How a Small Non-Profit Human Rights Group Uses R

Megan Price, Ph.D.

benetech
TECHNOLOGY SERVING HUMANITY

BARUG April 10, 2012

1. What We Do
2. How R Helps
   - Statistics
   - Visualization
   - Work Flow
3. Why it Matters

The Benetech Human Rights Program (HRP)

- Benetech - `http://www.benetech.org/`
  - Literacy - Bookshare, Route 66
  - Environment - Miradi, City Options
  - New Projects - Social Coding 4 Good
  - Human Rights
- Martus
  - `https://www.martus.org/`
- Human Rights Data Analysis Group (HRDAG)
  - `https://www.hrdag.org/`

Partners


CDHES
ABA CEELI
Promoting the Rule of Law
international center for TRANSITIONAL JUSTICE
HUMAN RIGHTS WATCH
WITNESS TO TRUTH:
Report of the
Sierra Leone
Truth & Reconciliation Commission
FREEDOM HOUSE
CAMPAIGN for GOOD GOVERNANCE
KOMISYON NASYONAL VERITE AK JISTIS
COMMISSION NATIONALE DE VERITE ET DE JUSTICE
CAVR
TIMOR LESTE
AAAS
TRUTH AND RECONCILIATION COMMISSION
TRC
TRUTH, JUSTICE AND RECONCILIATION
Comisión de la verdad y reconciliación
AHPN
Archivo Histórico de la Policía Nacional
-Guatemala-
FAFG
GUATEMALA
MEMORIA DEL SILENCIO
TZ'INIL NA'TAB'AL
Network for Human Rights Documentation - Burma
ND - Burma
truth reconciliation commission
TRUTH
THE ROAD TO
RECONCILIATION
CIIDH
Centro Internacional para Investigaciones en Derechos Humanos
Documentation & Research
Ensaaf
END IMPUNITY. ACHIEVE JUSTICE.
ਇਨਸਾਫ਼
MONUSCO
United Nations Organization Stabilization Mission in the Democratic Republic of the Congo
INTERNATIONAL CRIMINAL TRIBUNAL FOR THE FORMER YUGOSLAVIA
UNITED NATIONS
Comisión Colombiana de Juristas
EAST · WEST MANAGEMENT INSTITUTE
equitas
The National Security Archive
The George Washington University

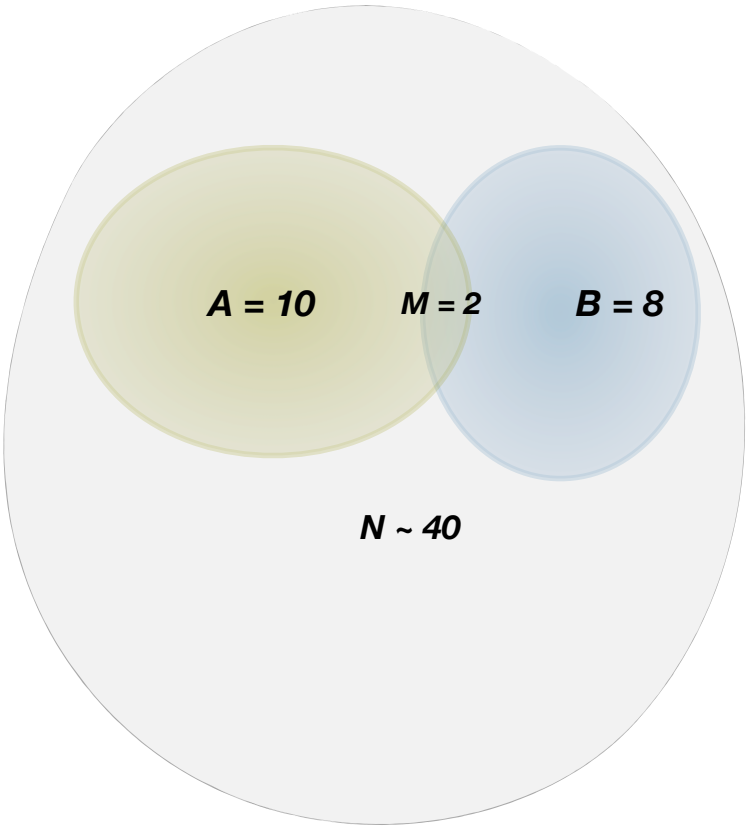
How Many . . .

