Mark Cotton

DS-SEA-06

PREVENTING GENOCIDE

Using Data Science to identify early indicators of Future State led mass killings

Table of Contents

[1. Background 2](#_Toc481536753)

[2. Problem statement and hypothesis 2](#_Toc481536754)

[3. Description of your data set and how it was obtained 3](#_Toc481536755)

[4. Description of any pre-processing steps you took 3](#_Toc481536756)

[5. What you learned from exploring the data, including visualizations 5](#_Toc481536757)

[6. How you chose which features to use in your analysis 5](#_Toc481536758)

[7. Details of your modeling process, including how you selected your models and validated them 6](#_Toc481536759)

[8. Your challenges and successes 6](#_Toc481536760)

[9. Possible extensions or business applications of your project 6](#_Toc481536761)

[10. Conclusions and key learnings 6](#_Toc481536762)

[11. References 6](#_Toc481536763)

# Background

On April 6, 1994, Rwandan President Juvenal Habyarimana, a Hutu, was shot down above Kigali airport. In the following 100 days, between 800,000 to 1,000,000 (perhaps even 2 M) Tutsi and moderate Hutus were slaughter were killed by fellow Rwandans, with violence being encouraged by the state. This averaged to 6 men, women and children were murdered every minute of every hour of every day for 3 months1.

Since the Holocaust during World War II, there have been numerous mass killings classified as genocide, from atrocities in Cambodia to Rwanda to Bangladesh to name a few, resulting in the slaughter of millions of people. According to GenocideWatch.org, genocides and other mass murders killed more people in the twentieth century than all the wars combined.

When the killings began in Rwanda, the international community responded inadequately and was not able to prevent the hundreds of thousands of deaths that followed. Did the world see this catastrophe coming? Could this all have been prevent or at least could the damage have been minimized? The biggest question is this: how do we prevent history from repeating itself again?

What can be done to prevent genocide? From a report by the Genocide Prevention Task Force from the United States Institute of Peace,

*“The first major element of a comprehensive system to prevent genocide and mass atrocities is a reliable process for accessing risks and generating early warning of potential atrocities... Effective early warning does not guarantee successful prevention, but if warning is absent, slow, inaccurate, or indistinguishable from the 'noise' of regular reporting, failure is virtually guaranteed.2”*

What follows, per the Early Warning Project, is that genocide is nearly always carried out by a country’s own police and military forces, requiring external (international) intervention. To make predictions before a country has fallen in to a large-scale genocide, the marker to watch for is state-led mass killings. State-led mass killings occur when a country uses its own forces to kill over 1000 non-combatant civilians in a 12-month time period (in a country of over 500,000).

# Problem statement and hypothesis

Identifying early indicators that mass killings are going to occur will allow a better opportunity for prevention.

Can data science machine learning principles be used to create a model to predict countries with a high likelihood of having a new state-led mass killing occur before the killings occur? The EWP already has created model to make these predictions. My hypothesis is that from within this data, certain combinations of features will be highly indicative of new mass killings occurring. I predict that lingual fractionalization, ethnic fractionalization, GDP, trade openness, country age, availability of natural resources, and history of mass killings will have strong correlations with new mass killings occurring. The model created will be in the form a logistic regression, predicting a 1 or 0 for new mass killings occurring in the following year, with some confidence percentage.

# Description of your data set and how it was obtained

The data set used for this report is all directly pulled from a Github repo, where the Early Warning Project (EWP) has aggregated data from multiple sources. The EWP runs statistical risk assessment with the goal of preventing mass atrocities. The data from publicly available sources and was cleaned by EWP. This is a time series dataset spanning 178 countries from 1945-2014. This report is using the pre-processed dataset which includes 9330 observations (rows) with 276 variables (columns). The file “EWP Data Dictionary 20140909.pdf” describes the sources and variables. The information below is copied from that document outlining the sources:

