Final project Econometrics

Make Words, Not War! The impact of information dissemination on conflicts

Louise Guez, Joachim Schwartz

November 19, 2024

1 Research Question

There is a large literature in economics dealing with the aftereffects of natural resources: is it a bless or a curse? One strand of this literature focuses on the consequences of natural resources on conflicts. More precisely, some papers focus on the discovery of a new natural resource and its impacts on conflicts at the local scale. One of the main channel for this is the political behaviour of citizens and rents capture by political leader. Armand et al. (2020) investigate how the size of this impact changes when information dissemination about the discovery and its possibilities varies. They analyse a promising natural gas discovery in Mozambique and implement a randomized control trial (RCT) to observe how different types of information dissemination across communities modify political behaviour and lead to violent events or not.

This paper aims at verifying thanks to simulations the simple OLS model proposed by Armand et al. (2020) to assess their research question. Among their several analyses, we only focus on their measurement of the effect led by information dissemation on conflicts occurence. For our replication, we use panel data at the individual level from Armand et al.(2020)'s replication packages, and especially data dealing with their RCT. To measure local variation in violence, we use international event-based datasets following Armand et. al (2020): the Armed Conflict Location and Event Data Project (ACLED ; Raleigh et al., 2010) and the Global Database on Events, Location and Tone (GDELT ; Leetaru and Schrodt, 2013).

In Armand et al. (2020, the randomization process (RCT) was effective in creating comparable groups within the experiment. Two treatments were implemented : in the first group of randomly-selected communities, only local political leaders received the information module (*leader treatment*). In the second group, the information module was delivered to both local leaders and citizens, targeting communities at large while aiming to provide higher levels of accountability (*community treatment*). In the control group, there was no dissemination's effects organized. To conduct their analyzes, the author considered different outcomes variables define as Y_{ij} (for community j and individual i). In this case individual i can be a local leader or a citizen. Outcomes defined at the community level are handled in the same manner as those defined at the level of the local leader. In order to simplify our project, we have chosen to focus on a singular outcome Y_{ij}, specifically the one pertaining to violence. More precisely, we have decided to focus on the presence of violent events at the community level. In doing so, the empirical strategy deals with the outcomes related to georeferenced violent events from specific data sources (ACLED,GDELT). First, we treat them as dummy variables as in Armand et al. (2020): each dependent variable is an indicator variable, taking value 1 if a violent event was recorded within 5 km of the community.

This paper is structured as follows. First, we partially replicate the results from Armand et al. (2020) dealing with violent events. Then, we implement both a fake data and a real data simulation to assess the relevance of the model used by the authors. We also add an extension through the introduction of a categorical variable taking into account the distance of the violent events to the community.

2 Replication

2.1 Model

The model used in Armand et al. and that we replicate is the following one (OLS + lag):

$$Y_{ijt} = \beta_0 + \beta_1 T_{1j} + \beta_2 T_{2j} + \gamma Y_{ijt-1} + \delta X_{ij} + \varepsilon_{ij} \tag{1}$$

With the following variables:

- Y_{ijt} is the a dummy variable that takes 1 if violent event occur in year t within 5km to the community j where individual i lives.
- Y_{ijt-1} is a lag variable that takes 1 if violent event occur in year $t-1$ within 5km to the community j where individual i lives.
- T_{1j} is a dummy variable that takes 1 if community j received the leader treatment.
- T_{2j} is a dummy variable that takes 1 if community j received the community treatment.
- X_{ij} is a vector of leader and community controls.

2.2 Results

We partially retrieve in Table 2 the results from Table 1 in Armand et al. (2020). Importantly, we find that the community treatment is significant in reducing the presence of violent event within 5km to the community. However, our coefficients are different. This is likely due to the fact that we do not use all the controls they put in their regression. For sake of simplicity, we did not take into account the following controls which were compelx indexes calculated previously (and in another do file) by the authors: infrastructure, natural resources, main ethnicity of the community.

