Predicting The Scale of Deadly Political Violence In Sub-Saharan Africa And The Maghreb

Jasper Linke

June 15, 2019

Introduction

This project uses a least square error machine leaning algorithm to predict the level of deadly political violence in Sub-Saharan Africa and the Maghreb between January 1997 and June 2019. The data used for this exercise is the African event data set of the Armed Conflict Location and Event Data Project (ACLED) published by the Peace Research Institute Oslo in June 2019. The data set was retrieved from the ACLED website on June 15 and can be found in this Git hub repository: https://github.com/jprlink/edx.git

This brief report will inform about the main steps undertaken in the project: 1) data cleaning and filtering, 2) data exploration, 3) preparation of train and test tests, 4) model training and selection, 5) regularization, 6) test of best-performing model, and 7) discussion of findings.

Methods and analysis

The following packages were loaded for the project:

```
library(tidyverse)
library(caret)
library(readxl)
library(lubridate)
```

The African event data set is imported as Excel file to the project and converted into a data frame to facilitate further analysis.

```
## 'data.frame':    184030 obs. of  27 variables:
##  $ ISO             : num  12 12 12 12 12 12 12 12 12 12 ...
##  $ EVENT_ID_CNTY   : chr  "ALG1" "ALG2" "ALG3" "ALG4" ...
##  $ EVENT_ID_NO_CNTY: num  1 2 3 4 5 6 7 8 11 10 ...
##  $ EVENT_DATE      : POSIXct, format: "1997-01-01" "1997-01-02" ...
##  $ YEAR            : num  1997 1997 1997 1997 1997 ...
##  $ TIME_PRECISION  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ EVENT_TYPE      : chr  "Violence against civilians" "Violence against civilians" "Violence agains
##  $ SUB_EVENT_TYPE  : chr  "Attack" "Attack" "Attack" "Attack" ...
##  $ ACTOR1          : chr  "GIA: Armed Islamic Group" "GIA: Armed Islamic Group" "GIA: Armed Islamic G
##  $ ASSOC_ACTOR_1   : chr  NA NA NA NA ...
##  $ INTER1          : num  2 2 2 2 2 1 2 2 2 1 ...
##  $ ACTOR2          : chr  "Civilians (Algeria)" "Civilians (Algeria)" "Civilians (Algeria)" "Civilian
##  $ ASSOC_ACTOR_2   : chr  NA NA NA NA ...
##  $ INTER2          : num  7 7 7 7 7 2 7 7 7 2 ...
##  $ INTERACTION     : num  27 27 27 27 27 12 27 27 27 12 ...
##  $ REGION          : chr  "Northern Africa" "Northern Africa" "Northern Africa" "Northern Africa" ..
##  $ COUNTRY         : chr  "Algeria" "Algeria" "Algeria" "Algeria" ...
##  $ ADMIN1          : chr  "Tipaza" "Relizane" "Saida" "Blida" ...
##  $ LOCATION        : chr  "Douaouda" "Hassasna" "Hassi El Abed" "Blida" ...
##  $ LATITUDE        : num  36.7 36.1 35 36.5 36.7 ...
##  $ LONGITUDE       : num  2.789 0.883 -0.29 2.829 2.789 ...
##  $ GEO_PRECISION   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ SOURCE          : chr  "www.algeria-watch.org" "www.algeria-watch.org" "www.algeria-watch.org" "w
##  $ SOURCE_SCALE    : chr  "National" "National" "National" "National" ...
##  $ NOTES           : chr  "5 January: Beheading of 5 citizens in Douaouda (Tipaza)." "Two citizens w
##  $ FATALITIES      : num  5 2 2 16 18 4 23 7 1 0 ...
##  $ TIMESTAMP       : num  1.55e+09 1.55e+09 1.55e+09 1.55e+09 1.55e+09 ...
```

The data frame currently entails 27 variables and both violent and non-violent events. We therefore create a new data set "events" that only entails deadly events and is limited to key variables we are interested in: The the perpetrator of violence, the target of violence, the first administrative unit where the event was observed (usually the state level), and finally the day of the event.

As political decision-makers might rather ask analysts to determine the likely magnitude of deadly events than the precise number of deaths, a new ordinal variable "DEATHS" is created based on four categories of fatality rates: 1-4, 5-9, 10-24 and 25+ deaths caused by a particular event.

As second variable "WEEK" is constructed by rounding the event dates to their week. This takes into account the fact that many days only see one deadly event happening.

```
##   1-4   5-9 10-24   25+
## 40491  8347  8105  3870
```

As it could be expected, deadly events tend to fall into the lowest category of death rates, and the frequency becomes much lower towards the categories of more extreme death rates. However, there is still significant variability across the categories.

