Gov 2001: Replication of Original Paper

Sima Biondi, Priyanka Sethy, Natalie Ayers

2022-11-05

Introduction

We are replicating "Redemption through Rebellion: Border Change, Lost Unity, and Nationalist Conflict." (2022), by Lars-Erik Cederman, Seraina Rüegger, and Guy Schvitz. The data and code for this paper are in the Harvard Dataverse. We are proposing to extend this paper in three ways: first, by including a variable of the conflict instigator to identify whether the paper's proposed mechanism, which relys on the rebels instigating conflict, is correct or if there may be other, state-initiated mechanisms at work. We also play to include a measure of topological features of the ethnic group's terrain to account for previous research findings of terrain's importance in civil conflict prediction (e.g., Carter et al. 2019, Buhaug and Gates 2002; Buhaug, Gates, and Lujala 2009; Fearon and Laitin 2003). Finally, we hope to test the validity of an additional mechanism of state consolidation by expanding Cederman et al.'s list of conflicts to include intrastate and extrastate/interstate conflicts which occur between a state and ethnic groups outside of the state boundaries.

The main findings of the Cederman et al. paper which will be important for our extensions are from their logistic regression models, which they run to test the impact of aggregate group factors on the onset of conflict. These results are presented for presented for post-WWII in Table 1, for post-1886 in Table 2, and as a predicted probability model in Figure 7. They also develop fixed-effects models in their robustness checks to reduce omitted variable bias and act as an additional test for changes in fractionalization over time, and they present these results in Table 3. We replicate these 4 tables and figures in our analysis, as we plan to use the same models in our extensions to provide for a direct comparison of the results.

We have so far noticed only one discrepancy between the code and data the authors provided and their published results: in Table 3, the authors find estimated coefficients of 0.058 (se 0.025) for Fractionalization and 0.067 (se 0.026) for Fractionalization increase since 1946, while our replication of their results using both their original stata code and the re-factored R code provide values of 0.046 (se 0.021) and 0.055 (se 0.024) for Fractionalization and Fractionalization increase since 1946, respectively.

Results

```
# Set working directory to own project folder or use here function
#setwd("./redemption-through-rebellion-dataverse_files/")

## Load dataset
an.df <- read.dta(here("redemption-through-rebellion-dataverse_files", "epr_segment_level_analysis.dta")

## Code "peaceyears" variable: years since last conflict, squared, cubed
an.df$pys <- an.df$peaceyears
an.df$pys2 <- an.df$peaceyears^2
```

```
an.df$pys3 <- an.df$peaceyears^3

## Subset dataset: only politically relevant groups (EPR definition), exclude
## dominant and monopoly groups and groups without settlement area in GeoEPR,
an.df.sub.allyears <- an.df %>%
  filter(isrelevant==1,
         status_monopoly==0,
         status_dominant==0,
         !is.na(seg_area_sqkm),
         !is.na(onset_do_flag))
```