* mkl = Early Warning Project’s Episodes of State-Led Mass Killing Dataset
* wdi = World Bank’s World Development Indicators (via R package ’WDI’)
* mev = Center for Systemic Peace’s Major Episodes of Political Violence (<http://www.systemicpeace.org/inscrdata.html>)
* pol = Polity IV (<http://www.systemicpeace.org/inscrdata.html>)
* imr = U.S. Bureau of the Census, International Division (via PITF)
* cmm = Center for Systemic Peace’s List of Coups d’Etat (<http://www.systemicpeace.org/inscrdata.html>)
* cpt = Powell and Thyne’s Coups d’Etat, 1950 to Present (<http://www.uky.edu/~clthyn2/coup_data/home.htm>)
* cou = An amalgamation of cmm and cpt data
* pit = PITF Problem Set (i.e., episodes of instability) (<http://www.systemicpeace.org/inscrdata.html>)
* dis = Center for Systemic Peace’s Discrimination Data Set (via PITF)
* imf = International Monetary Fund’s World Economic Outlook Report (<http://www.imf.org/external/pubs/ft/weo/2014/01/weodata/index.aspx>)
* elf = Anderson’s Ethnic Fractionalization Data (<http://www.anderson.ucla.edu/faculty_pages/romain.wacziarg/papersum.html>)
* hum = Farris et al.’s Latent Human Rights Protection Scores (<http://humanrightsscores.org/>)
* fiw = Freedom House’s Freedom in the World Data Set (<https://freedomhouse.org/report-types/freedom-world#.VA7Sufk7u-M>)
* aut = Geddes, Wright, and Frantz’s Authoritarian Regimes Dataset (<http://sites.psu.edu/dictators/>)

# Description of any pre-processing steps you took

The EWP uses three different models to predict new mass killings. For the “general” model, Random Forests is used with “no feature selection involved. These models were specified a priori to represent existing theories on the causes of mass killings, so this [R] script is only used to demonstrate that the resulting models and the average of forecasts from them do have real predictive power.” Below is the description of the specific features selected ahead a priori:

* mkl\_start: Onset of any episodes of state-led mass killing
* reg\_afr: US Dept State region: Sub-Saharan Africa
* reg\_eap: US Dept State region: East Asia and Pacific
* reg\_eur: US Dept State region: Europe and Eurasia
* reg\_mna: US Dept State region: Middle East and North Africa
* reg\_sca: US Dept State region: South and Central Asia
* reg\_amr: US Dept State region: Americas
* mkl\_ongoing; Any ongoing episodes of state-led mass killing
* mkl\_ever: Any state-led mass killing since WWII (cumulative)
* countryage\_ln: Country age, logged
* wdi\_popsize\_ln: Population size, logged
* imr\_normed\_ln: Infant mortality rate relative to annual global median, logged
* gdppcgrow\_sr: Annual % change in GDP per capita, meld of IMF and WDI, square root
* wdi\_trade\_ln: Trade openness, logged
* ios\_iccpr1: ICCPR 1st Optional Protocol signatory
* postcw: Post-Cold War period (year ≥ 1991)
* pol\_cat\_fl1: Autocracy (Fearon and Laitin)
* pol\_cat\_fl2: Anocracy (Fearon and Laitin)
* pol\_cat\_fl3: Democracy (Fearon and Laitin)
* pol\_cat\_fl7: Other (Fearon and Laitin)
* pol\_durable\_ln: Regime duration, logged (Polity)
* dis\_l4pop\_ln: Percent of population subjected to state-led discrimination, logged
* elf\_ethnicc1: Ethnic fractionalization: low
* elf\_ethnicc2: Ethnic fractionalization: medium

Note: Final category not needed: elf\_ethnicc3: ethnic fractionalization: high

* elf\_ethnicc9: Ethnic fractionalization: missing
* elc\_eleth1: Salient elite ethnicity: majority rule
* elc\_eleth2: Salient elite ethnicity: minority rule

Note: Final category not needed: elc.elethc : Salient elite ethnicity (yes/no)

* elc\_eliti: Ruling elites espouse an exclusionary ideology
* cou\_tries5d: Any coup attempts in past 5 years ((t-4) to (t))
* pit\_sftpuhvl2\_10\_ln: Sum of max annual magnitudes of PITF (Political Instability Task Force) instability other than genocide from past 10 yrs ((t-9) to (t)), logged
* mev\_regac\_ln: Scalar measure of armed conflict in geographic region, logged
* mev\_civtot\_ln: Scale of violent civil conflict, logged

From the features listed above, there were 14563 nulls, with 2669 nulls in wdi\_trade\_ln, 2295 in gdppcgrow\_sr, 1685 in wdi\_popsize\_ln.