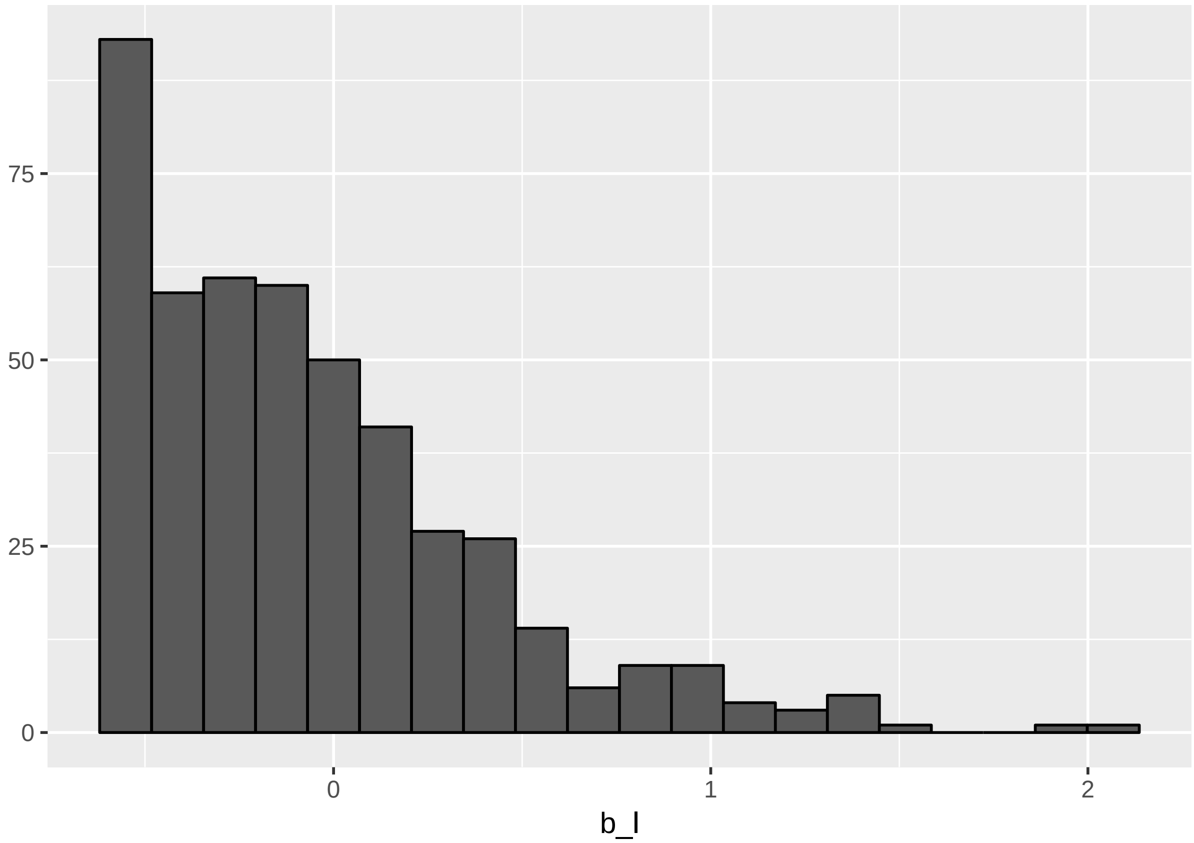
To facilitate further analysis, the explanatory variables are converted into factors and the death rate variable into an integer with values 1-4.

90 percent of the event data set are allocated to a training and 10 percent to a test test. By using repeatedly the semi-joint function, it is ensured that all levels of the key variables feature both in the test and the train set.

Based on the training set, several models are evaluated with regard to their predictive power, measured through the root-mean-square error (RMSE).

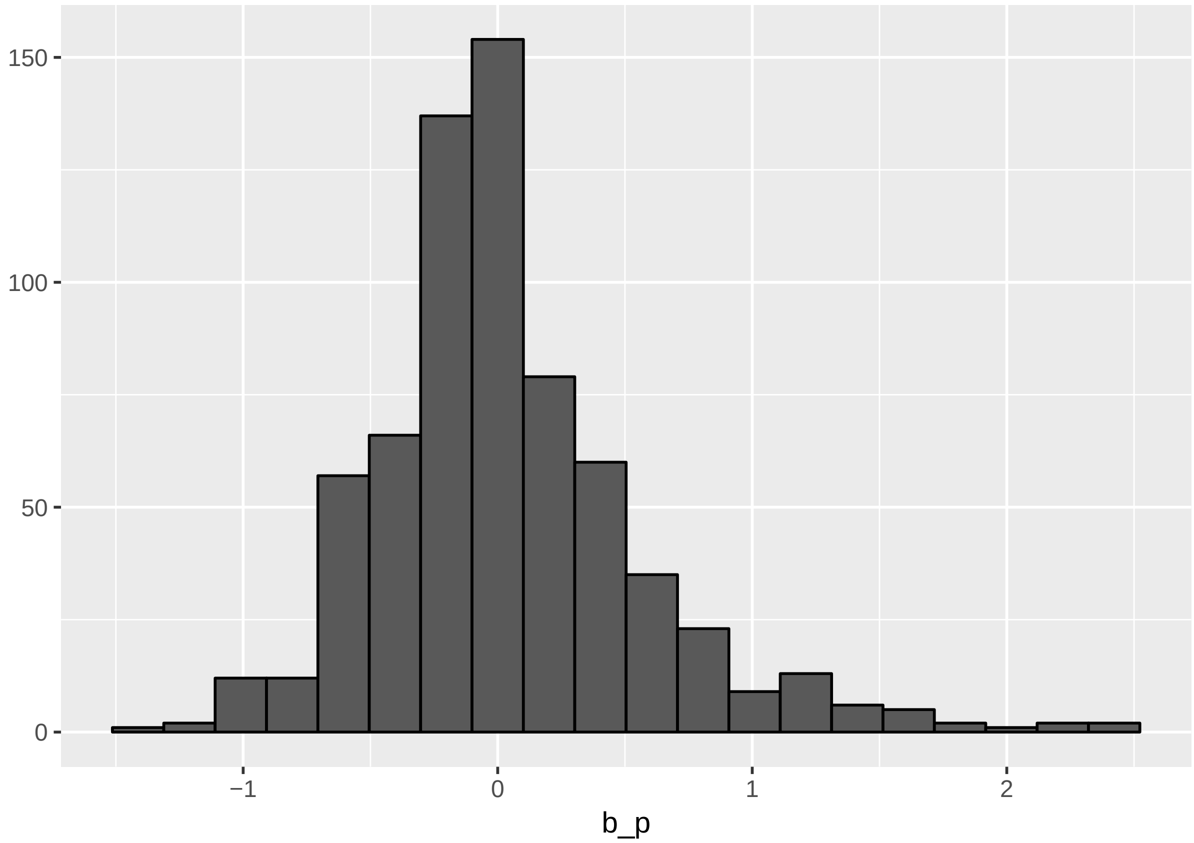
A first "naive" model predicts the death rate only based on the mean death rate of deadly events. The model results in an RMSE of 0.9360634.