Table 1

The table below replicates Table 1 in Cederman et al. (2022), providing the results of 3 logit models with robust standard errors clustered at the AG level. Our results align exactly with the published results, which the authors originally obtained from Stata using `logit` with `cluster`. They show that having an AG divided as a binary measure has a significant, positive relationship with conflict onset (1), while testing the impact of a continuous measure of territorial fractionalization and increased fractionalization since 1946 both show even larger, positive relationships with conflict onset. Given the intent of the authors to identify whether fractionalization is relevant to conflict onset, the logit model they apply seems relevant here.

```
# define variables for use in all models, including DV
vars_logit <- c("onset_do_flag", "ln_ag_area_sqkm", "ag_incidence_flag_lag",
                "status_excl", "downgraded2",
                "rbal", "warhist", "ln_capdist", "ln_rgdppc_lag",
                "ln_pop_lag",  "colonial_past", "ln_state_age",
                "pys", "pys2", "pys3")
# identify AG "treatment" variables to be analyzed separately
treat.vars <- c("split", "tfrac", "tfrac_incr_post1946")

# for each "treatment" variable, run logit model
for(treat_var in treat.vars){
  full_lhs <- append(treat_var, vars_logit)
  onset_logit <- glm(onset_do_flag ~ .,
                     data=an.df.sub.allyears[,full_lhs], family="binomial")
  assign(paste0("logit_",treat_var), onset_logit)
}

# include clustered standard errors for each generated model
logit_split_coefs <- coeftest(logit_split,
                              vcov. = vcovCL(logit_split,
                                             cluster = an.df.sub.allyears$ag_id))
logit_tfrac_coefs <- coeftest(logit_tfrac,
                              vcov. = vcovCL(logit_tfrac,
                                             cluster = an.df.sub.allyears$ag_id))
logit_tfrac_1946_coefs <- coeftest(logit_tfrac_incr_post1946,
                                   vcov. = vcovCL(logit_tfrac_incr_post1946,
                                                  cluster = an.df.sub.allyears$ag_id))

# Display results of logit model
stargazer(logit_split_coefs, logit_tfrac_coefs, logit_tfrac_1946_coefs,
          type="latex",
```

```
          dep.var.labels.include = FALSE,
          omit = c("pys", "pys2","pys3"),
          title = "Replication of Table 1",
          dep.var.caption = "Civil Conflict Onset",
          covariate.labels = c("Divided group", "Fractionalization",
                               "Frac. incr. since 1946", "Territory sq.km, log",
                               "Ongoing conflict, lag", "Exclusion",
                               "Downgraded", "Relative size", "War history",
                               "Distance to capital, log", "GDP, lag, log",
                               "Population, log","Colonial history",
                               "State age, log"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Dec 09, 2022 - 15:41:26

Table 2

The below table replicates Table 2 in Cederman et al. (2022), providing the results of logit models testing the impact of increasing fractionalization over three historical periods on conflict onset. This analysis is a direct extension of their Table 1 results, aiming to trace trends in fractionalization's impact on conflict onset through history, particularly identifying the impact of border changes around the WWII cut-off. These results also support their theory, with post-1946 fractionalization producing a stronger impact than pre-1946, but both significant and positive. Our results generated in R match exactly their published results generated in Stata.

However, this analysis relies on applying the GeoEPR ethnic settlement data from 1946 to all years from 1886-1946, which the authors acknowledge may be suspect. To alleviate concerns, they geocode a 1918 map of ethnic groups in Europe to re-run the tests for Europe, and geocode data on precolonial ethnic groups in Africa to re-run the test for African settlements. Both of these robustness checks still find significant, positive effects of historical fractionalization on conflict onset, but given the difficulties and likely inaccuracies involved with geocoding qualitative maps, the African data being only a "rough approximation" (pg 38), and the European data not reflecting anything pre-1918, there can still be doubts remaining as to the authors' ability to make these historical claims. As such, while we will replicate this analysis with our extensions, we give more weight to the Table 1 results.

```
# define "treatment" variables
treat.vars.2 <- c("tfrac_incr", "tfrac_incr_pre1946")

# Create models 2.1 and 2.3
for(treat_var in treat.vars.2){
  full_lhs <- append(treat_var, vars_logit)
  onset_logit <- glm(onset_do_flag ~ .,
                     data=an.df.sub.allyears[,full_lhs], family="binomial")
  assign(paste0("logit_2_",treat_var), onset_logit)
}

# Create model 2.2
combined_lhs <- append(c("tfrac_incr_pre1946","tfrac_incr_post1946"), vars_logit)
logit_2_prepost1946 <- glm(onset_do_flag ~ .,
                       data = an.df.sub.allyears[,combined_lhs], family="binomial")
# generate RSEs for all three models
logit_2_tfrac_coefs <- coeftest(logit_2_tfrac_incr,
                                vcov. = vcovCL(logit_2_tfrac_incr,
```

Table 1: Replication of Table 1

	Civil Conflict Onset		
	(1)	(2)	(3)
Divided group	0.581***		
	(0.182)		
Fractionalization		1.559***	
		(0.353)	
Frac. incr. since 1946			2.576***
			(0.752)
Territory sq.km, log	−0.051	−0.064	0.027
	(0.053)	(0.047)	(0.045)
Ongoing conflict, lag	0.100	−0.038	0.122
	(0.421)	(0.377)	(0.442)
Exclusion	0.870***	0.845***	0.872***
	(0.287)	(0.270)	(0.283)
Downgraded	1.248***	1.295***	1.300***
	(0.271)	(0.265)	(0.271)
Relative size	1.042**	0.889**	0.877**
	(0.440)	(0.425)	(0.389)
War history	0.686***	0.678***	0.694***
	(0.109)	(0.108)	(0.098)
Distance to capital, log	0.142	0.142	0.139
	(0.087)	(0.093)	(0.087)
GDP, lag, log	0.218**	0.254**	0.115
	(0.110)	(0.109)	(0.109)
Population, log	−0.722***	−0.723***	−0.727***
	(0.187)	(0.178)	(0.182)
Colonial history	0.501	0.454	0.545**
	(0.315)	(0.304)	(0.270)
State age, log	0.098	0.076	0.163
	(0.131)	(0.128)	(0.116)
Constant	−3.370***	−3.255***	−3.747***
	(1.091)	(1.023)	(1.046)

Note: *p<0.1; **p<0.05; ***p<0.01