Table 1—Violence

	Presence of violent events		
	ACLED (1)	GDELT (2)	ACLED + GDELT (3)
(T1) Leader treatment	−0.025	−0.017	−0.047
	(0.031)	(0.028)	(0.035)
	[0.61–0.61]	[0.61–0.61]	[0.31-0.40]
(T2) Community treatment	−0.057	−0.054	−0.085
	(0.028)	(0.026)	(0.032)
	[0.08–0.16]	[0.08–0.16]	[0.03–0.05]
Observations	206	206	206
R^2	0.275	0.733	0.656
Mean (control group)	0.055	0.091	0.127
T1 = T2 (p-value)	0.245	0.145	0.223
T1 = T2 (adjusted p-value, row-level)	0.226	0.200	0.226
T1 = T2 (adjusted p-value, table-level)	0.458	0.376	0.458
Lagged dependent variable	Yes	Yes	Yes

Table 1: Results from Armand et al. (2020)

	ACLED	GDELT	ACLED + GDELT
Treatment 1	−0.034	−0.019	−0.050
	(0.033)	(0.029)	(0.037)
Treatment 2	−0.050	−0.039	−0.063*
	(0.032)	(0.025)	(0.033)
Num.Obs.	206	206	206
R2	0.200	0.696	0.610

* p < 0.1, ** p < 0.05, *** p < 0.01
Standard errors clustered at household level
All regressions include baseline controls

Table 2: Replication: Impact of information campaign on violence (5km radius)

3 Simulations

Our goal is to run various simulations in order to retrieve the same rough estimates for our effect of interest (violent event according to ACLED + GDELT). At first sight, the simple OLS model used by the auhors is surprising. Our simulations are a way to analyze its relevance. First, we generate a fake data simulation and try to retrieve the effect of interest for this outcome. Then, we use real data simulation to investigate further for potential pitfalls.

3.1 Fake data simulation

We use a simple data generating process designed in order to estimate the following model:

$$Y_{ijt} = \beta_0 + \beta_1 T_{1j} + \beta_2 T_{2j} + \gamma Y_{ijt-1} + \varepsilon_{ij} \tag{2}$$

It is quite the same than (1), but here we assume taht there is no other (control) variables that could potentially influence our outcome.

To generate fake data, we need to input some baseline parameters. You can refer to the quarto document for the precise derivation of each parameters. Here, you will find the justification for each parameters.

To generate our fake data according to the DGP above, we need the following parameters:

- $\beta_0 = -0.05$ according to our replication
- $\beta_1 = -0.05$, according to the Armand et al. (2020), and consistent with our replication
- $\beta_2 = -0.09$, according to the Armand et al. (2020), and consistent with our replication
- $\gamma = -0.4$, according to our replication
- $\sigma_\epsilon = 0.2$, according to the RMSE of our replication, which was (almost) effective to retrieve the true effect
- $p_{dev} = 0.03$, according to the ACLED + GDELT data, the proportion of observations where a violent event took place is 3%.
- $p_{treat_{t1}} = 0.25$, according to RCT data
- $p_{treat_{t2}} = 0.5$, according to RCT data

Because our replication was effective in retrieving the same rough estimates than in the paper bu with a simpler model, we use it to assess values for some of our baseline parameters, and especially the effects of interest ($beta_1$ and $beta_2$). Some of other parameters are derived thanks to the data exploration and calibration, which is available in the Quarto document. Because our generating process do not create a dummy outcome at first, we use a threshold to allocate observation to value 1 or 0. This threshold is 0.25 and allows us to have about 8% of the observations which have the outcome equals to 1, as in the real data.

The fake data set allows us to estimate our model and to see if we retrieve the same effects of interest when we run it multiple times. Figure 1 and Figure 2 show that we partially retrieve the effect of Armand et al (2020) for the leader treatment. But most of the estimates are far away from the hypothetic effect of β_2 (community treatment). It could highlight some weaknesses of the model, but it could also come from a potential dichotomy between the DGP and the estimated model, due to the way we generated the dummy outcome (cf. supra).