A second model is based on the location (the first administrative unit) where deadly events occur. It would be plausible that areas with certain political, social, economic and environmental characteristics are more prone to a certain level of deadly violence than others. As indicated by the lower RMSE, adding a location effect b_l outperforms the "naive" model significantly. The histogram shows that knowing the location tends to shift the prediction somewhere between 0 and -0.5 death rates from the mean (1.5).



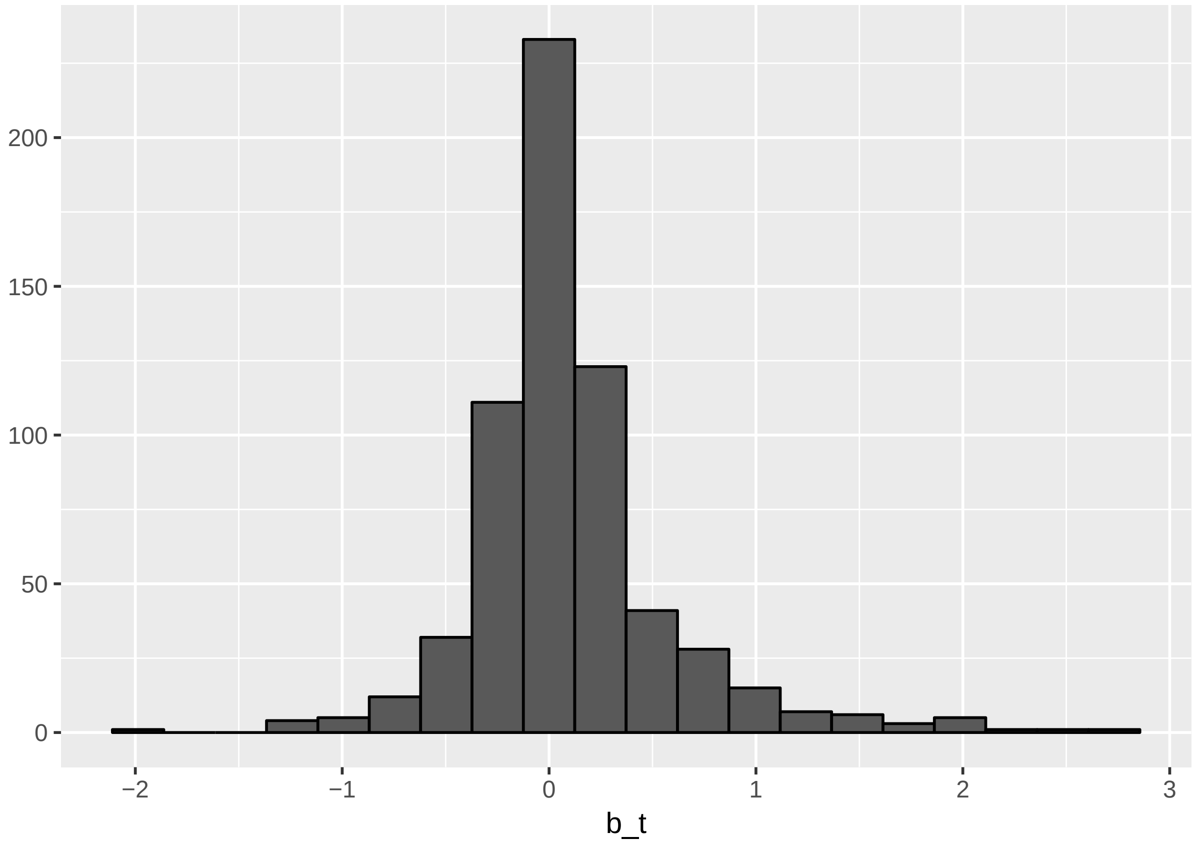
method	RMSE
Just the average	0.9360634
Location Effect Model	0.8623693

A third model is based on the assumption that knowing the armed actor perpetrating an attack can further improve the prediction of the magnitude of fatalities. As seen in the histogram, the predictions show some variation around the predictions of the previous model (indicated by the 0). However, the effect of knowing the perpetrator further improves the prediction, as indicated by the lower RMSE.



