Prediction of armed conflict: What can we say about uncertainty in such models?

Presentation to the  'Godt Hjort' seminar

Håvard Hegre[1,2]

[1]Peace Research Institute Oslo (PRIO)
[2]Department of Peace and Conflict Research, Uppsala University

5 December 2023

Stability and Change

Project at Center for Advanced Studies, Oslo, 2022–2023

Directed by  and  Centre for Advanced Study at the Norwegian Academy of Science and Letters CAS OSLO

Stability and Change – Statistics and peace research

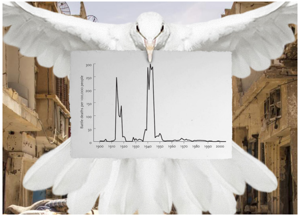
Is the world becoming more peaceful?

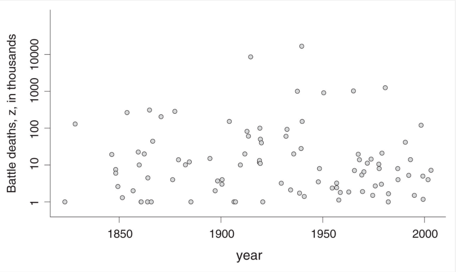
- – are ‘better angels’ getting the upper hand?

Cunen, Hjort & Nygård, 2020: ‘Statistical sightings of better angels’

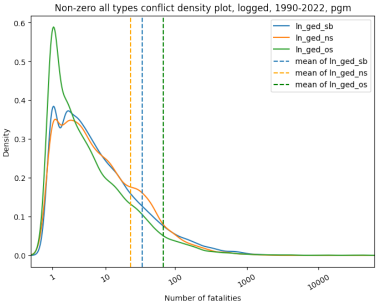
- Fatalities in wars over the 1823–2005 – is there a point where death counts change?
- Point of maximal change in 1950 – Korean war
- Upper quartile of the battle-deaths distribution decreases from 63,545 before to 14,943 after

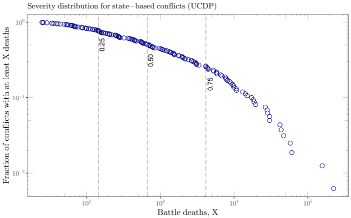
Challenge: Fatality counts in war have power-law distribution





Topics in *Stability and Change*





- Change points
- Power-law distributions
- Migration
- Forecasting armed conflict: VIEWS
  - Predicting with uncertainty
  - Given the near-power law distribution
  - *AND* zero-inflation: 99.5% zeros
- What characterizes a good prediction model?

VIEWS: Violence and Impacts Early Warning System

At PRIO and Uppsala University

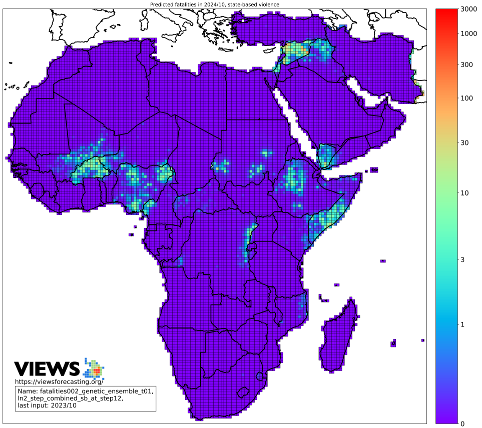
- with Céline, Gudmund, Jonathan Williams

Forecasting problem:

- Fatalities in armed conflict
- At country and geographical level
- 36 months into the future
- Currently as point predictions

Optimized through MSE on $ln(Y+1)$

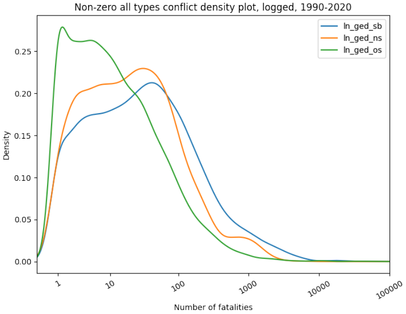
- Log transformation beneficial
    - to avoid underpredictions
    - and reduce influence of extreme cases
- Or is it?