Total nulls in filtered data: 14563

mkl\_start 0

reg\_afr 0

reg\_eap 0

reg\_eur 0

reg\_mna 0

reg\_sca 0

reg\_amr 0

mkl\_ongoing 0

mkl\_ever 0

countryage\_ln 0

wdi\_popsize\_ln 1685

imr\_normed\_ln 966

gdppcgrow\_sr 2295

wdi\_trade\_ln 2669

ios\_iccpr1 994

postcw 0

pol\_cat\_fl1 144

pol\_cat\_fl2 144

pol\_cat\_fl3 144

pol\_cat\_fl7 144

pol\_durable\_ln 175

dis\_l4pop\_ln 0

elf\_ethnicc1 0

elf\_ethnicc2 0

elf\_ethnicc9 0

elc\_eleth1 934

elc\_eleth2 934

elc\_eliti 934

cou\_tries5d 360

pit\_sftpuhvl2\_10\_ln 1507

mev\_regac\_ln 267

mev\_civtot\_ln 267

To attempt to clean this data, nulls are replaced with the average values:

dft = train.fillna(train.mean())

The issue with this process of imputing missing data is that the mean value may be very misrepresentative for a given country. For example, Yugoslavia has no population size information for any years – is the mean going to be a fare estimate? This country turned in to Bosnia and Herzegovina, Croatia, Federal Republic of Yugoslavia (later Serbia and Montenegro), Macedonia, and Slovenia. Yugoslavia had mass killings start in 1945 and 1991 and Bosnia and Herzegovina in 1992. Even though mass killings happened in Yugoslavia before the country split in 1991, none of these countries have “history” of mass killings (mkl\_ever = 0). (Bosnia and Herzegovina had mass killings start in the first year and therefore have mkl\_ever=1.) Most striking to me from this data is that Yugoslavia (mkl\_ever = 1) split in to multiple countries including the Federal Republic of Yugoslavia (mkl\_ever = 0) which later became Serbia and Montenegro (mkl\_ever = 1) and then became the separate countries Serbia (mkl\_ever = 0) and Montenegro (mkl\_ever = 0). *Both the Federal Republic of Yugoslavia and (Serbia and Montenegro) never had mkl\_start=0 and yet Serbia and Montenegro have history of mass killings.*

# What you learned from exploring the data, including visualizations

Some holes in data – information for history of country starts at 0 generally (ie, mkl\_ever) if the country is born in the dataset.

# How you chose which features to use in your analysis

Features selected a priori. Would like to use additional features but want to refine model to be more than 13% accurate with recommended features first. (Don’t want to optimize a broken model – want to find out what is broken first and fix that.)

# Details of your modeling process, including how you selected your models and validated them

Logistic regression using defined features. Confusion matrix using to refine regression – want to capture more true positives to make sure my model can catch all mkl\_starts.

# Your challenges and successes

Nulls may require categorical modelling (KNN, DBSCAN clustering, ideally, model values for each country.) If possible, use info about which countries became other countries (or use correlations between yrborn and yrdied to assume association when possible). Also, model year over year change per feature for making estimates?

# Possible extensions or business applications of your project

Help further refine Early Warning Project models to influence results.

# Conclusions and key learnings

Certain features had high corr with new mkl:

* Blah
* Blah

Certain additional features were added to model to increase accuracy:

* Blahg

Certain features were unimportant and therefore removed:

* That one

# References

1 <http://survivors-fund.org.uk/resources/rwandan-history/statistics/>  
2Preventing Genocide, A Blueprint for U.S. Policymakers, Madeleine K. Albright, William S. Cohen  
<https://www.usip.org/genocide-prevention-task-force/view-report>