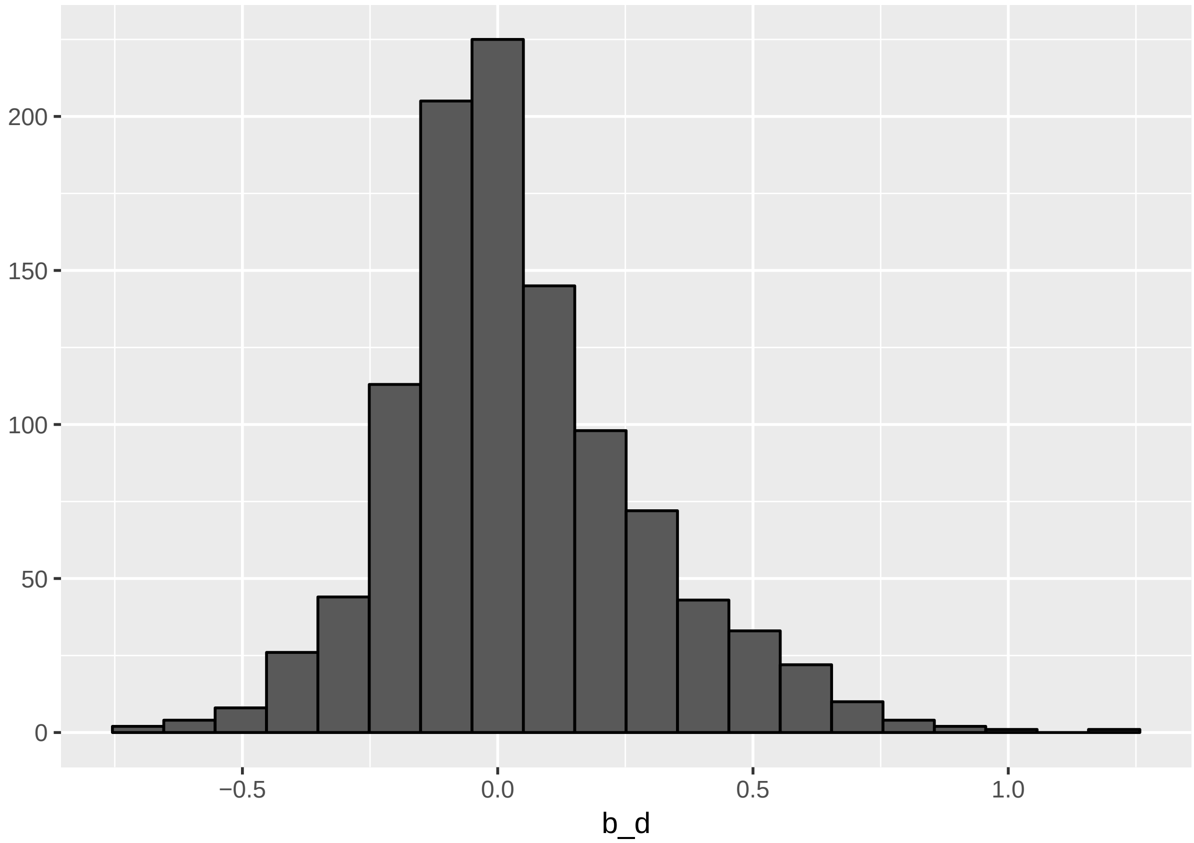
method	RMSE
Just the average	0.9360634
Location Effect Model	0.8623693
Location + Perpetrator Effects Model	0.8090158

A fourth model is constructed to include an effect b_t, based on knowing the potential targets of the deadly violence. The assumption is that some groups of people more often suffer from higher or lower levels of deadly violence that others. As seen in the histogram, the predictions show low variation and are close to those of the previous model. However, adding the target effect further improves the prediction, as indicated by the lower RMSE.



method	RMSE
Just the average	0.9360634
Location Effect Model	0.8623693
Location + Perpetrator Effects Model	0.8090158
Location + Perpetrator + Target Effects Model	0.7870915

