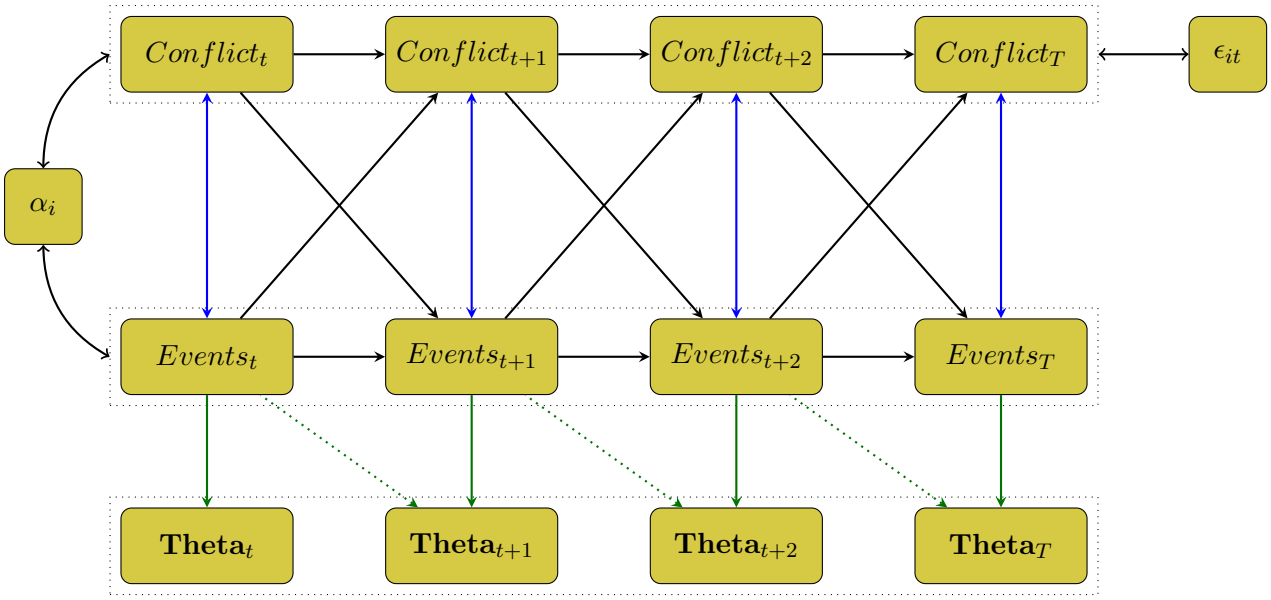

Table 2: Initial Panel Models: Armed Conflict

	Pooled	**FE**	**FEInteract**
	Armed Conflict		
Constant	-0.2615***	-0.2195***	-0.2193***
	(0.0609)	(0.0742)	(0.0795)
Topic 2 Share	0.2588***	0.2625**	0.2920***
	(0.0767)	(0.1038)	(0.1107)
Topic 3 Share	0.1760**	0.2651*	0.2395
	(0.0857)	(0.1419)	(0.1542)
Topic 4 Share	0.5330***	0.3668***	0.4174***
	(0.1304)	(0.1193)	(0.1224)
Topic 5 Share	0.1846***	0.2132**	0.1819*
	(0.0686)	(0.0841)	(0.1003)
Topic 6 Share	0.3911**	0.1447	0.2605*
	(0.1710)	(0.1233)	(0.1413)
Topic 7 Share	0.1748**	0.2595*	0.2531*
	(0.0680)	(0.1362)	(0.1459)
Topic 8 Share	0.3066***	0.2768***	0.2715***
	(0.0763)	(0.0883)	(0.0894)
Topic 9 Share	0.1820***	0.1710*	0.1878**
	(0.0641)	(0.0883)	(0.0891)
Topic 10 Share	0.2764***	0.2381**	0.2535**
	(0.0955)	(0.1146)	(0.1238)
Topic 11 Share	1.0280***	0.8449***	0.8301***
	(0.1670)	(0.1935)	(0.2011)
Topic 12 Share	0.2100***	0.2001**	0.1826*
	(0.0793)	(0.0925)	(0.1011)
Topic 13 Share	0.2398**	0.2287**	0.2037**
	(0.0966)	(0.0950)	(0.0985)
Topic 14 Share	0.3517***	0.2141**	0.1217
	(0.0991)	(0.1018)	(0.1186)
Topic 15 Share	0.2953***	0.2827***	0.2915**
	(0.0803)	(0.1091)	(0.1263)
Included Effects:	Time	Entity, Time	Entity, Time
R-Squared:	0.042	0.015	0.023
Observations:	4486	4486	4348

