Using Latent Variable Models to Estimate the Prevalence of Sexual Violence in Armed Conflict: An Introduction

Jule Krüger,
Center for Political Studies, University of Michigan

Workshop on Empirical and Computational Social Sciences in India

Ashoka University, December 15, 2018

This workshop builds on ongoing research with

Ragnhild Nordås

and

Christopher Fariss

Wartime Sexual Violence

- Includes the use of rape and other forms of sexual violence
- Constitutes a severe human rights problem
- Is difficult to observe and document as a practice

A lack of systematic data impedes empirical analysis with regard to extent, spatiotemporal trends, and patterns.

Why Conflict-Related Sexual Violence Is Hard to Measure

- Shame, fear of retaliation, stigma and rejection due to socio-cultural taboos
- Inconsistency in testimony and lack of clear narrative due to trauma-induced memory loss
- Differing conceptualizations and language used to refer to sexual violence events
- Perpetrators' incentives to conceal activity and evade accountability for war crimes
- Blending of state actors and institutions with regard to the perpetration and reporting of these crimes

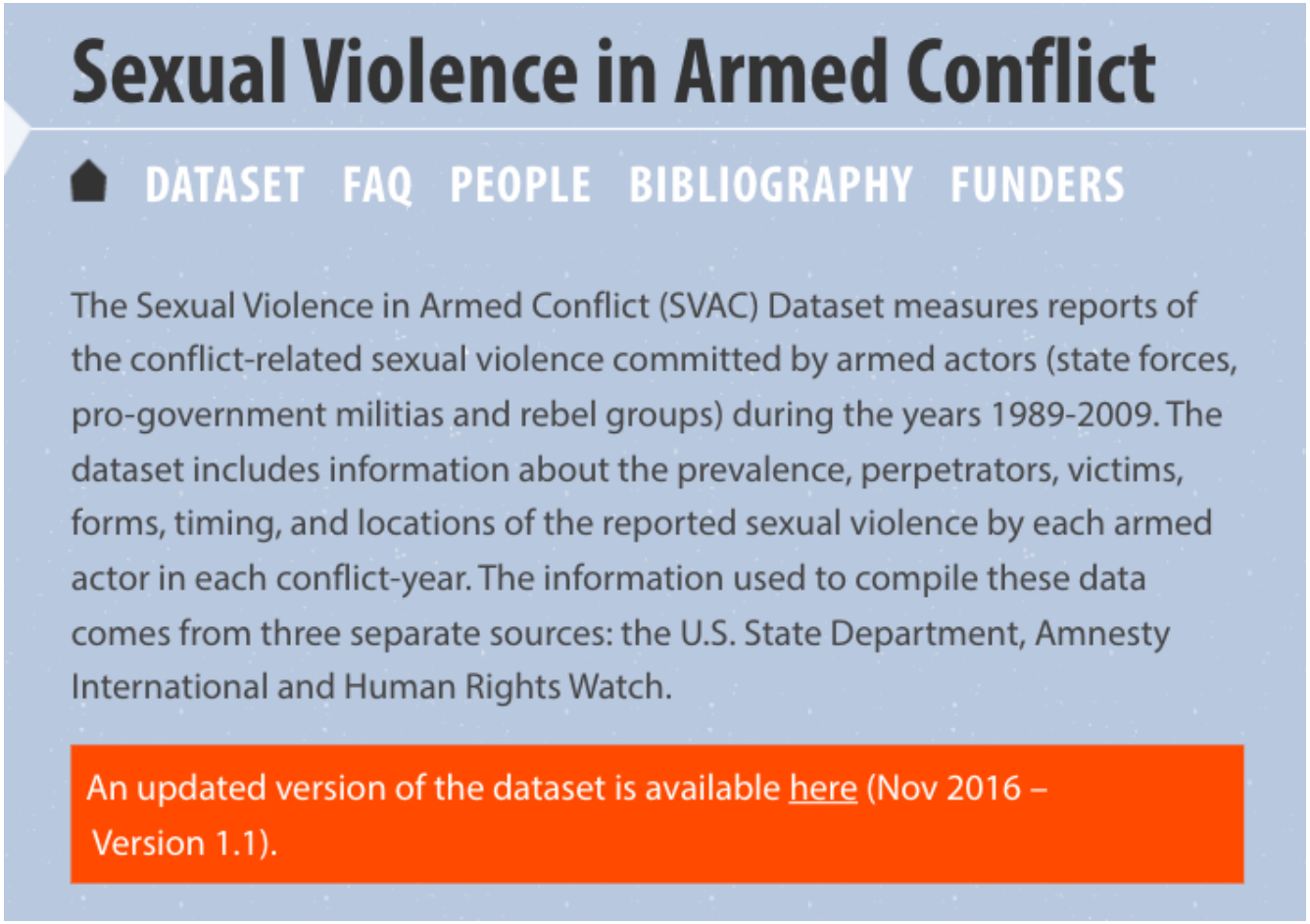
All of these issues vary over space and time.