```
                                            cluster = an.df.sub.allyears$ag_id))
logit_2_tfrac_pre1946_coefs <- coeftest(logit_2_tfrac_incr_pre1946,
                                  vcov. = vcovCL(logit_2_tfrac_incr_pre1946,
                                                 cluster = an.df.sub.allyears$ag_id))
logit_2_prepost1946_coefs <- coeftest(logit_2_prepost1946,
                                vcov. = vcovCL(logit_2_prepost1946,
                                               cluster = an.df.sub.allyears$ag_id))


# display results
stargazer(logit_2_tfrac_coefs, logit_2_prepost1946_coefs, logit_2_tfrac_pre1946_coefs,
          type="latex",
          dep.var.labels.include = FALSE,
          omit = c("pys", "pys2","pys3"),
          title = "Replication of Table 2",
          dep.var.caption = "Civil Conflict Onset",
          covariate.labels = c("Fractionalization incr. since 1886",
                               "Frac. incr. before 1946",
                               "Frac. incr. since 1946",
                               "Territory sq.km, log",
                               "Ongoing conflict, lag", "Exclusion",
                               "Downgraded", "Relative size", "War history",
                               "Distance to capital, log", "GDP, lag, log",
                               "Population, log","Colonial history",
                               "State age, log"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Dec 09, 2022 - 15:41:27

Table 3

The table below replicates Table 3 in Cederman et al. (2022), which displays the results of two fixed effects models to account for potential AG omitted effects and, they argue, provide a better test of their hypothesis that stronger fractionalization should lead to greater chance of conflict.

Our results from an exact re-running of their provided Stata code with their data match to the R code below, giving us confidence in these results with the provided data; however, the results do not match the published results in the paper: they are off only slightly, but the impacts reported in the paper are slightly higher for both Fractionalization and Fractionalization increase since 1946 (0.058 (se 0.025) and 0.067 (se 0.026) as compared to our results of 0.046 (se 0.021) and 0.055 (se 0.024)). We are not certain why this is the case, but imagine perhaps the data published in the Dataverse was not the exact same as that used to generate the publication - it would be a miniscule difference, though, to generate these differences. It may also be the case that the authors used different data for only this table, or perhaps re-ran all of their other analyses with updated data and did not update this one.

That said, they (and we) find significant, positive impacts of fractionalization on conflict onset using this linear fixed effects model, and it does match their goal of testing for any omitted effects of AG characteristics. One potential concern with using it as evidence that within-AG changes over time have an impact, though, is the scarcity of conflict data points for each individual AG, prompting caution with any interpretations.

```
# identify "treatment" variables
treat.vars.3 <- c("tfrac", "tfrac_incr_post1946")
```

Table 2: Replication of Table 2

	Civil Conflict Onset		
	(1)	(2)	(3)
Fractionalization incr. since 1886	1.449***		
	(0.372)		
Frac. incr. before 1946		1.039***	0.918**
		(0.371)	(0.388)
Frac. incr. since 1946		2.749***	
		(0.764)	
Territory sq.km, log	−0.009	0.003	0.005
	(0.049)	(0.048)	(0.046)
Ongoing conflict, lag	0.073	0.037	0.120
	(0.432)	(0.448)	(0.434)
Exclusion	0.886***	0.888***	0.859***
	(0.297)	(0.290)	(0.294)
Downgraded	1.259***	1.288***	1.253***
	(0.268)	(0.270)	(0.269)
Relative size	0.894**	0.867**	0.937**
	(0.419)	(0.380)	(0.459)
War history	0.682***	0.677***	0.700***
	(0.102)	(0.096)	(0.110)
Distance to capital, log	0.113	0.133	0.113
	(0.091)	(0.092)	(0.088)
GDP, lag, log	0.203*	0.165	0.187*
	(0.107)	(0.110)	(0.108)
Population, log	−0.722***	−0.717***	−0.736***
	(0.173)	(0.172)	(0.180)
Colonial history	0.409	0.520*	0.369
	(0.297)	(0.267)	(0.324)
State age, log	0.104	0.154	0.080
	(0.125)	(0.114)	(0.134)
Constant	−3.532***	−3.786***	−3.387***
	(1.014)	(1.015)	(1.061)