Cluster Robust Standard Errors

Table 3: Initial Panel Models: Civil War

	Pooled	**FE**	**FEInteract**
		Civil War	
Constant	-0.0674	-0.1259**	-0.1293**
	(0.0844)	(0.0596)	(0.0622)
Topic 2 Share	0.0641	0.1564*	0.1643**
	(0.0860)	(0.0804)	(0.0831)
Topic 3 Share	0.0750	0.2212**	0.2147**
	(0.1008)	(0.0913)	(0.0948)
Topic 4 Share	0.2844**	0.2748***	0.3124***
	(0.1109)	(0.0876)	(0.0995)
Topic 5 Share	0.0580	0.1866***	0.2019**
	(0.0875)	(0.0666)	(0.0787)
Topic 6 Share	0.0698	0.0017	0.0650
	(0.1343)	(0.1277)	(0.1146)
Topic 7 Share	0.0074	-0.0088	-0.0226
	(0.0891)	(0.0913)	(0.0899)
Topic 8 Share	0.0001	0.0671	0.0508
	(0.1003)	(0.0702)	(0.0694)
Topic 9 Share	0.0534	0.1182*	0.1336**
	(0.0898)	(0.0629)	(0.0659)
Topic 10 Share	0.0218	0.0919	0.0988
	(0.0990)	(0.0841)	(0.0870)
Topic 11 Share	0.4905***	0.5722***	0.5778***
	(0.1579)	(0.1574)	(0.1611)
Topic 12 Share	0.0533	0.1118	0.1106
	(0.0952)	(0.0742)	(0.0772)
Topic 13 Share	0.0352	0.1570**	0.1615**
	(0.0900)	(0.0730)	(0.0728)
Topic 14 Share	0.0712	0.0972	0.0856
	(0.1183)	(0.0762)	(0.0843)
Topic 15 Share	0.0732	0.1601**	0.1157
	(0.0814)	(0.0661)	(0.0745)
Included Effects:	Time	Entity, Time	Entity, Time
R-Squared:	0.036	0.020	0.027
Observations:	5062	5062	4924