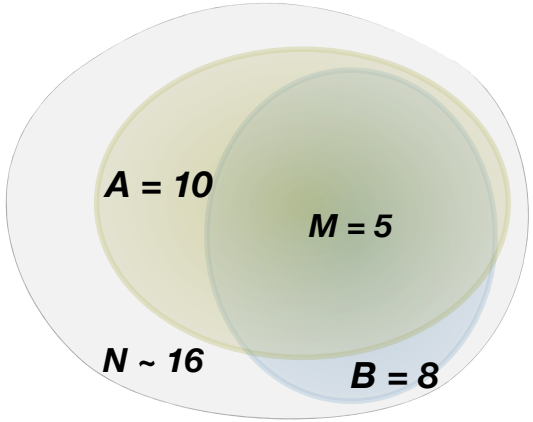
- Conflict-related deaths in Timor Leste between 1974 and 1999?
- Kosovars were killed between March and June 1999?
- Documents with information about disappearances are in the Guatemalan National Police Archive?

Statistics

- Summary statistics
- Regression
- Multiple Systems Estimation (MSE)
- Bayesian Model Averaging (BMA)
- Probabilistic Samples

How Do You Know What You Don't Know?





rcapture

Sophie Baillargeon and Louis-Paul Rivest

- Closed populations
  - Independent lists
  - Time dependent
  - Heterogeneous capture probabilities
  - Behavioral response
  - Chao
  - Darroch
  - Gamma
  - Poisson
- Open populations

rcapture

- Closed populations
  - Time dependent
  - Heterogeneous capture probabilities
  - Poisson

⇒`closedpCI.t(X, dfreq=FALSE, m=c("Mt","Mth"), h=c("Chao","Poisson","Darroch","Gamma"), mX=NULL, alpha=0.05)`

rcapture

Table: Hypothetical Distribution of Records into Numerous Sources

Source A	Source B	Source C	Number of Records
1	0	0	n_{100}
0	1	0	n_{010}
1	1	0	n_{110}
...			
0	0	0	n_{000}

Models

$$n_{000} = \frac{\hat{m}_{111}\hat{m}_{100}\hat{m}_{010}\hat{m}_{001}}{\hat{m}_{110}\hat{m}_{101}\hat{m}_{011}} \tag{1}$$

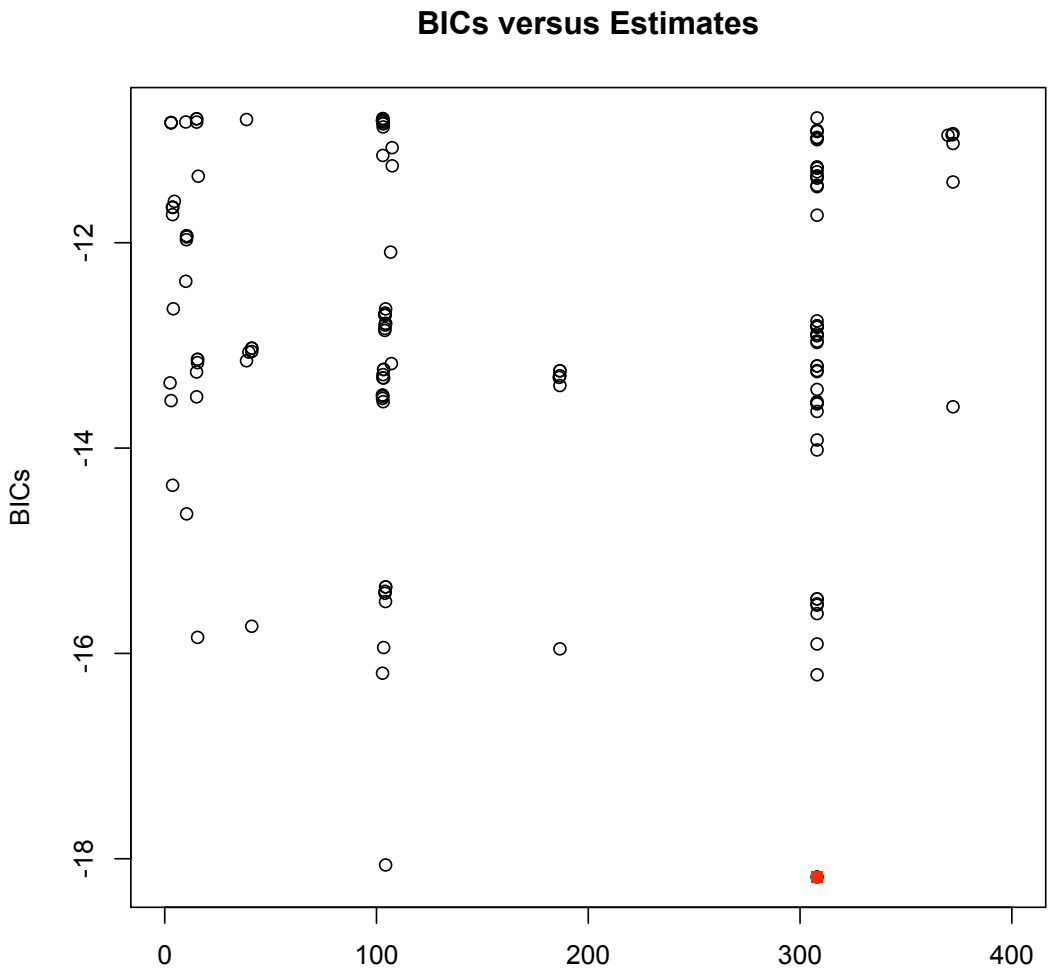
$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} \tag{2}$$