Note: *p<0.1; **p<0.05; ***p<0.01

```r
# run fixed effects models for both treatment variables
# felm is very picky with equation inputs, so had to manually
# paste each equation into each regression
for(treat_var in treat.vars.3){
  full_lhs <- append(treat_var, vars_logit)
  full_lhs <- append(full_lhs,"ag_id")

  # use to create formula again if need to, copy and paste
  #felm_formula <- as.formula(paste("onset_do_flag ~ ",
  #                           paste(full_lhs[!full_lhs %in% c("onset_do_flag","ag_id")],
  #                           collapse="+")))

  d <- an.df.sub.allyears[,full_lhs]
  if(treat_var == "tfrac"){
    # regression for `tfrac` variable
    onset_tfrac_fe_reg <- felm(onset_do_flag ~ tfrac + ln_ag_area_sqkm + ag_incidence_flag_lag +
                                 status_excl + downgraded2 + rbal + warhist + ln_capdist +
                                 ln_rgdppc_lag + ln_pop_lag + colonial_past + ln_state_age +
                                 pys + pys2 + pys3 | ag_id | 0 | ag_id, data = d)
  }
  else{
    # regression for `tfrac_incr_post1946` variable
    onset_tfrac_post1946_fe_reg <- felm(onset_do_flag ~ tfrac_incr_post1946 + ln_ag_area_sqkm +
                                          ag_incidence_flag_lag + status_excl + downgraded2 +
                                          rbal + warhist + ln_capdist + ln_rgdppc_lag +
                                          ln_pop_lag + colonial_past + ln_state_age +
                                          pys + pys2 + pys3 | ag_id | 0 | ag_id, data = d)
  }
}

# display results
stargazer(onset_tfrac_fe_reg, onset_tfrac_post1946_fe_reg,
          type="latex",
          dep.var.labels.include = FALSE,
          omit = c("pys", "pys2","pys3"),
          title = "Replication of Table 3",
          dep.var.caption = "Linear Probability Models: Civil Conflict Onset",
          covariate.labels = c("Fractionalization",
                               "Frac. incr. since 1946",
                               "Territory sq.km, log",
                               "Ongoing conflict, lag", "Exclusion",
