quantil | matemáticas aplicadas

Datos de la Comisión de la Verdad: Estimación del subregistro de víctimas

Septiembre 2023

Human Rights Data Analysis Group

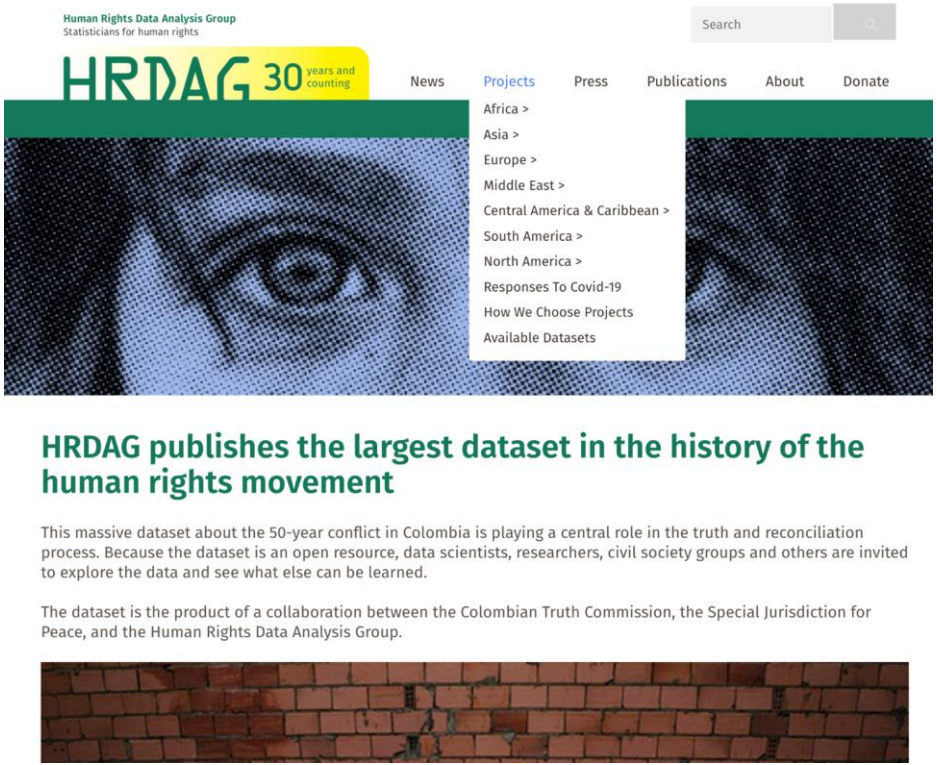
JEP JURISDICCIÓN ESPECIAL PARA LA PAZ · COMISIÓN DE LA VERDAD · HRDAG Statisticians for human rights

Informe metodológico del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística*

18 de agosto de 2022**

Acerca de HRDAG

- Organización sin ánimo de lucro que usa la estadística para el estudio de violaciones a derechos humanos.
- Desde su fundación en 1991, ha participado en 10 comisiones de la verdad.



Fuente: https://hrdag.org/

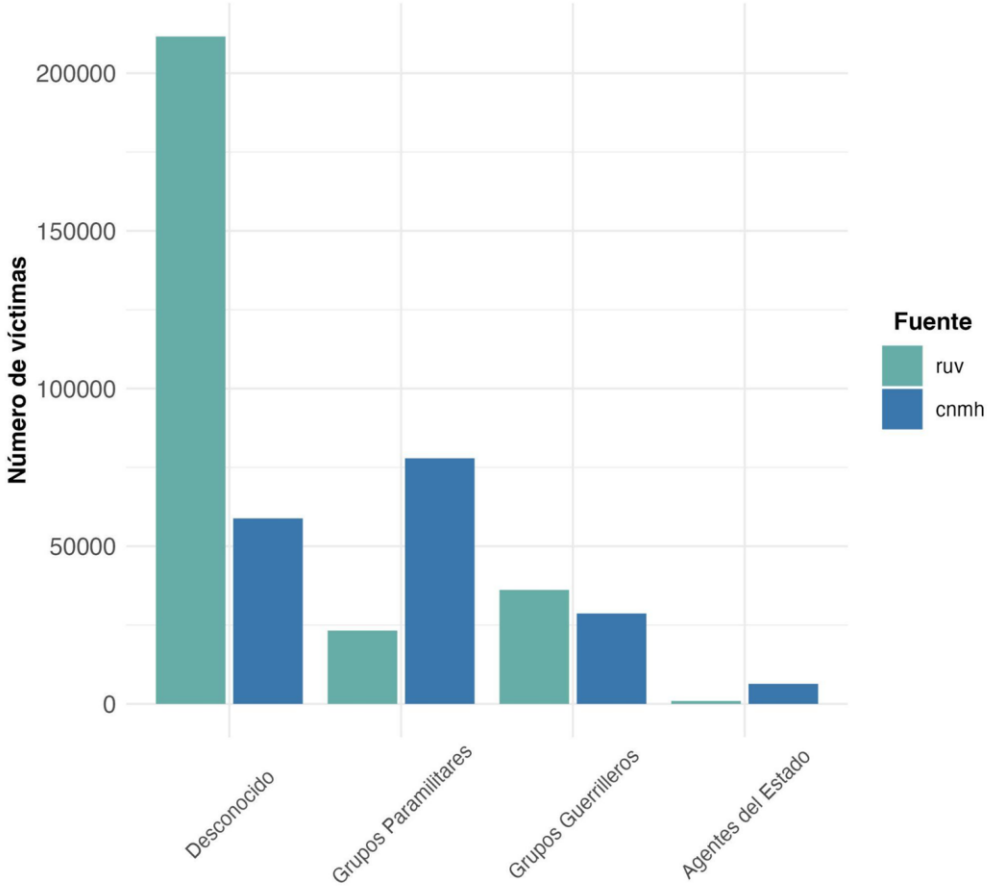
HRDAG
Statisticians for human rights

Motivación

- Toda base de datos muestra solo una imagen del mundo.
- Cada una de ellas tiene limitaciones.
- Podemos saber sobre los patrones de documentación, pero, ¿qué pasa con los **patrones de violencia**?

