SAND2020-5141C

Assessing Extreme Value Analysis to predict rare events from the Global Terrorism Database

J. Gabriel Huerta, Lekha Patel, Lyndsay Shand, J. Derek Tucker, William Miller*

*Department of Statistics Sciences, Sandia National Laboratories, Albuquerque, NM USA

Sandia National Laboratories

Abstract

Extreme value methods have commonly been used to predict and quantify uncertainty around environmental or climatological events that could have high impact on human casualties or costs (e.g., earthquakes, hurricanes, flooding, wildfires). In this work, our focus is to study the number of casualties as the variable of interest, from the Global Terrorism Database (GTD) for a particular region and time frame and characterize extreme events via finding observations that exceed a given threshold and fitting a Generalized Pareto Distribution (GPD) to these exceedances. We assess whether the goodness of fit of the GPD parameters are adequate for our framework. We also provide graphical representations of predicted 95% and 99% quantiles based on our models and compare these to the actual data. The results of these analyses are a building block into the development of a representative Bayesian hierarchical model that fully characterizes the spatial-temporal relationships present in extreme events from the GTD.

GTD database: Introduction

- The GTD database[a] contains all known attacks in 199 different countries between the years 1970-2018.
- Data for each attack include: specific attack regions within countries, attack time (day, month and year), attack type with description (e.g. suicide bombing), target type and subtype (e.g. military), nationality of attackers, used weapon types (e.g. firearms), number of people killed and wounded, property damage (monetary value), attack time duration and ransom descriptions.
- Define an event as the number of casualties = number of people killed + number wounded from each terrorism attack.
- Our question is whether we can quantify the occurrence of extreme terrorism events using extreme value analysis.
- Our method is to develop a model that analyzes the risk of extreme terrorism events at an arbitrary space/time location.

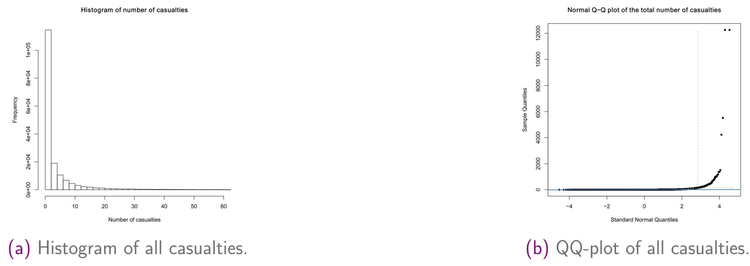
[a]This is an open-source database which can be downloaded at https://www.start.umd.edu/gtd/.

Data challenges

- The database contains many missing entries on one or more of the explanatory covariates, number of casualties and the time stamps of events.
- Many of spatial locations of events are also non-exact, to the approximate longitude/latitude of the nearest city.
- Many countries, especially those from before 2000, have underreported attacks, in terms of intensity and number of reported causalities.
- Data limitations make EVA difficult for certain countries e.g. the USA and UK from which there is not enough data to pursue significant analysis.
- Insufficient data renders EVA via the block maximum approach inappropriate, enabling subsequent analysis to be done through the threshold exceedance method.

Motivation for EVA

- Across all countries and years, empirical data indicates a zero-inflated heavy tailed distribution on the number of total casualties.



(a) Histogram of all casualties. (b) QQ-plot of all casualties.

Figure 1: Plots of all global casualties during 1970-2018.

- The countries which have quoted the most deadly attacks in the database are: Iraq, Pakistan, Afghanistan, India and Colombia.

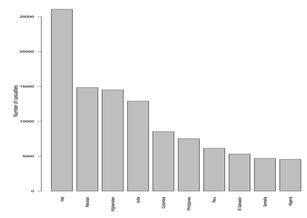


Figure 2: Histogram of causalities per country.

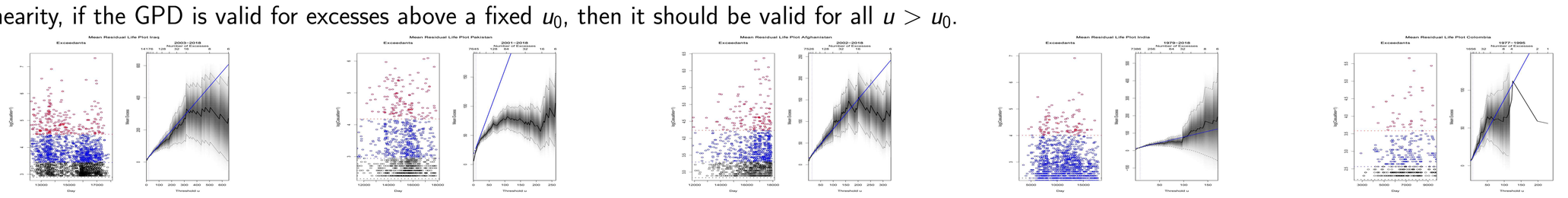
- Empirical CDFs and quantile plots suggest the distributions between pairs of countries (Iraq/Afghanistan, India/Pakistan/Colombia) are similar and heavy tailed.