                               "Downgraded", "Relative size", "War history",
                               "Distance to capital, log", "GDP, lag, log",
                               "Population, log","Colonial history",
                               "State age, log"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Dec 09, 2022 - 15:41:27

Figure 7

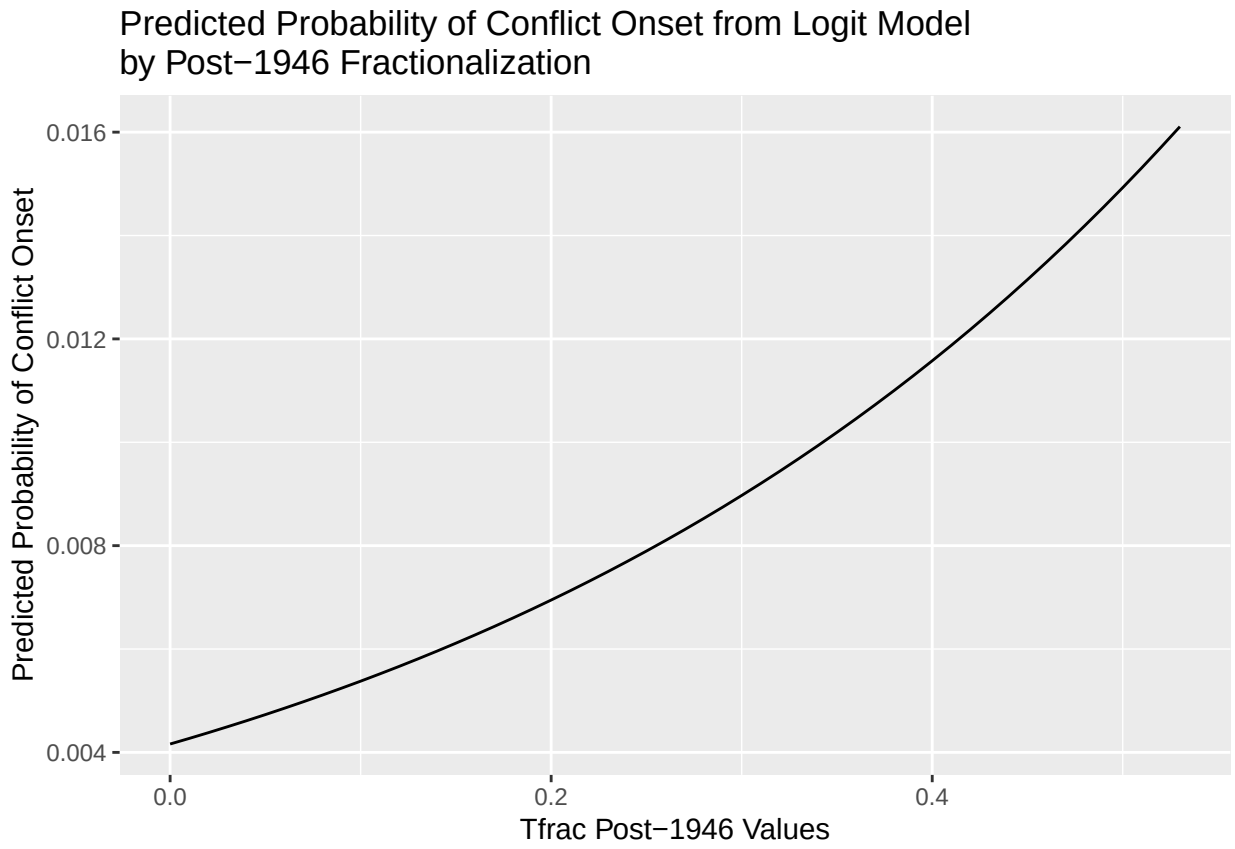
The figure below replicates Figure 7 in Cederman et al. (2022), which provides predicted probabilities of conflict onset as the post-1946 fractionalization increases from 0 to its maximum value, according to

Table 3: Replication of Table 3

	Linear Probability Models: Civil Conflict Onset	
	(1)	(2)
Fractionalization	0.046**	
	(0.021)	
Frac. incr. since 1946		0.055**
		(0.024)
Territory sq.km, log	−0.010	−0.009
	(0.010)	(0.010)
Ongoing conflict, lag	−0.015**	−0.015**
	(0.006)	(0.006)
Exclusion	0.004*	0.004
	(0.003)	(0.003)
Downgraded	0.020***	0.020***
	(0.007)	(0.007)
Relative size	0.012	0.012
	(0.007)	(0.007)
War history	0.011	0.011
	(0.007)	(0.007)
Distance to capital, log	0.001	0.001
	(0.001)	(0.001)
GDP, lag, log	0.001	0.001
	(0.002)	(0.002)
Population, log	−0.008***	−0.008***
	(0.003)	(0.003)
Colonial history	0.008	0.008
	(0.005)	(0.005)
State age, log	0.002	0.003
	(0.002)	(0.002)
Observations	27,105	27,105
R^2	0.052	0.052
Adjusted R^2	0.037	0.037
Residual Std. Error (df = 26677)	0.087	0.087
Note:		*p<0.1; **p<0.05; ***p<0.01

the original logit model developed for use in Table 1. The authors use it to indicate that their results are substantively important, demonstrating the magnitude of the predicted increase in conflict onset with increasing territorial fractionalization. The results of our R code match closely with the figure generated by the authors in Stata. Given the difficulties with making precise estimates from empirical conflict models, we are treating this figure as fairly supplemental, rather than central, to our extension - it provides a visual that is more easily interpretable, but the actual predicted values should be taken with caution, as there are wide confidence intervals around these predictions.