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} \tag{3}$$

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)} \tag{4}$$

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)} + u_{13(ik)} \tag{5}$$

How to Choose Models?



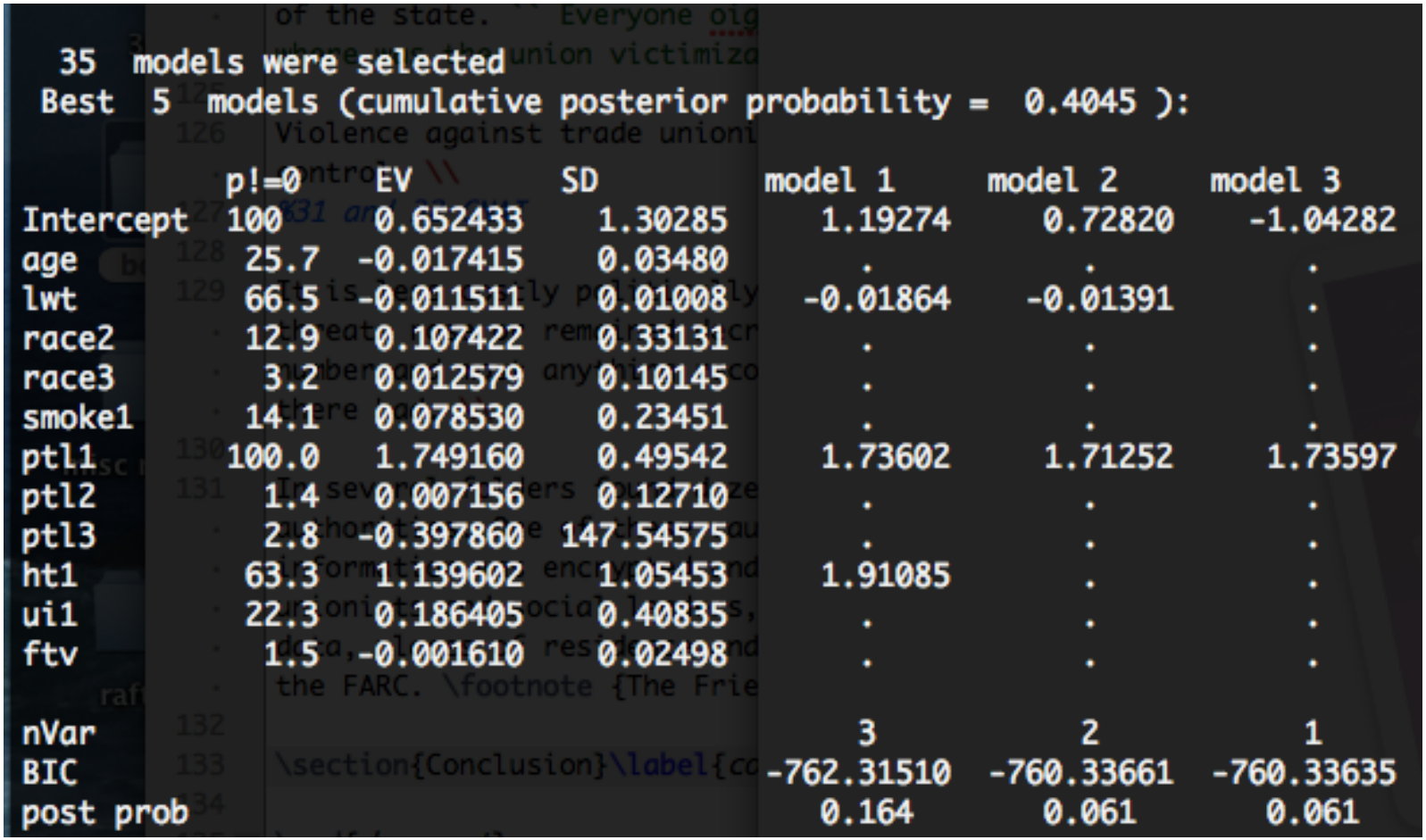
BMA

Adrian Raftery, Jennifer Hoeting, Chris Volinsky, Ian Painter, Ka Yee Yeung

`http://www2.research.att.com/~volinsky/bma.html`

- Provides Bayesian Model Averaging (BMA) for linear models, generalizable linear models, and survival models
- Searches model space using branch-and-bound algorithm (`leaps` package)
- `bic.glm(x, y, glm.family, nbest = 150)`

BMA



```
 35  models were selected
 Best  5  models (cumulative posterior probability =  0.4045 ):

          p!=0    EV         SD          model 1     model 2     model 3
Intercept  100    0.652433     1.30285     1.19274     0.72820    -1.04282
age         25.7  -0.017415    0.03480         .           .           .
lwt         66.5  -0.011511    0.01008    -0.01864    -0.01391         .
race2       12.9   0.107422    0.33131         .           .           .
race3        3.2   0.012579    0.10145         .           .           .
smoke1      14.1   0.078530    0.23451         .           .           .
ptl1       100.0   1.749160    0.49542     1.73602     1.71252     1.73597
ptl2        1.4    0.007156    0.12710         .           .           .
ptl3        2.8   -0.397860  147.54575         .           .           .
ht1        63.3    1.139602    1.05453     1.91085         .           .
ui1        22.3    0.186405    0.40835         .           .           .
ftv         1.5   -0.001610    0.02498         .           .           .

nVar                                            3           2           1
BIC                                      -762.31510  -760.33661  -760.33635
post prob                                   0.164       0.061       0.061
```