Why the Observation of Wartime Sexual Violence May Improve over Time

- Increasing international focus
- Changing norms and perceptions of survivors
- Recent challenges to societal taboos
- Growing initiatives to empower survivors to speak out
- Changes in the wording of sexual violence experiences leading to more explicit descriptions
- Growth in documentation efforts paired with improved documentation practices

While these trends vary across space, we will likely see higher reporting rates in some places over time.

How We Currently Measure Wartime Sexual Violence



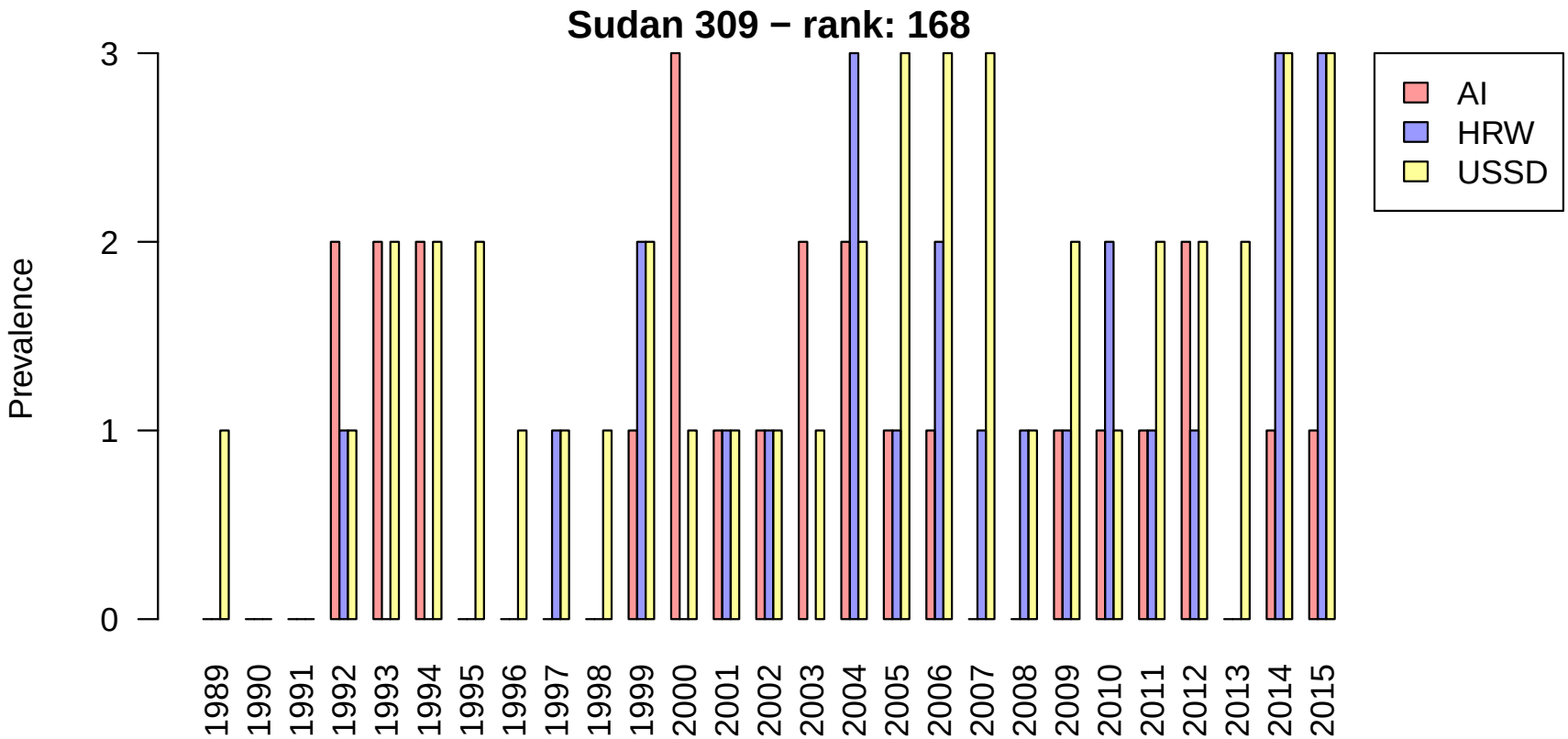
Three ordinal SVAC variables based on human-coded annual human rights reports http://www.sexualviolencedata.org/