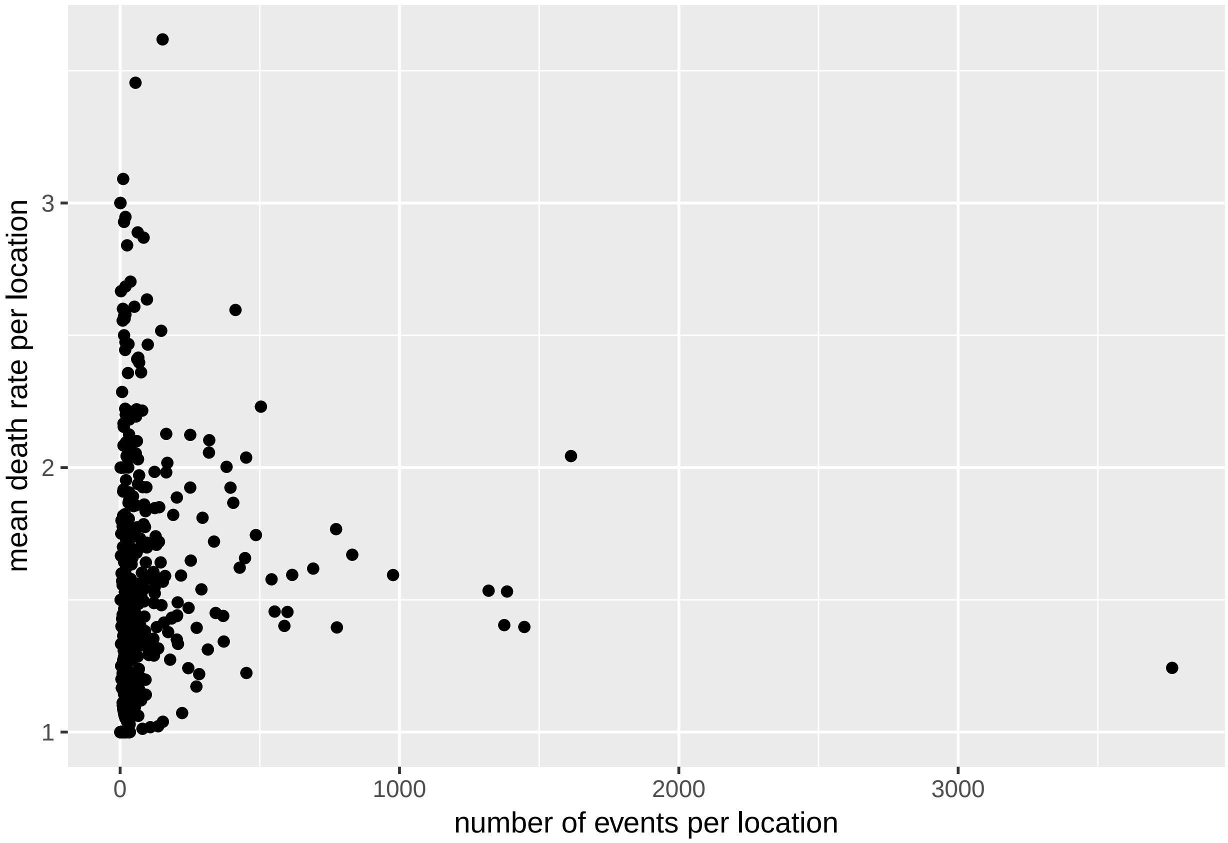
A fifth model is constructed by adding a time effect, based on the week of an event, to the previous model. The assumption is that the level of violence might vary over time. Adding a time effect significantly improves the predictive power compared to the previous model, as indicated by the lower RMSE.



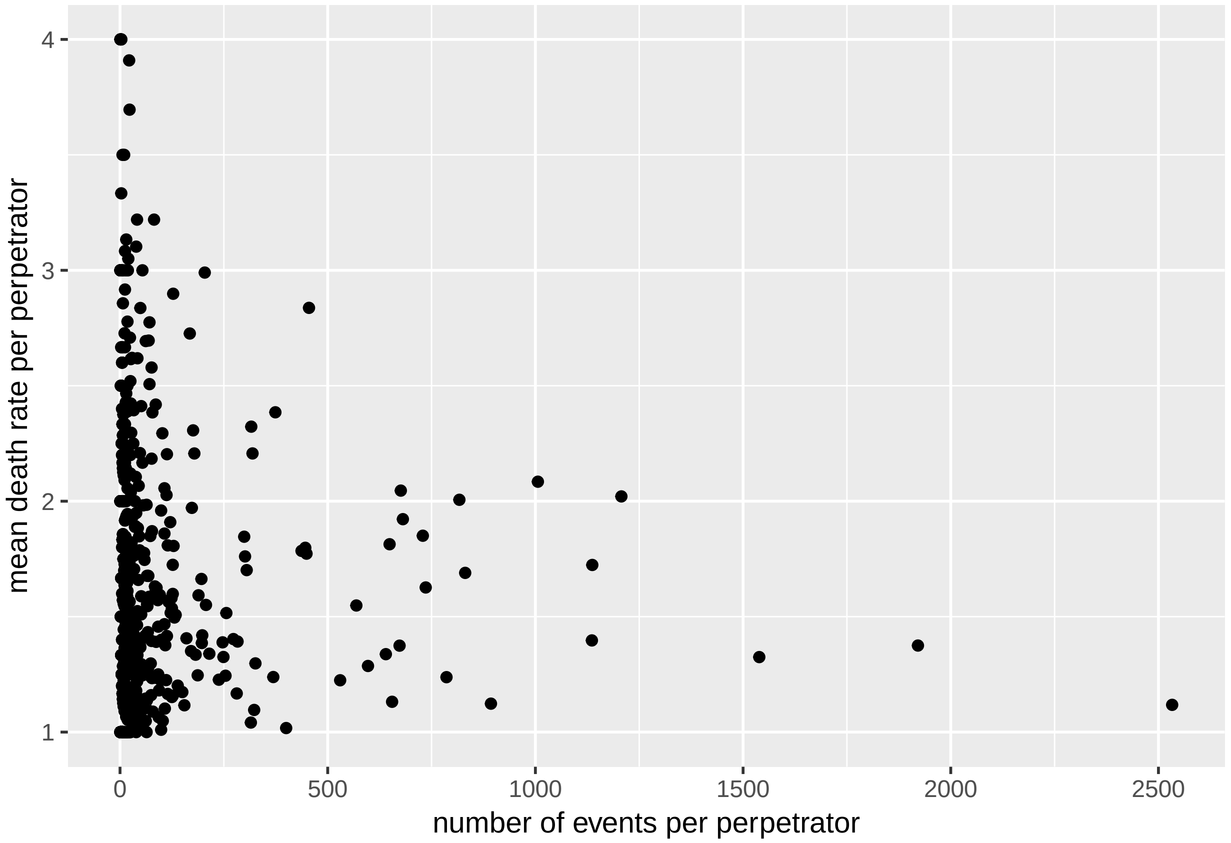
method	RMSE
Just the average	0.9360634
Location Effect Model	0.8623693
Location + Perpetrator Effects Model	0.8090158
Location + Perpetrator + Target Effects Model	0.7870915
Location + Perpetrator + Target + Event Date Effects Model	0.7651697