Guatemalan National Police Archive - One Big Sampling Problem

Guatemalan National Police Archive - One Big Sampling Problem

survey

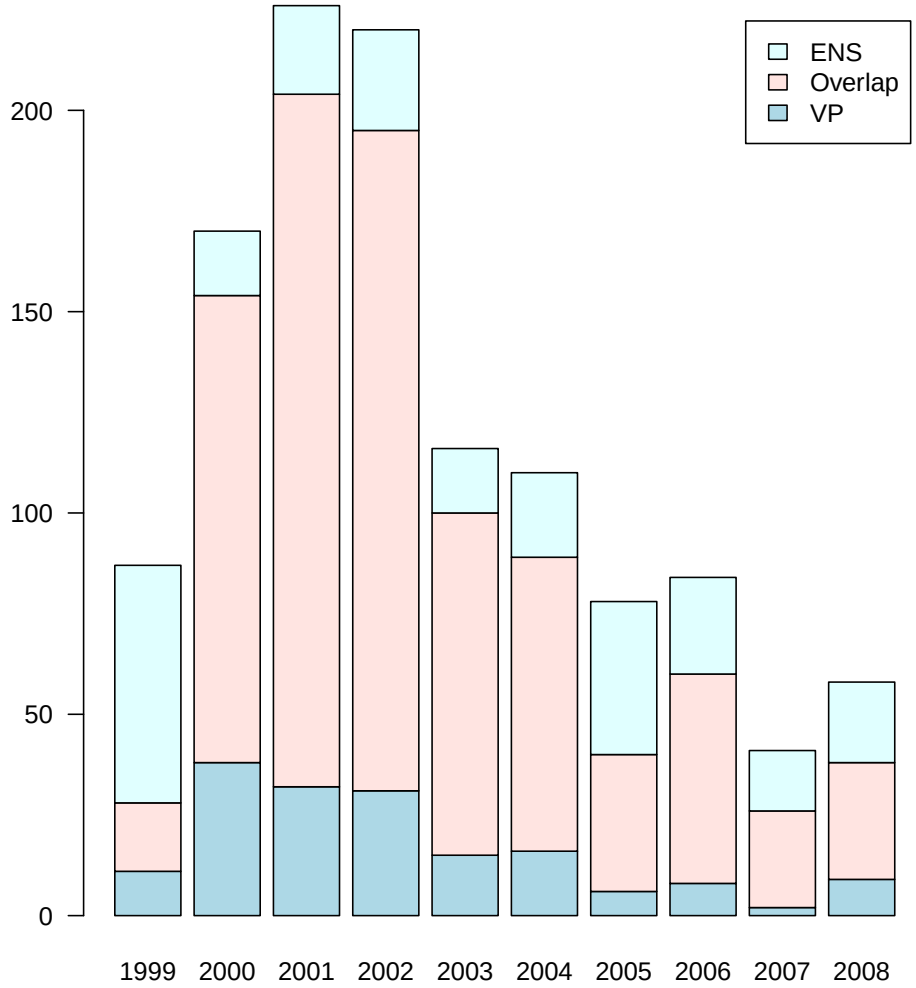
Thomas Lumley - `http://faculty.washington.edu/tlumley/survey/`

- Variety of summary statistics for entire sample or specific subsamples
- Calculates variances using Taylor linearization or replicate weights (BRR, jackknife, bootstrap, multistage bootstrap, or user-supplied)
- Numerous sample designs (multi-stage, with and without replacement, PPS)
- Post-stratification, raking, weight trimming

survey

- svydesign
- subset
- svytotal
- svyratio

Barplots


ENS
Overlap
VP
200
150
100
50
0
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008