Machine-learning models

Core algorithms: Decision-tree models

- Random forests (XGB implementation)
- Gradient boosting models (XGB/LGB/sklearn implementations)

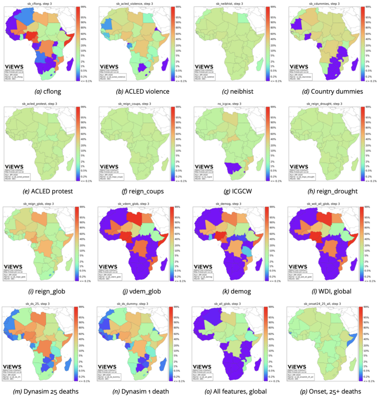


Distribution of outcome challenge – Solutions:

- Predicting $log(Y+1)$
- Hurdle models (Fritz et al 2022)
  1. Learn probability of non-zero observations $\hat{p}_{nz} = p(Y > 0)$
  2. Learn number of fatalities if non-zero $\hat{Y}_{nz} = Y|Y > 0$
  3. Combined prediction $\hat{Y} = \hat{p}_{nz} \times \hat{Y}_{nz}$
- Markov models (Randahl & Vegelius 2022)

Ensembles of constituent model predictions

Figure B-3. Predicted probabilities, constituent models, **sb**, $s = 3$ (month 483, March 2020), based on data up to December 2019

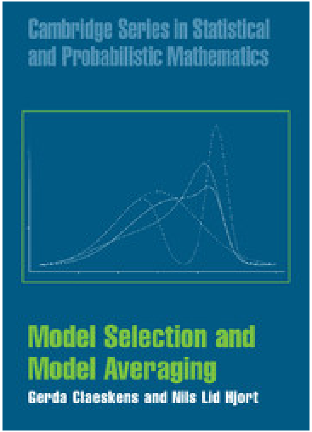


Models combining algorithms and feature sets

- Final model: an ensemble
- Optimal weights found by a genetic algorithm
- Optimized on log MSE per country month

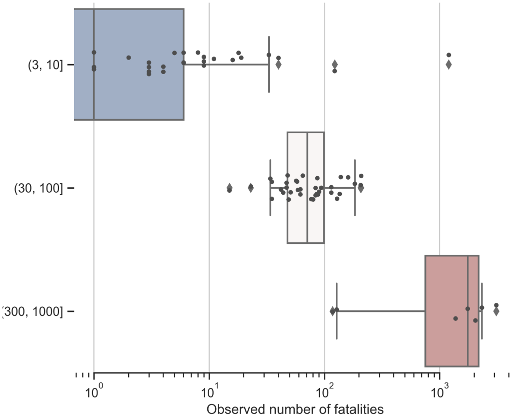
We are unsure what is the best optimization criterion

- What should be our 'Focused Information Criterion'?



How well do we predict?

MSEs at country level between .25 and .75



How many were killed per country if we predict the following 12 months into the future:

- 300–1000 fatalities:
  - all are above 100, and 90% are above 800
- 30–100 fatalities:
  - 90% are between 30 and 200
- 3–10 fatalities:
  - 50% are 1 or higher, median observation is 1, and 95% are below 30

We underpredict onset cases

Read more: Forecasting Fatalities (Hegre et al. 2022) (Model: *fatalities001*)

Optimizing on log MSE at country month level not such a good idea!

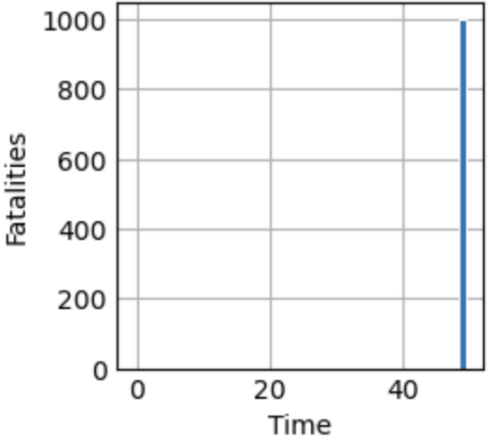
An example: only 'structural' predictors

Consider the situation:

- We observe 49 months of no violence, then 1,000 deaths
- We have no time-varying predictors
- Common in our application!

What is the best prediction per month?

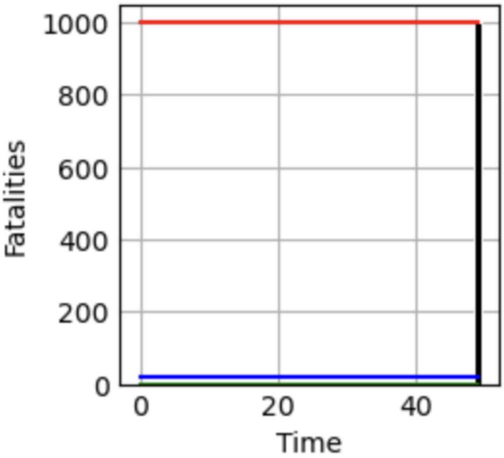
- Nils' immediate response: '20?'