```
pred_prob_lhs <- append("tfrac_incr_post1946", vars_logit)
pred_prob_post1946_onset_logit <- glm(onset_do_flag ~ .,
                   data=an.df.sub.allyears[,pred_prob_lhs], family="binomial")

predprob_logit_tfrac_post1946_coefs <- coeftest(pred_prob_post1946_onset_logit,
                                vcov. = vcovCL(pred_prob_post1946_onset_logit,
                                               cluster = an.df.sub.allyears$ag_id))

# holding everything except tfrac_incr_post1946 at their means
mean_var_vals <- c(1, map(an.df.sub.allyears[,pred_prob_lhs[3:16]], function(x) mean(x, na.rm=TRUE)))
names(mean_var_vals)[1] <- c("(Intercept)")
mean_var_vals <- unlist(mean_var_vals)

# get max val of tfrac_incr_post1946 to properly specify range:
max_tfrac_post1946 <- max(an.df.sub.allyears$tfrac_incr_post1946, na.rm=TRUE)

# specify range of tfrac_incr_post1946 over which to predict
tfrac_post1946_range <- seq(0,max_tfrac_post1946, 0.01)

pred_prob_onset <- function(dyn_coef){
  full_coefs <- as.list(c(mean_var_vals[1],tfrac_incr_post1946=dyn_coef,mean_var_vals[2:15]))
  return(predict(pred_prob_post1946_onset_logit, full_coefs, type = "response"))
}

pred_probs_onset_tfrac_post1946 <- lapply(tfrac_post1946_range, pred_prob_onset)

pred_probs_onset_tfrac_post1946_df <- data.frame(pred_probs=unlist(pred_probs_onset_tfrac_post1946))
pred_probs_onset_tfrac_post1946_df$tfrac_incr_post1946 <- tfrac_post1946_range

# without Confidence Intervals
ggplot(data = pred_probs_onset_tfrac_post1946_df,
       aes(x = tfrac_post1946_range, y=pred_probs)) +
  geom_line() +
  xlab("Tfrac Post-1946 Values") +
  ylab("Predicted Probability of Conflict Onset") +
  labs(title = "Predicted Probability of Conflict Onset from Logit Model \nby Post-1946 Fractionalizati
```



Status of Extension Analysis

Instigator

The inclusion of an instigator field is not possible using the UCDP conflict data which Cederman et al. employ in their analysis. Therefore, we plan on joining the UCDP data with the Correlates of War Intra-State wars dataset (and potentially other datasets with instigator values, using xSub) to add this instigator field. We expect to lose a number of data points by performing this join, both because some data points won't be in the COW dataset and because some will be unable to join with confidence, but we expect there to be no systemic bias in the 2nd cause, and we expect any systemic bias from the first cause to have minimal to no impact on the mechanism of fractionalization's impact on conflict onset which we will be trying to test.

We plan on re-running the data creation with this updated dataset and re-running the same models above with an indicator variable for which side began the conflict and separating the dataset into two based on which side instigated. We expect the dataset creation to be the most time-intensive component of this part of the extension, given the code to run the models is already created, but we don't expect this to be unfeasible.

Topographic Features

We hope to use two datasets to test the impact of terrain features in our analysis. The first is the measure of terrain ruggedness published in "Terrain ruggedness and land cover: Improved data for most research designs" by Andrew Shaver, David B. Carter, and Tsering Wangyal Shawa (2019). This data is available at the 1 x 1 km level, so we can join it into our existing data with spatial joins to include in our analysis. While this measure is relatively new, it has already made an extensive impact on the literature, including with use

in other conflict research (eg, Carter, Shaver, and Wright, 2019), so it will allow us to compare and validate our results. We've downloaded the ruggedness data from the Harvard dataverse.

We'll also use the PRIO-GRID data to include the average mountainous terrain in our analysis, which we've downloaded from the PRIO-GRID site. We expect this to be less precise and informative than the ruggedness measure above, but given the extensive historical use of 'mountainous terrain' as a variable in conflict research, we feel it will provide for interesting comparison with historical analyses (eg, Fearon and Laitin, 2003).

Extrastate Conflict Inclusion

We'll be able to use the original UCDP dataset with revised filtering to test this mechanism, so there isn't new data to gather, only slight modifications to the filters used in generating the dataset.

References

Cederman, Lars-Erik, Seraina Rüegger, and Guy Schvitz. 2022. "Redemption through Rebellion: Border Change, Lost Unity, and Nationalist Conflict." American Journal of Political Science 66 (1): 24–42. https://doi.org/10.1111/ajps.12634.

Carter, David B., Andrew C. Shaver, and Austin L. Wright. 2019. "Places to Hide: Terrain, Ethnicity, and Civil Conflict." The Journal of Politics, September. https://doi.org/10.1086/704597.

Fearon, James D., and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." American Political Science Review 97 (1): 75–90. https://doi.org/10.1017/S0003055403000534.

Shaver, A., Carter, D. B., & Shawa, T. W. (2019). Terrain ruggedness and land cover: Improved data for most research designs. Conflict Management and Peace Science, 36(2), 191–218. https://doi-org.ezp-prod1.hul.harvard.edu/10.1177/0738894216659843

Sources Used

https://stackoverflow.com/questions/16498849/logistic-regression-with-robust-clustered-standard-errors-in-r

https://www.r-bloggers.com/2021/05/clustered-standard-errors-with-r/

https://statisticsglobe.com/name-variables-in-for-loop-dynamically-in-r

https://unc-libraries-data.github.io/R-Open-Labs/Extras/Stargazer/Stargazer.html