Barplots

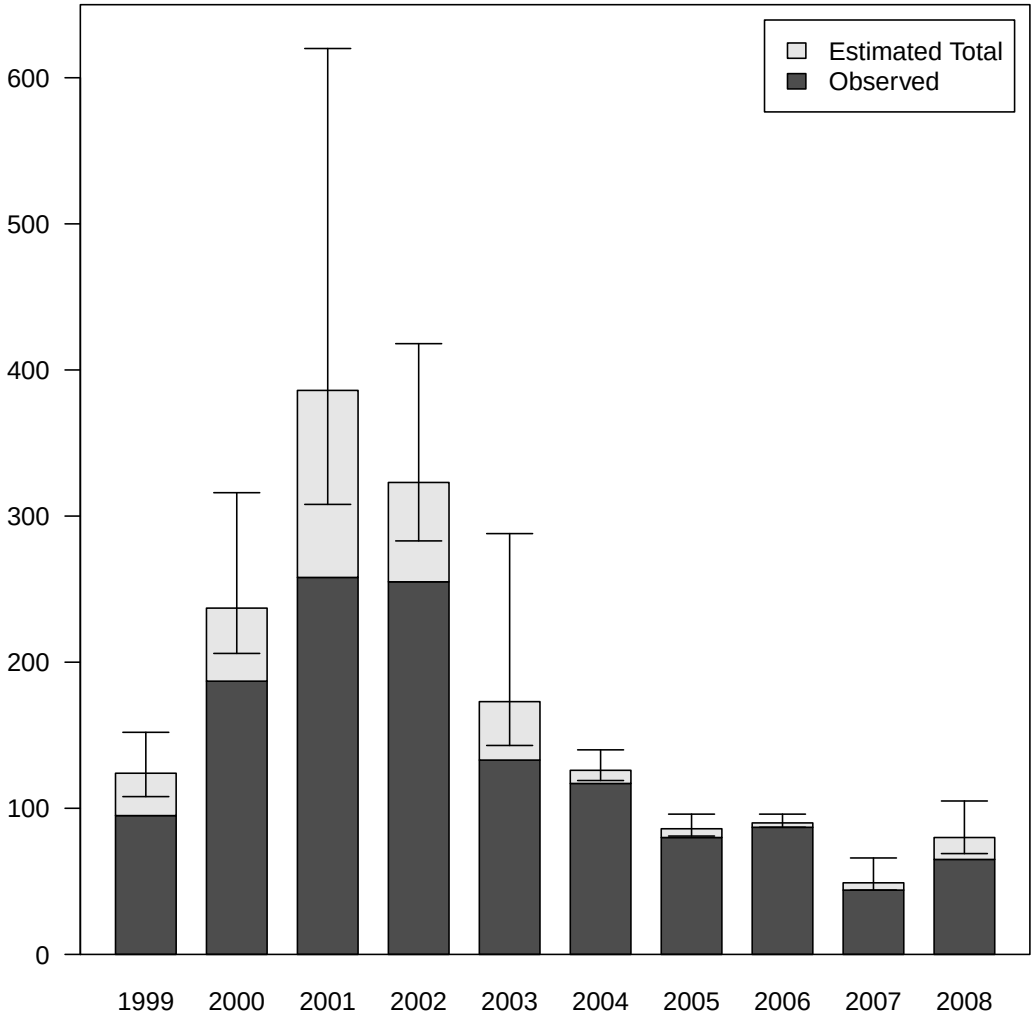
