How a Small Non-Profit Human Rights Group Uses R

Megan Price, Ph.D.

Human Rights Data Analysis Group
everybody counts.

BARUG March 17, 2014

Partners


CUHES
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
CAVR
TIMOR LESTE
CAMPAIGN for GOOD GOVERNANCE
KOMISYON NASYONAL VERITE AK JISTIS
COMMISSION NATIONALE DE VERITE ET DE JUSTICE
AAAS
TRUTH AND RECONCILIATION COMMISSION
TRC
TRUTH, JUSTICE AND RECONCILIATION
Comisión de la verdad y reconciliación
GUATEMALA MEMORIA DEL SILENCIO
TZ'INIL NA'TAB'AL
AHPN
Archivo Histórico de la Policía Nacional
-Guatemala-
FAFG
Network for Human Rights Documentation - Burma
ND - Burma
CIIDH
Centro Internacional para Investigaciones en Derechos Humanos
truth reconciliation commission
TRUTH
THE ROAD TO
RECONCILIATION
Documentation & Research
Ensaaf
END IMPUNITY. ACHIEVE JUSTICE.
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-killings in Syria?
- Kosovars were killed between March and June 1999?
- Documents with information about disappearances are in the Guatemalan National Police Archive?

Comparisons

- Did violence in Homs peak in Feb, March, or April 2012?
- Were more Indigenous or Non-Indigenous people killed in Guatemala?
- Do union members suffer more violence in Colombia than the general population?

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)


Arabic
فاطمة كريم
2011 - 1985
دوما
VDC
مركز توثيق الانتهاكات في سوريا
Violations Documentation Center in Syria
Home | About | Join our mailing list | Contact us | Reports and testimonies | Martyrs | Detainees | Missing | Regime fatalities
Filter by
Province Sex Status Display Start Date End Date This Date
Show All Show All Show All Show All 2011-03- 2013-11-
Cause of Death
Explosion Shelling Field Execution Shooting Kidnapping - Execution Kidnapping - Torture Kidnapping - Torture - Execution
Detention - Execution Detention - Torture Detention - Torture - Execution Un-allowed to seek Medical help Warplane shelling Other
Chemical and toxic gases
Name Family Status Area Occupation Notes Martyrdom location Age Rank
Submit
Latest Martyrs (80187)
Beginning 2 3 4 5 6 ..... End
Name | Status | Sex | Province | Area | Date of death | Cause of Death
Wafa Yones Darwish | Civilian | Adult - Male | Damascus Suburbs | Zabadany: Kfer Yabus Village | 2013-11-30 | Field Execution
Abo Omer Bibers | Non-Civilian | Adult - Male | Damascus | Midan | 2013-11-30 | Shooting

Other Languages

- python
- YAML
- SQL
- LaTeX, MultiMarkdown

Sweave or knitr

```
<<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 or knitr

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.

Statistics

- Summary statistics
- Multiple Systems Estimation (MSE)
- Regression
- Probabilistic Samples

data.table

Matthew Dowle

http://datatable.r-forge.r-project.org/

- Fast aggregation of large data
- Fast ordered joins
- Fast add/modify/delete of columns by group

data.table

```
org <- data.table(org)

only.SOHR <- org[,list(fatal.sohr=sum(onlySOHR)),
by=list(year, governorate)]