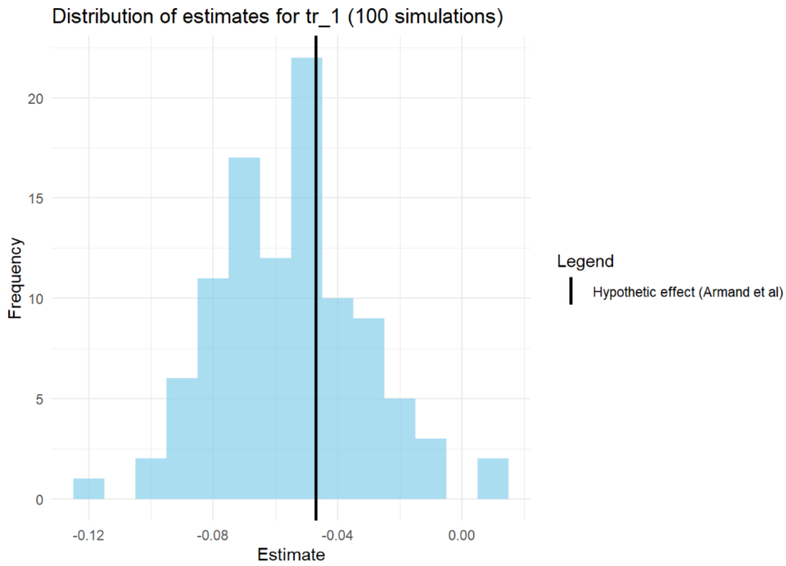

Figure 1: Leader treatment effect for 100 fake data simulations

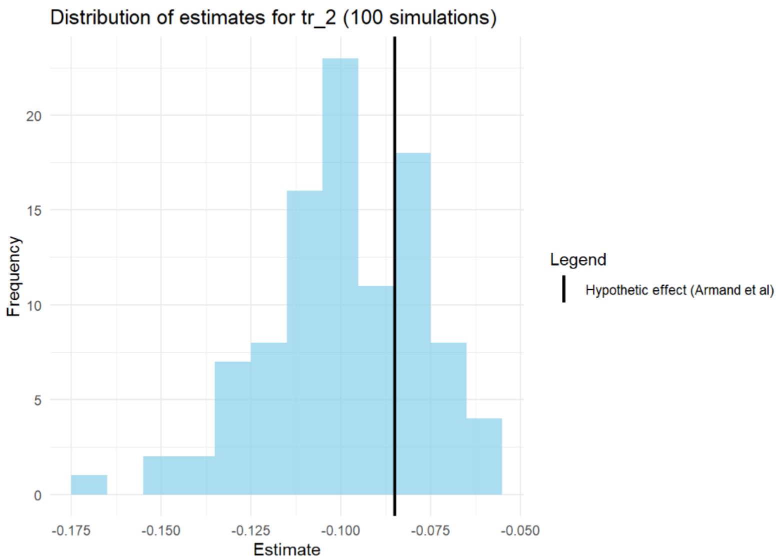

Figure 2: Community treatment effect for 100 fake data simulations

Then, simulating for various sample sizes, we provide an evolution of the potential statistical power depending on the sample size. Figure 3 shows poor results in terms of statistical power. It means that we hardly detect an existing effect, especially when the sample size is low. However, we may think that a huge sample size could make the model powerful enough. In Armand et al. (2020), they have about 2000 observations by year. Thus, we may think their model present a satisfactory statistical power.

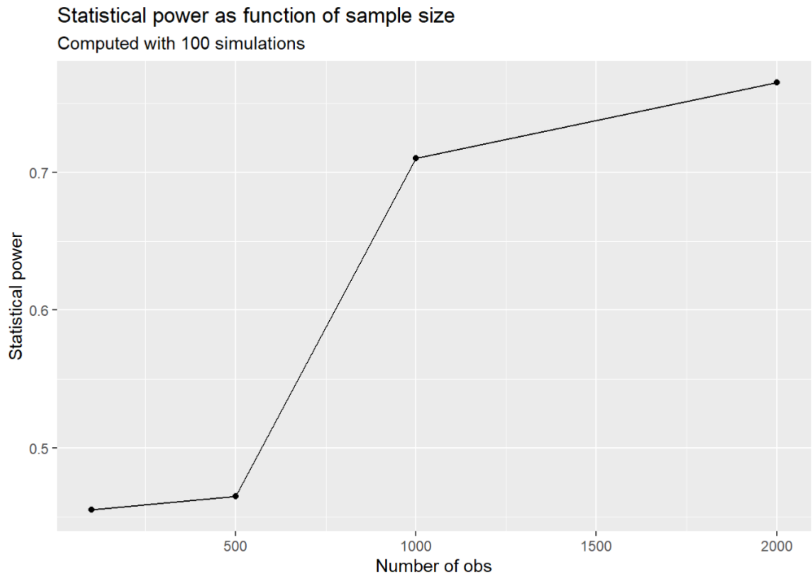

Figure 3: Statistical power and sample sizes

3.2 Real data simulation

The goal of this part is to check the robustness of the model and to investigate for some potential pitfalls.

3.2.1 Normal distribution of errors

Based on our replication, which is our best way to approximate the model of Armand et al. (2020), we investigate the normal distribution of errors. This could be determinant in so far as they use a simple OLS model, but with a binary outcome.

We use the Shapiro-Wilk test, which is often used for small samples.