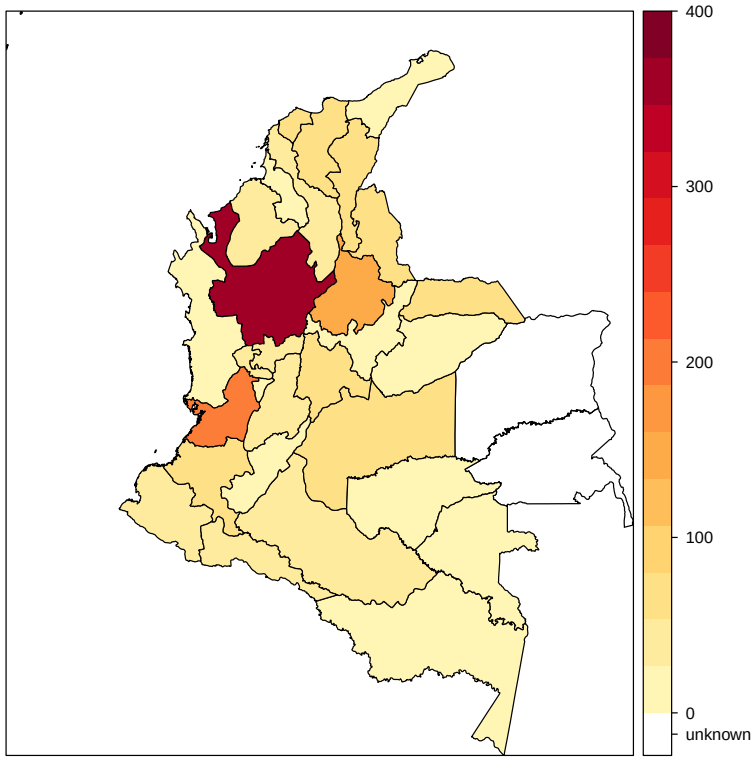
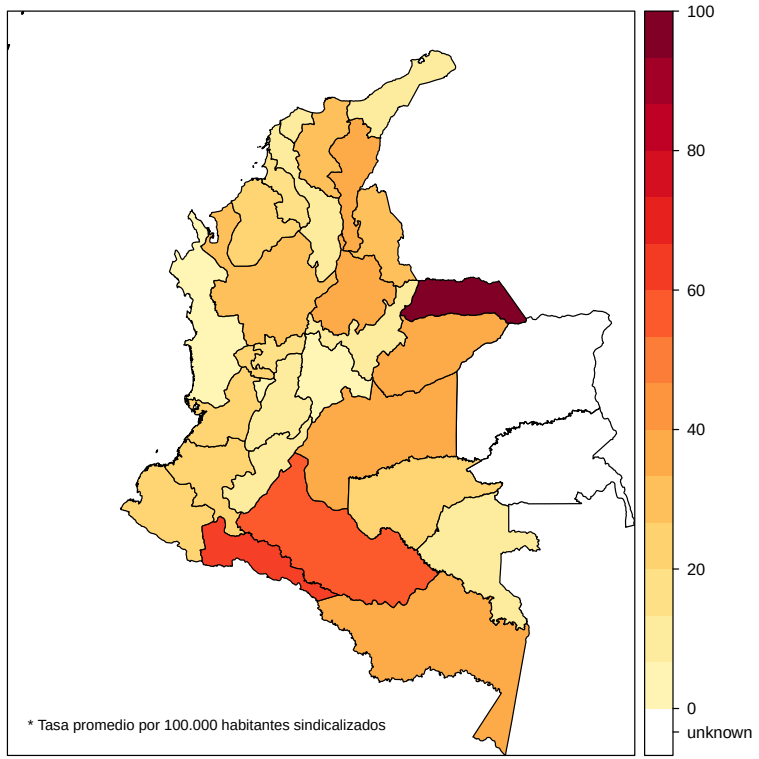