Cluster Robust Standard Errors

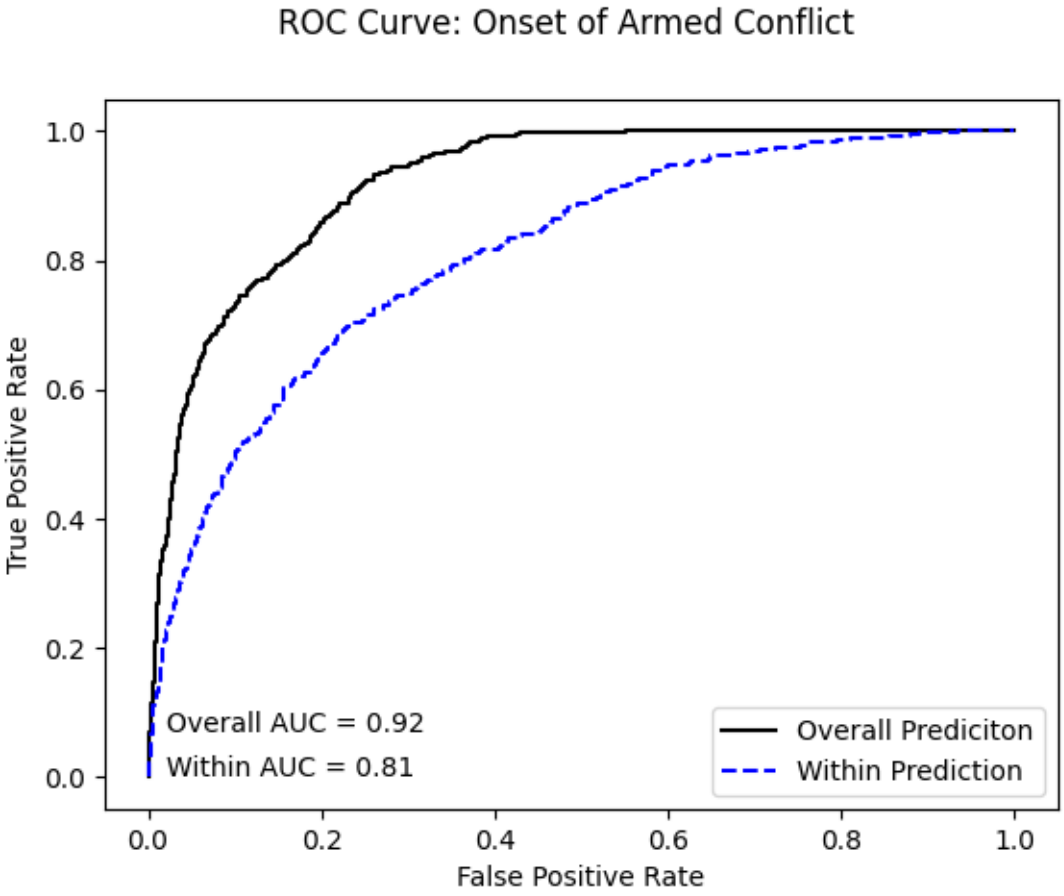

Figure 2: ROC Curve

Table 4: Blundell-Bond System Models

	GMM
	Armed Conflict
Lag of Topic 2 Share	0.0546
	(0.0405)
Lag of Topic 3 Share	0.1603
	(0.0981)
Lag of Topic 4 Share	-0.0218
	(0.0838)
Lag of Topic 5 Share	0.1205**
	(0.0555)
Lag of Topic 6 Share	0.3156***
	(0.0971)
Lag of Topic 7 Share	0.2127**
	(0.1084)
Lag of Topic 8 Share	-0.0522
	(0.0465)
Lag of Topic 9 Share	-0.1058***
	(0.0378)
Lag of Topic 10 Share	0.0705
	(0.0528)
Lag of Topic 11 Share	0.3721***
	(0.1006)
Lag of Topic 12 Share	0.0105
	(0.0559)
Lag of Topic 13 Share	0.0540
	(0.0746)
Lag of Topic 14 Share	-0.0588
	(0.0566)
Lag of Topic 15 Share	0.0306
	(0.0444)
Lag of Armed Conflict	0.5891***
	(0.0234)
Included Effects:	Time, Entity
R-Squared:	0.403
Observations:	8954
Cluster Robust Standard Errors	

	GMM
	Civil War
Lag of Topic 2 Share	0.0418
	(0.0268)
Lag of Topic 3 Share	-0.1096***
	(0.0395)
Lag of Topic 4 Share	0.0353
	(0.0362)
Lag of Topic 5 Share	0.0079
	(0.0250)
Lag of Topic 6 Share	0.0237
	(0.0509)
Lag of Topic 7 Share	-0.0342
	(0.0269)
Lag of Topic 8 Share	-0.0232
	(0.0278)
Lag of Topic 9 Share	-0.0312
	(0.0226)
Lag of Topic 10 Share	-0.0432
	(0.0265)
Lag of Topic 11 Share	0.3930***
	(0.0668)
Lag of Topic 12 Share	-0.0430
	(0.0307)
Lag of Topic 13 Share	-0.0522*
	(0.0298)
Lag of Topic 14 Share	-0.0334
	(0.0360)
Lag of Topic 15 Share	0.0142
	(0.0310)
Lag of Civil War	0.6720***
	(0.0141)
Included Effects:	Time, Entity
R-Squared:	0.197
Observations:	8954
Cluster Robust Standard Errors	

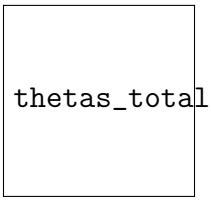

Figure 3: Topic Shares over Time

References

[!h] 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.