Let's import the data into R and take a look at it:

```
$: cd ~/git/SVAC-LVM-tutorial/import
$: open -a Rstudio src/import-check-data-main.R
```

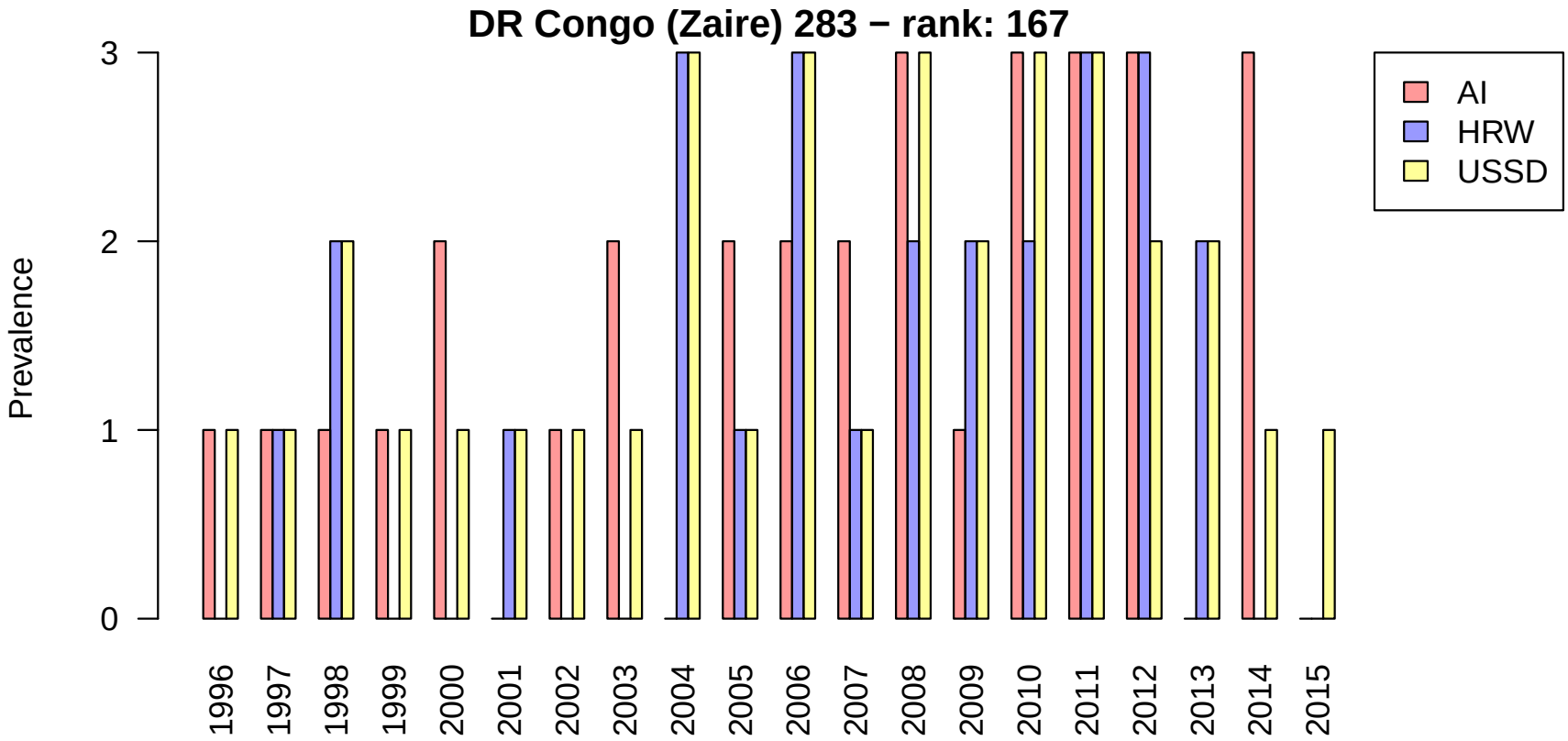
In this tutorial, we will only look at reported SVAC with respect to state forces ('GOV').