Estimated Total
Observed
600
500
400
300
200
100
0
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008

Figure: Estimates Killings Between 1999 to 2008. Total and Rates





(a) Point Estimates of Trade Unionists Killed by Department

(b) Rates of Estimated Killings by Unionized Population*

Libraries

- maptools
- gpclib
- RColorBrewer

Idealized Workflow

1. Data 'from the sky' (import)
2. 'Processing' (clean, parse, translate, canonicalize)
3. Summary/Descriptive Statistics (individual or compare)
4. Inference (MSE)
5. Outputs/Deliverables (write)

Other Languages

- Sweave
- xtable
- SQL
- YAML

Sweave

```
<<echo=FALSE>>=
   load("input/magic-numbers.Rdata")
@

...

Using this data and our MSE developments, we estimate that
there were between
\Sexpr{yrs_agg_kill_lowF}--\Sexpr{yrs_agg_kill_highF}
killings in Casanare in 2000-2007 and between
\Sexpr{yrs_agg_disp_lowF}--\Sexpr{yrs_agg_disp_highF}
disappearances in 1998-2005.
```

Sweave

Using this data and our MSE developments, we estimate that there were between 3,944–9,983 killings in Casanare 2000-2007 and between 1,270–5,552 disappearances in 1998-2005.

xtable

```
library(xtable)

make_xtable <- function(x, caption_text, table_name) {
    print('str(x) in make.xtable')
    print(str(x))
    x <- xtable(x, caption = caption_text)
    digits(x) <- 3
    print(x, file=sprintf('output/%s.tex', table_name),
        table.placement='H', caption.placement = 'top',
include.rownames=TRUE)
    return(TRUE)
}
```

xtable

```
% latex table generated in R 2.8.0 by xtable 1.5-5 package
\begin{table}[H]
\begin{center}
\caption{Proportion of Documents Authored by Each Group}
\begin{tabular}{rrr}
  \hline
 & Proportion & SE \\
  \hline
  gaid & 0.016 & 0.008 \\
  dept & 0.240 & 0.075 \\
  allpn & 0.766 & 0.030 \\
   \hline
\end{tabular}
\end{center}
\end{table}
```