How can we further improve the predictive power of our model? We can observe that the number of events per location, perpetrator, target and week tends to be very low and some locations, perpetrators, targets and weeks with high mean death rates only saw one or two events.

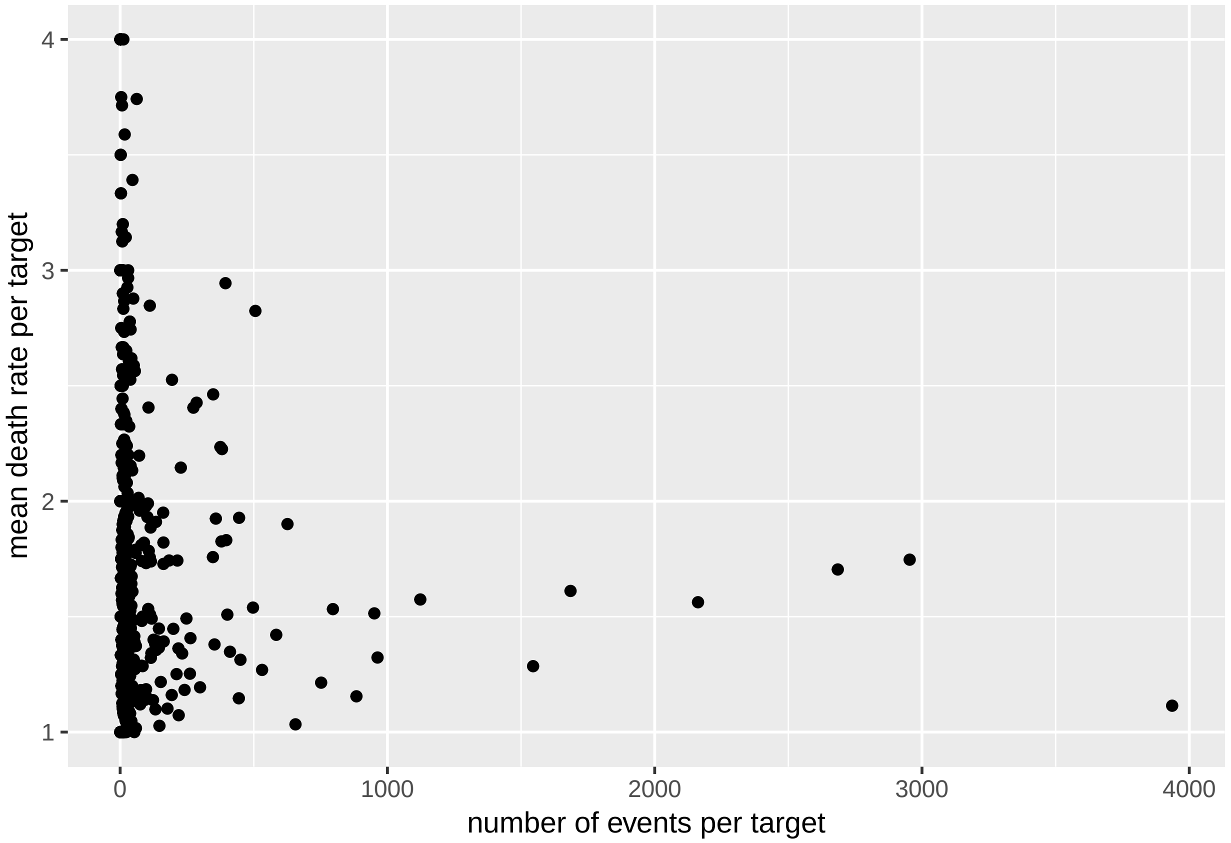
ADMIN1	events	rate
Bie	152	3.618421
Gash Barka	55	3.454546
Sennar	11	3.090909
Djibouti	1	3.000000
Loh-Djiboua	1	3.000000
Sud-Comoe	1	3.000000