SVAC Provides Three Indicators that Report Prevalence of Wartime Sexual Violence



Reported level of engagement in wartime sexual violence by state forces in Sudan according to three sources est

Another Case...



Reported level of engagement in wartime sexual violence by state forces in DRC according to three sources est

Let's make some more barplots of reported state behavior for other armed conflict cases:

```
$: cd ../visualize
$: open -a Rstudio
   src/barplot-sources-by-conflict.R
```

For Every Conflict-Year, We Have Three Sources Reporting on SVAC Prevalence

- For many years, the three prevalence measures diverge (e.g., Sudan 2006, DRC 1998)
- In other years, the three sources seem to agree (e.g., Sudan 2001-2, DRC 2011)
- In a considerable number of years (when looking at the entire dataset), the three sources do not report any SV (e.g., Sudan 1990-1)

How do we deal with converging/diverging information across the three sources? Shall we average across them, or should we choose the most common/lowest/highest level?

What If, for a Given Conflict-Year, We Understand

- True SVAC prevalence as a *latent trait* that can't be observed directly but estimated using observed outcomes
- Available human rights reports as imperfect measures of the latent level due to observational challenges
- Human-coded SVAC variables as imperfect measures of human rights reports due to perceptual coding error
- Information convergence/divergence across sources as a measure of certainty regarding SVAC prevalence

The logic of latent variable models (LVM) follows precisely this conceptual approach to measurement.

The Added Value of Latent Variable Models

- Leverage information from multiple sources
- Provide probabilistic estimates of a latent trait, i.e., SVAC in our case
- Express our uncertainty regarding the estimates of the latent trait through credible intervals
- Compute the estimated latent trait at the interval-level (instead of ordinal), which simplifies subsequent analysis
- Enable direct probabilistic comparisons across conflict-years and cases

Parametrizing a LVM to Estimate SVAC, I

(Cf. Schnakenberg and Fariss (2014:7-10) for details.)

We assume that the observed human rights reports for each conflict-year are functions of a unidimensional latent variable θ that represents the level of SVAC.

For each conflict-year observation, we index conflicts with i and years with t.

For each model, we have three ordinal indicators J with levels 0 (no reports), 1 (some), 2 (several/many), and 3 (massive).

Parametrizing a LVM to Estimate SVAC, II

The observed values of each indicator (or, "item") are denoted as y_{itj} for a given conflict-year and assumed to depend on θ_{it}.

Using these observed values, our goal is to estimate θ_{it}, i.e., the latent SVAC prevalence in conflict i in year t.

For each item (i.e., indicator), we estimate an "item discrimination" parameter β_j and a set of $K_j - 1$ difficulty cut-points $(\alpha_{jk})_{k=1}^{k_j}$. (We will plot these cut-points later on.)

Parametrizing a LVM to Estimate SVAC, III

There is also an error term ε_{itj} for each item, which in our case represents observational challenges and coding errors. We assume that the error terms are independently drawn from a logistic distribution.

The error term expresses the likelihood of our model.

Estimating a Latent Variable Model

Following our parametrization, we can derive a probability distribution for a given response to item j, and a likelihood function for β, α and θ given the data.

If you are interested in the math, please refer to Schnakenberg and Fariss (2014: 7-8). For time constraints, we will limit our tutorial to LVM implementation in R.

Let's Run Some LVMs!

```
$: cd ../estimate
$: open -a Rstudio src/estimate-static-SVAC.R
```