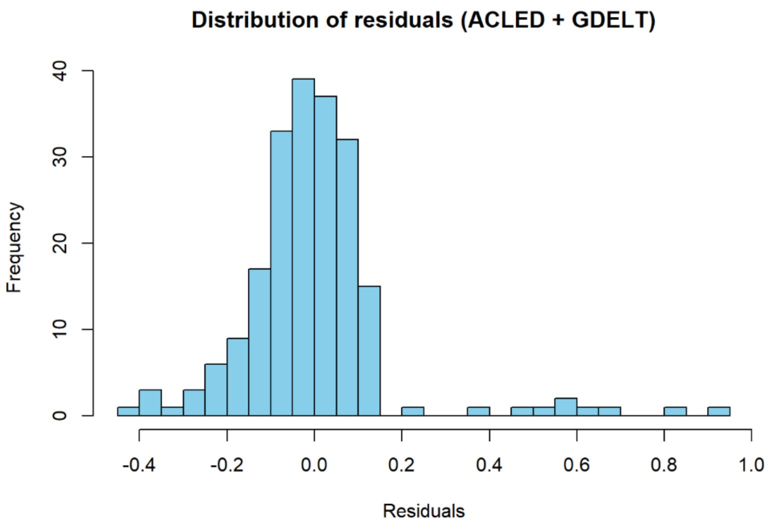

Figure 4: Normal distribution of residuals

Overall, we observe that the distribution is relatively symmetric and centered around 0, suggesting that the model is generally unbiased. The few extreme values on the right might indicate outliers. In summary, the assumption of normality of residuals is approximately met.

3.2.2 Bootstrap

Now, we simulate the model used for the replication (simple model + alsmost all the controls) on several samples randomly generated from the real data of Armand et al. (2020). It allows us to compare the estimates we find in the replication and the ones found by Armand et al. (2020) with the ones provided by various simulations on real data.