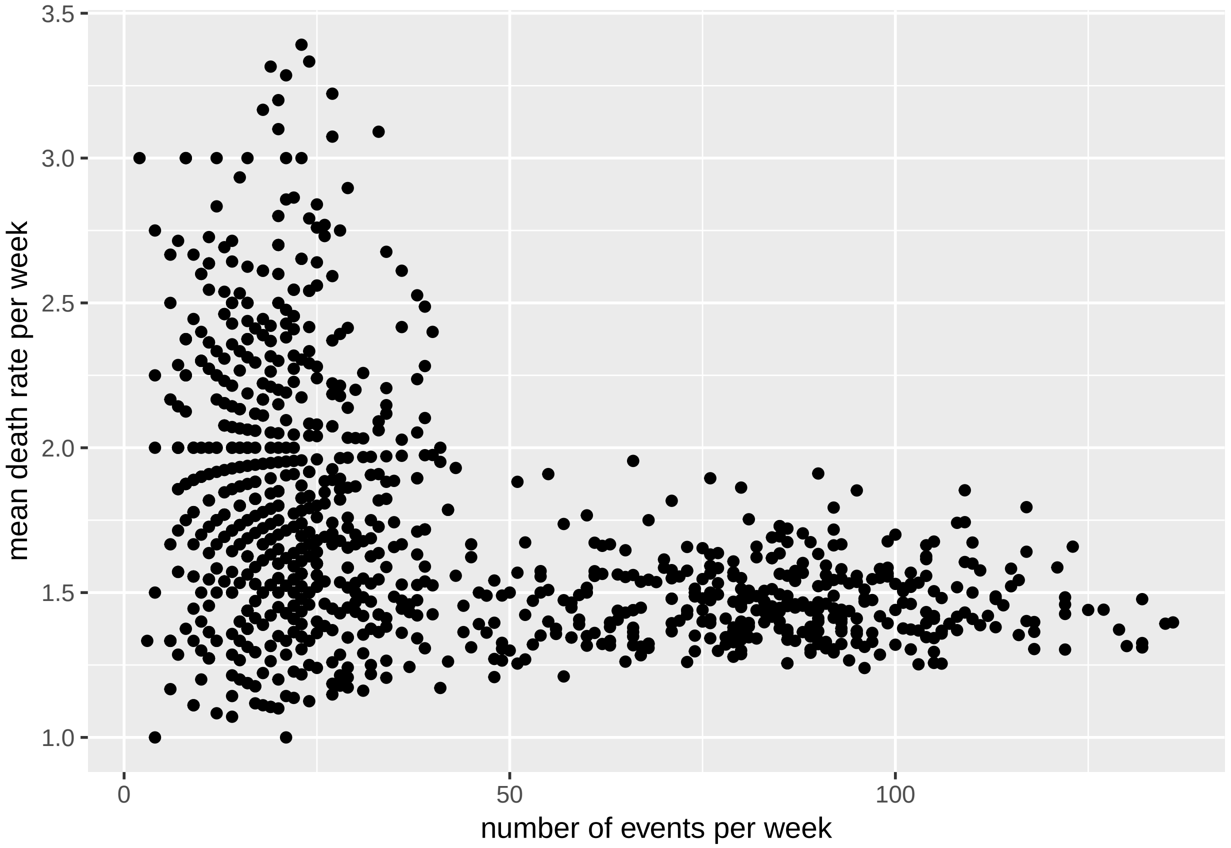
ACTOR1	events	rate
Abuok Communal Militia (South Sudan)	1	4.000000
BDK: Bunda Dia Kongo	1	4.000000
ECOMOG: Economic Community of West African States Monitoring Group	1	4.000000
Messalit Ethnic Militia (Kenya)	3	4.000000
MNJTF: Multinational Joint Task Force	3	4.000000
Lou Nuer Ethnic Militia (Sudan)	22	3.909091