What is the best prediction per month?

What is the best point prediction per month?

- 0? (best in most months)
- 1000? (best in most interesting month)
- 20? (best calibrated at large – 1,000 predicted deaths over the 50 months)
- 0.15 – the exponential of ln(1,000)/50?



What do the metrics say?

- Comparing performance of various constant-level point predictions

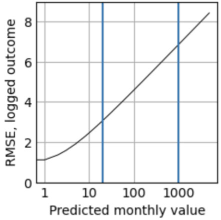
Which models do metrics prefer?

- RMSLE at country month: Models close to zero
- MSE at country month: Very uncertain about what is best, weakly preferring 20/month
- RMSLE at entire period: Prefers 20/month

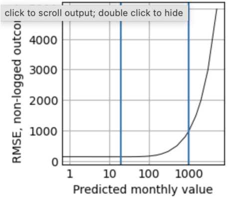
Conceivably, the best prediction is 98% 0 and 2% 1000

- Prediction with uncertainty necessary to select a good model!

Logged outcome:



Nonlogged outcome:



How to span the uncertainty of armed conflict forecasts

Aim:

- To produce VIEWS forecasts as probability distributions over possible fatality counts

Solution:

- Formulating models capturing:
  - uncertainty about model specification
  - sample variation/statistical uncertainty
  - uncertainty of input data

Construct ensembles:

- Extract draws from the probability distributions models imply
- Weight models by CRPS (?)
- At various temporal and spatial resolutions

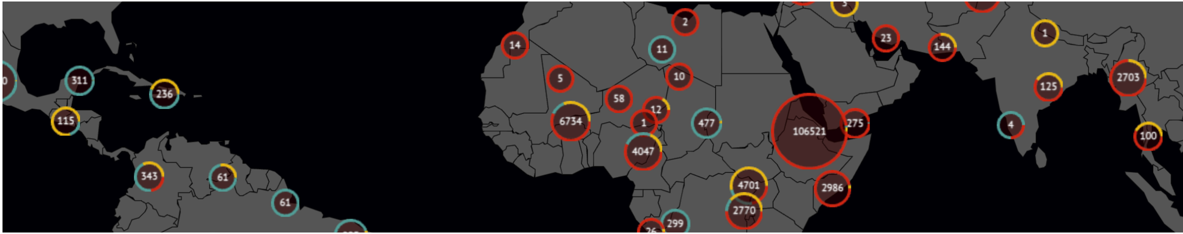
Sample variation

A model-agnostic approach

- Produce predictions for hold-out partition
- Bin them in ranges
- Draw prediction errors at random within bins as estimate of uncertainty
- *Stability and Change* solution: Conformal prediction

Quantile regression

Uncertainty regarding input data: UCDP uncertainty

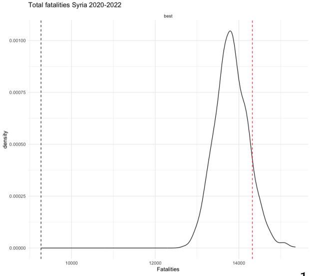


How many did really die in each conflict?

- Uppsala Conflict Data Program
- UCDP publishes only counts they can solidly verify
- But UCDP knows more people died

Solution: Complementing UCDP's 'best' estimates with probability distributions over the true values

- Distribution obtained through an expert elicitation



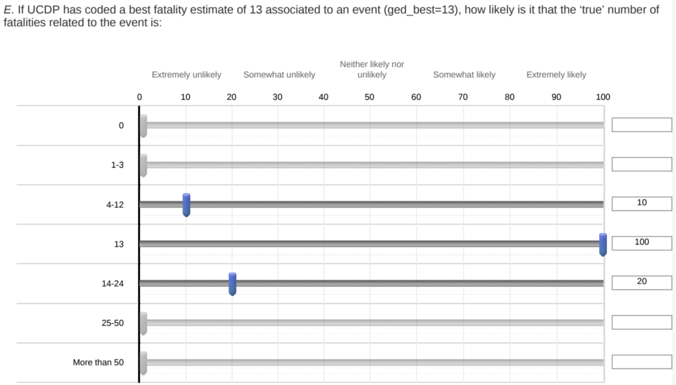
Capturing UCDP coder uncertainty through expert elicitation