not.SOHR <- org[,list(fatal.other=sum(notSOHR)),
by=list(year, governorate)]
```

`rcapture` (Multiple Systems Estimation)

Sophie Baillargeon and Louis-Paul Rivest

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

rcapture (Multiple Systems Estimation)

- 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)`

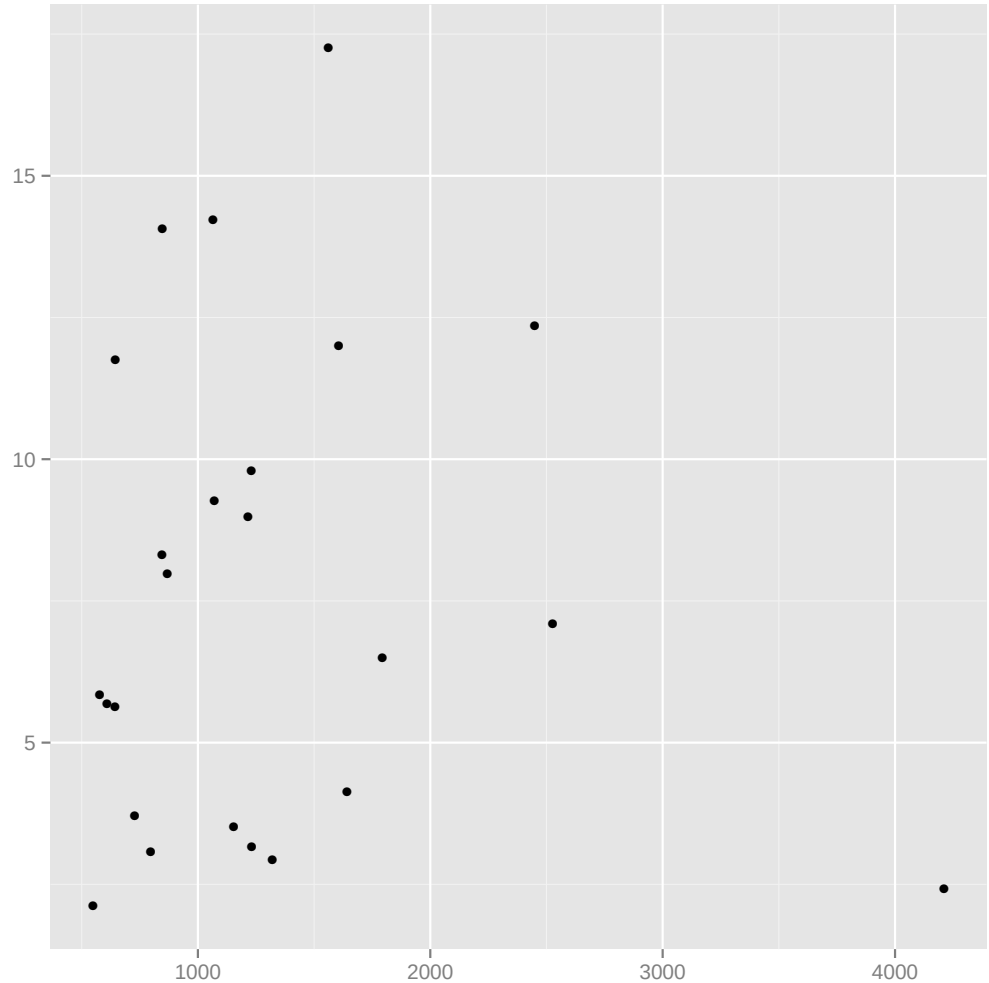
`rcapture` (Multiple Systems Estimation)

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}

How to Choose Models?

(note: internal diagnostic of preliminary results; x-axis is estimate, y-axis is -BIC)



Guatemalan National Police Archive - One Big Sampling Problem

Guatemalan National Police Archive - One Big Sampling Problem

survey (Probabilistic Samples)

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

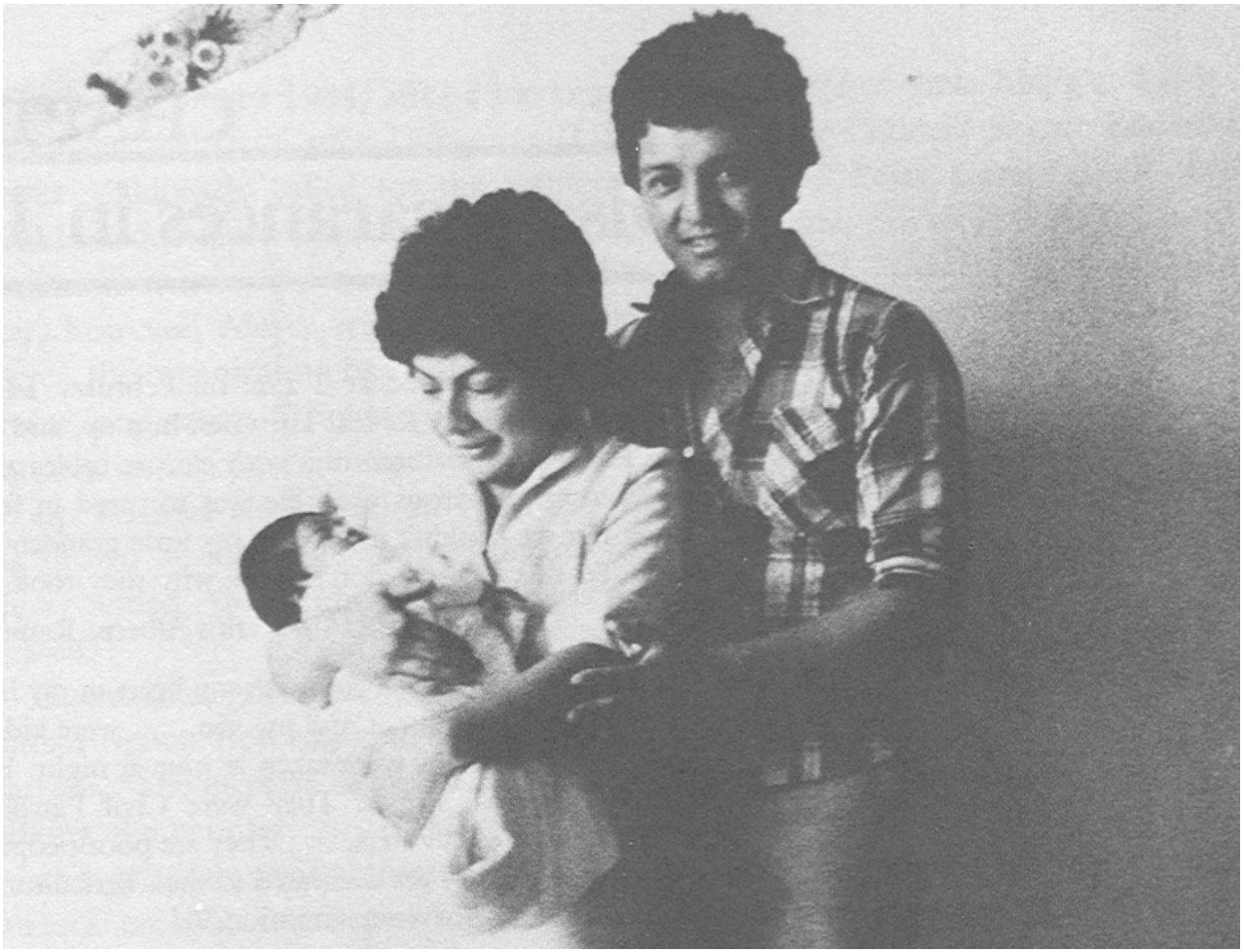
survey (Probabilistic Samples)

- `svydesign`
- `subset`
- `svytotal`
- `svyratio`

How Many . . .

- Conflict-killings in Syria?
  - ??
- 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)

Edgar Fernando García



Edgar Fernando García

Thank You!

meganp@hrdag.org
https://www.hrdag.org/
@hrdag