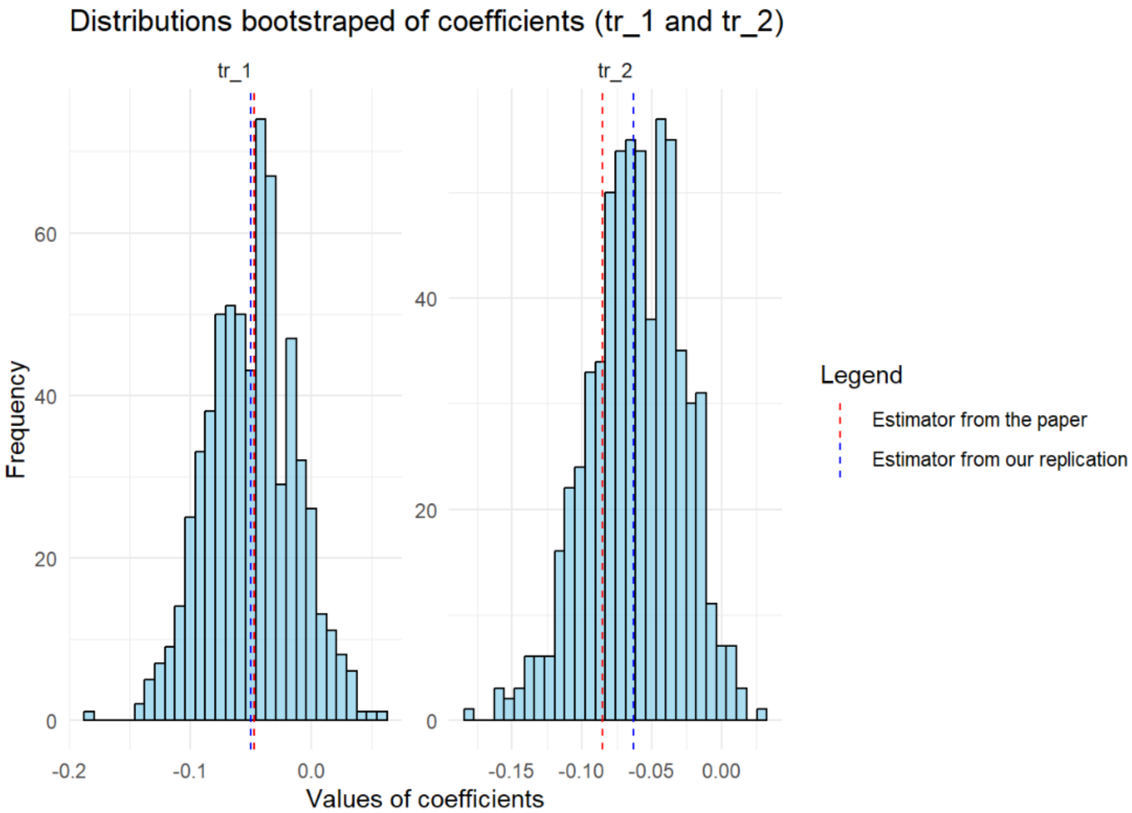

Figure 5: Bootstrap

We observe the empirical distribution of the coefficients by simulating multiple samples for the variables tr1 and tr2 on the outcome devc5km. These distributions appear to be mostly centered around both the theoretical and empirical effects. The estimate for Treatment 1 (tr1) is relatively well aligned with the replication results, closely approximating the original effect. However, this alignment is weaker for the estimate of Treatment 2 (tr2), where the replicated estimate deviates more noticeably from the theoretical effect. Overall, we find that our replication captures the effect of Treatment 1 more accurately than that of Treatment 2, indicating greater difficulty in replicating the original effect of Treatment 2.

3.2.3 Extension with a categorical variable

In order to add a categorical variable in our analysis, we modify the main outcome to take into account the distance of violent events to the community. It allows to understand better how treatments push violent events away from the community. The outcome "presence of violent events according to ACLED + GDLET" is now a categorical variable and takes four possible values:

- "*close*" when a violent event occurred within 3km to the community
- "*med*" when a violent event occurred between 3 and 5km away from the community
- "*away*" when a violent event occurred between 5 and 7km away from the community
- "*none*" when no violent event was recorded within 7km.

We investigate if being treated decrease the probability to have a violent event close to community. To handle this new categorical outcome, we use an ordered logit model. The goal is to see if we retrieve quite the same estimates for our effects of interest than with the model proposed by Armand et al. (2020) and which we replicated. This extension allows a more precise investigation of the treatment effects on violent events, but it is based on another model, which is justified by the new shape of our outcome. If we achieve to retrieve the same rough estimates, it would again demonstrate the robustness of Armand et al. (2020)'s simple OLS model.

Precisely, we use two ordered logit models with two different orders for our categorical outcome. The first one is: "*close*" < "*med*" < "*away*" < "*none*", whereas the second goes from "*none*" to "*close*". It allows us to disentangle between the effects of the treatments in increasing the probability to have a violent event far from its village from the effect of the treatments in reducing the probability to have a violent event close to its community.

What we find in Table 3 is not really consistent with previous results. Our extension does not allow to retrieve significant effects of the treatments in reducing the odds of facing a violent event close to its home (*resp*: increasing the odds of pushing away the violent event).

	Dependent variable:	
	Model 1	
	(1)	(2)
Treatment 1	0.849	−0.849
	(0.787)	(0.787)
Treatment 2	0.365	−0.365
	(0.579)	(0.579)
Observations	206	206
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table 3: Results of ordered logit models

Conclusion

In conclusion, we partially recovered the estimated coefficients for treatments 1 and 2 from Armand et al. (2020). However, we couldn't exactly recover the same effects due to missing controls. In our subsequent work, we focused on a single type of outcome, namely the occurrence of a violent event recorded by ACLED and GDELT databases more than 5km from a community. Regarding the simulation, we generally recovered the effect related to treatment 1 more than treatment 2, which isn't really consistent. Indeed, the paper predicts a greater effect for treatment 2 (namely the fact that leaders and citizens are aware of the information and could deliberate about it should theoretically reduce the number of violent events observed within 5km of the community). The result of our simulation is therefore not very consistent with the result found by Armand et al. (2020), namely that when information about natural resources is only communicated to leaders, the number of violent events found near a community decreases, which is not the result found in the paper. In this regard, our statistical power is quite low (0.7) for the theoretical number of observations taken by the authors (namely 2000). Finally, we conducted two types of tests to ensure the robustness of results by verifying first the normality of errors, allowing us to conclude that the model was able to estimate the outcome in question (a dummy) by OLS. And secondly, a bootstrap test confirms that our replication better captures the effects of Treatment 1 than 2. Finally, we considered a model extension by creating a categorical variable still concerning the same outcome by dividing it into 4 categories: "close" when a violent event occurred within 3km of the community, "med" when a violent event occurred between 3 and 5km away from the community, "away" when a violent event occurred between 5 and 7km away from the community, "none" when no violent event was recorded within 7km. However, the generated model doesn't allow us to find significant effects.

Reference

Armand, A., Coutts, A., Vicente, P. C., & Vilela, I. (2020). "Does information break the political resource curse? Experimental evidence from Mozambique". *American Economic Review*, 110(11), 3431-3453.