(a) Empirical CDF of casualties on log scale. (b) Empirical quantile plots of casualties on log scale.

Figure 3: CDF and quantile plots of casualties in the five most deadly countries during 1970-2018.

Data analysis & EVA

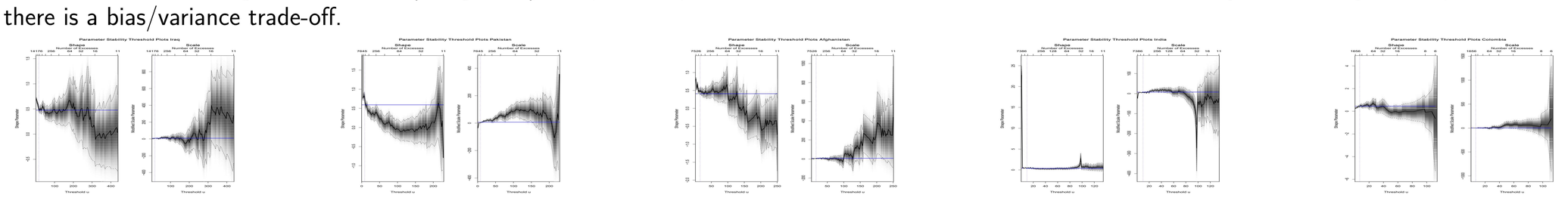
- General EVA via the threshold exceedance method requires a threshold u to be established apriori to the analysis.
- Looking at the most fatal countries within specific time periods, if the number of casualties X follows a GPD with shape parameter ξ and scale parameter σ, then then excess distribution $X - u$ follows a GPD with the same shape parameter and modified scale parameter $\sigma(u) = \sigma + \xi u$. The mean excess function is the expected value of $X - u$ given $X > u$, and should be linear in u.
- Due to linearity, if the GPD is valid for excesses above a fixed u_0, then it should be valid for all $u > u_0$.



(a) Iraq event exceedants 2003-2018. (b) Pakistan event exceedants 2001-2018. (c) Afghanistan event exceedants 2002-2018. (d) India event exceedants 1979-2018. (e) Colombia event exceedants 1977-1995.

Figure 4: For each country, event exceedants above the 90th empirical quantile (black), 95th quantile (blue) and 99th quantile (red) are shown (left) with their mean residual life plots at different thresholds (right).

- Since the GPD is valid for excesses above all $u > u_0$, the ξ and $\sigma(u)$ remain constant for higher thresholds.
- Can estimate the model at a range of thresholds (using MLE) and plot the parameter estimates. Thresholds can be determined by the stability of the parameter estimates, although there is a bias/variance trade-off.



(a) Iraq parameter estimates 2003-2018. (b) Pakistan parameter estimates 2001-2018. (c) Afghanistan parameter estimates 2002-2018. (d) India parameter estimates 1979-2018. (e) Colombia parameter estimates 1977-1995.

Figure 5: For each country, parameter estimates for the shape and modified scale of the GPD, are plotted at different thresholds. A stable parameter estimate (horizontal line) indicates a good fit of the corresponding GPD distribution.

- The most suitable threshold is both country and time dependent.
- Plots of the empirical mean excess functions and parameter stability indicate a good GPD fit for Iraq, Afghanistan and India. Pakistan has a poorer fit and insufficient data is highlighted for Colombia.

Conclusions & Future work

- While the GPD generally indicates a good fit with some countries' data, poorer fits with other countries highlight a need for a fully flexible spatio-temporal model. This model could also incorporate other covariates given in the GTD database.
- The threshold u could therefore be both spatio-temporal and covariate dependent, or as an additional unknown parameter in the model.
- Data challenges with the GTD database, with particular emphasis on missing and approximate data need to be succinctly addressed.
- Spatio-temporal models that include the bulk of the data as well as the extremes may alleviate the data limitations that are currently present.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. SAND NO. 2020–00000