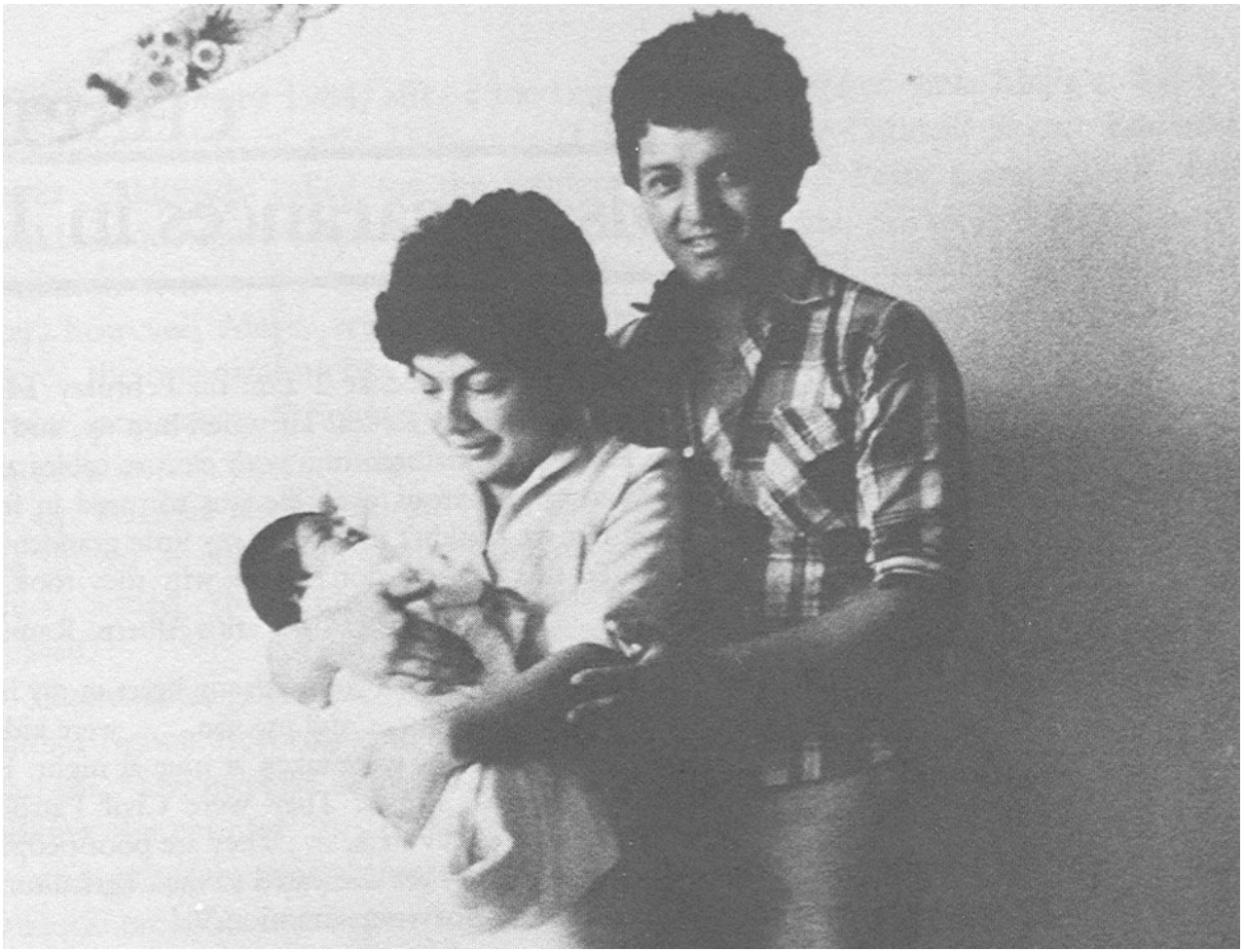
How Many . . .

- Conflict-related deaths in Timor Leste between 1974 and 1999?
  - 102,800 (+/- 12,000)
- Kosovars were killed between March and June 1999?
  - 10,356 (9,002, 12,122)
- Documents with information about disappearances are in the Guatemalan National Police Archive?
  - 414,542 (SE = 92,599)

Why it Matters

- Conflict-related deaths in Timor Leste between 1974 and 1999?
  - Historical Narrative
- Kosovars were killed between March and June 1999?
  - International Criminal Tribunal for the former Yugoslavia
- Documents with information about disappearances are in the Guatemalan National Police Archive?
  - Edgar Fernando García

Edgar Fernando García



Edgar Fernando García

Thank You!

meganp@benetech.org

https://www.hrdag.org/

Lincoln-Petersen Estimate

$$\widehat{Pr}(A) = \frac{|A|}{N}$$

$$\widehat{Pr}(B) = \frac{|B|}{N}$$

$$\widehat{Pr}(M) = \frac{|M|}{N}$$

$$\frac{|M|}{N} = \widehat{Pr}(M) = \widehat{Pr}(A \cap B) = \widehat{Pr}(A)\widehat{Pr}(B) = \frac{|A||B|}{N^2}$$

$$\frac{|M|}{N} = \frac{|A||B|}{N^2}$$

$$\widehat{N} = \frac{|A||B|}{|M|}$$