Tap into UCDP coders' excellent understanding of reporting uncertainty

- For a selection of coded event types
- Survey to elicit probability distributions across number of fatalities
- According to characteristics such as information situation, number of coded fatalities
- Fit parametric distribution to each survey response
- E.g., lognormal

The UCDP codebook stipulates how to handle imprecise reports

- e.g. *many* $\rightarrow 3$



Constructing a probability distribution of fatalities for new events

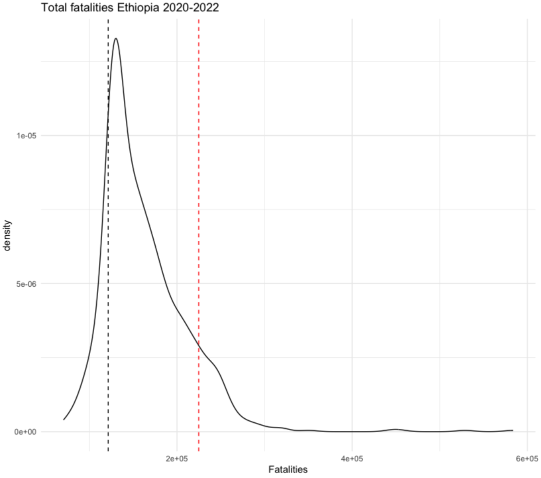
Use the probability distributions for the sample event types to construct an uncertainty model for all events

Possibly:

$$\hat{\theta}_{ij} = \beta ln(u_i) + \alpha_j + \gamma X + \delta Z + \epsilon_{ij}$$

where

- θ_{ij} is a distribution parameter (m_{ij}, σ_{ij})
- $ln(u_i)$ is log of the coded value
- α_j is a fixed term for the coder
- X is a set of metadata such as information context or type of report
- Z is a set of special values (e.g. 3, 25, 100)

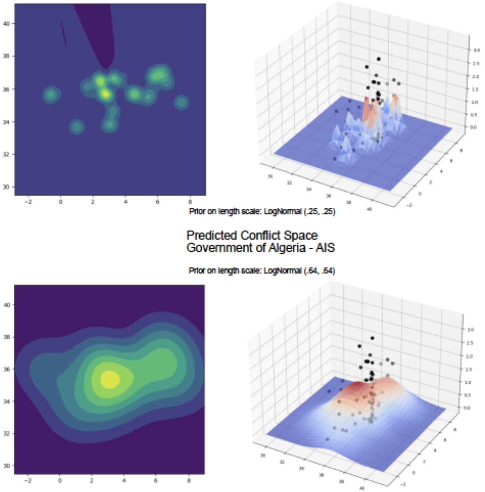


Modeling when and where did violence occur?

Mihai Croicu: 'Known geographic imprecision' – UCDP notes location is imprecise and assigns placeholder location

- Estimate the spatial probability distribution for each conflict
- Making use of actor data
- Randomly draw location based on distribution estimated using Gaussian Process model

Developed at CAS with great advice from Nils!



Nowcasting the UCDP data

Timeline of data collection:

- UCDP delivers monthly 'candidate' data
- In May every year, they publish final data for the preceding year

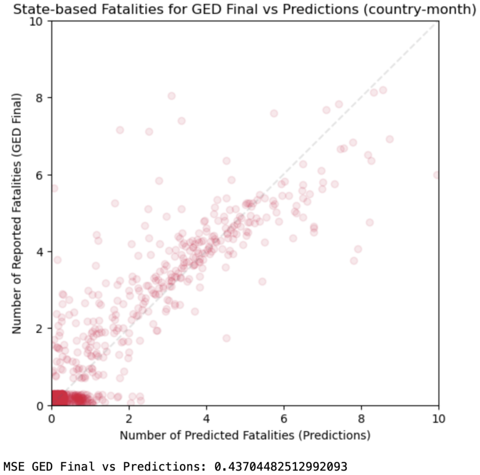
Candidate data are imperfect approximations to final data

- Solution: 'now-cast' final GED data

Current best model:

- A negative binomial country random effects model

Also developed at CAS!



Thanks, and congratulations to Nils!

Contact: hhegre@prio.org, views@pcr.uu.se

Website: https://viewsforecasting.org