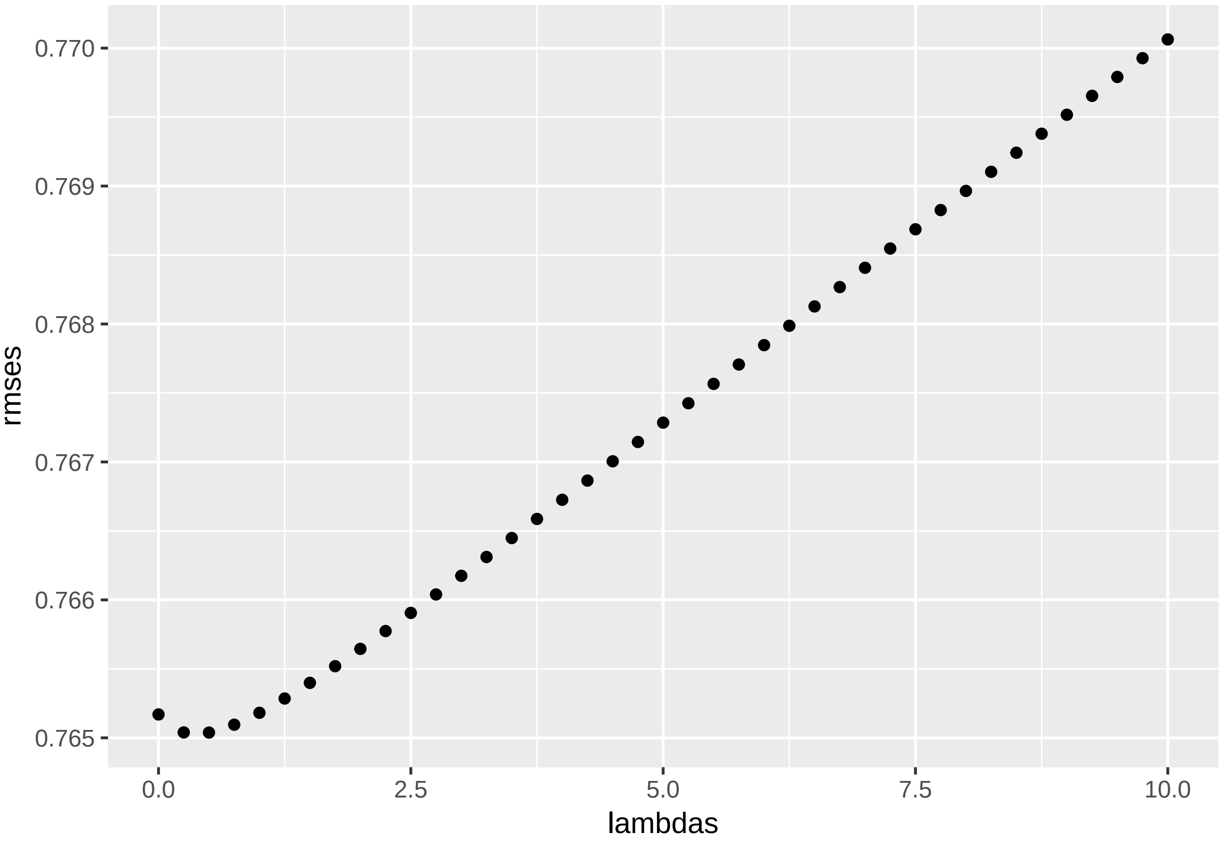
ACTOR2	events	rate
Alel Thony Communal Militia (South Sudan)	1	4
ANR: National Resistance Army	1	4
Benishangale Ethnic Militia (Ethiopia)	12	4
Former Military Forces of South Sudan (2011-)	1	4
FUC: United Front for Change	1	4
Irigwe Ethnic Militia (Nigeria)	1	4



WEEK	events	rate
1999-02-07	23	3.391304
1999-03-21	24	3.333333
1999-01-10	19	3.315790
1999-01-17	21	3.285714
1999-07-18	27	3.222222
1998-12-20	20	3.200000



To control for this variation effect, regularization techniques are applied to the previous model. Penalized regression may control for the total variability of the frequency of events across locations, perpetrators, targets and weeks. The penalty term lambda is a tuning parameter and we will choose it through cross-validation. As we see below, regularization slightly further decreases the RMSE of the previous model by adding Lambda = 0.5.



method	RMSE
Just the average	0.9360634
Location Effect Model	0.8623693
Location + Perpetrator Effects Model	0.8090158
Location + Perpetrator + Target Effects Model	0.7870915
Location + Perpetrator + Target + Event Date Effects Model	0.7651697
Regularized Location + Perpetrator + Target + Date Effect Model	0.7650384

The best-performing model is now applied to the test set. It turns out that the algorithm predicts the level of deadly violence even better for the test set than for the training set.

method	RMSE
Just the average	0.9360634
Location Effect Model	0.8623693
Location + Perpetrator Effects Model	0.8090158
Location + Perpetrator + Target Effects Model	0.7870915
Location + Perpetrator + Target + Event Date Effects Model	0.7651697
Regularized Location + Perpetrator + Target + Date Effect Model	0.7650384
Test of Regularized Location + Perpetrator + Target + Date Effect Model	0.6381861

Discussion of results

The results show that we can use location, perpetrator, target and date averages to predict the scale of fatalities in deadly political violence in Sub-Saharan Africa and the Maghreb with much more confidence

(0.20 RMSEs better) than by simply basing our guess on the mean death rate. The testing of the model with an independent sample actually led to better results than with the training sample.

Conflict early warning systems usually attempt to predict the occurrence of armed conflict on a yearly and national level. In these cases, conflict is simply assumed to “start” with 25 battle-related deaths per year. The present project took a different approach and dis-aggregated conflict into its deadly encounters per week in a specific location. Hence, predictions are much more timely and relevant to the local context, as they also consider violence much below the normal 25 deaths threshold.

There is, however, a caveat to the robustness of the findings due to the way how the samples were constructed. The original aim of the project was to predict all types of violent events, including those that do not result in any deaths. However, the ACLED data does not include a variable coding for violence per se. It therefore has to be kept in mind that the selection of only deadly events introduces significant bias to the sample.

Conclusion

This small project has demonstrated the potential usefulness of using machine learning techniques on dis-aggregated armed conflict data to predict the magnitude of the deadly consequences of conflict events. The final model includes variables on the location, perpetrator, target and date of a violent event.

The next step of this project would be to construct a forecasting system for deadly political violence by adding further predictors, including various political, social and economic variables.