HRDAG
Statisticians
for human rights

Víctimas de homicidio por fuente



Fuente: RUV y CNMH

HRDAG
Statisticians
for human rights

Acerca del proyecto conjunto CEV-JEP-HRDAG

Objetivos

- Hacer análisis estadísticos de los patrones de violencia dentro del conflicto armado.
- Tener en cuenta el sesgo y la varianza.
- Considerar y abordar la presencia de datos faltantes.
- 4 violaciones a derechos humanos: **homicidio, secuestro, reclutamiento y desaparición**.

HRDAG
Statisticians
for human rights

Acerca del proyecto conjunto CEV-JEP-HRDAG

Insumos

- **112** bases de datos.
- Provenientes de **44** fuentes.
- Bases de datos **especializadas** en conflicto armado o **generales.**

HRDAG
Statisticians
for human rights

Objetivo del seminario

- Ofrecer un panorama del proyecto con un enfoque en:
  - El proceso de deduplicación.
  - El proceso de imputación estadística.
  - El proceso de estimación del subregistro.

- Ofrecer un panorama de los datos resultantes y el paquete de R para su correcto uso.

Publicación de datos CEV-JEP-HRDAG:
https://microdatos.dane.gov.co/index.php/catalog/795

HRDAG
Statisticians
for human rights

Deduplicación o vinculación de registros

HRDAG
Statisticians
for human rights

JEP
JURISDICCIÓN ESPECIAL PARA LA PAZ
COMISIÓN DE LA VERDAD
HRDAG
Statisticians
for human rights

Informe metodológico del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística*

18 de agosto de 2022**

Vinculación de registros

- Hemos juntado más de 100 bases de datos
- Una víctima puede estar **registrada varias veces** en diferentes bases de datos, o en la misma.
- Puede haber pequeñas diferencias en el registro.

HRDAG
Statisticians
for human rights

Id registro	Nombre	Apellido	Año hecho	Etnia
1	Nicole	Gonzalez	1999	¿?
2	Nicolle	Gonzalez	1998	MESTIZO
3	Nicol	Gonzales	1999	MESTIZO
4	Maria	Sanchez	2011	INDIGENA
5	Marina	Sanches	2011	INDIGENA
6	Mariana	Sanchez	2011	¿?
7	Jose	Pineda	1990	¿?
8	Josue	Pineda	1990	¿?
9	Carlos	Alvarez	2015	MESTIZO

Vinculación de registros

- Encontrar todos los registros que corresponden a la misma víctima: **deduplicación**

- El proceso de deduplicación:

    1. Modelo de bloques
    2. Modelo de clasificación
    3. Modelo de agrupamiento

HRDAG
Statisticians
for human rights

Vinculación de registros

Modelo de bloques

- **Analizar cada par de registros** para ver si son la misma persona.
- Casi 13 millones de registros: **trillones de pares**.
- El objetivo de este paso es **reducir el problema** pero abarcando **todos los posibles registros que se refieran a la misma persona**.
- Se seleccionan **grupos de registros** que comparten algunas características para comparar dentro del grupo.

HRDAG
Statisticians
for human rights

Id registro	Nombre	Apellido	Año hecho	Etnia
1	Nicole	Gonzalez	1999	¿?
2	Nicolle	Gonzalez	1998	MESTIZO
3	Nicol	Gonzales	1999	MESTIZO
4	Maria	Sanchez	2011	INDIGENA
5	Marina	Sanches	2011	INDIGENA
6	Mariana	Sanchez	2011	¿?
7	Jose	Pineda	1990	¿?
8	Josue	Pineda	1990	¿?
9	Carlos	Alvarez	2015	MESTIZO

Id registro	Nombre	Apellido	Año hecho	Etnia
1	Nicole	Gonzalez	1999	¿?
2	Nicolle	Gonzalez	1998	MESTIZO
3	Nicol	Gonzales	1999	MESTIZO
4	Maria	Sanchez	2011	INDIGENA
5	Marina	Sanches	2011	INDIGENA
6	Mariana	Sanchez	2011	¿?
7	Jose	Pineda	1990	¿?
8	Josue	Pineda	1990	¿?
9	Carlos	Alvarez	2015	MESTIZO

Vinculación de registros

Modelo de clasificación

- En bloques más pequeños: identificar registros correferentes.
- **Medidas de comparación**: nos permiten estimar la probabilidad de que un par de registros se refiera a la misma persona.
- Datos de entrenamiento: el modelo aprende las **características que mejor predicen** etiquetas positivas y negativas del oráculo.
- La combinación de varias medidas da como resultado un ***score* de 0 a 1**: la ponderación de que sean la misma persona.

HRDAG
Statisticians
for human rights

Vinculación de registros

Modelo de clasificación

- Paso 1: La agrupación de registros da como resultado un **grupo de expansión**
- Paso 2: Para separar las agrupaciones se usa la distancia entre registros (1-*p*)
    - Si el modelo indica un puntaje de 0.8 entre pares, la distancia entre ellos es de 0.2

Vinculación de registros

Modelo de agrupamiento

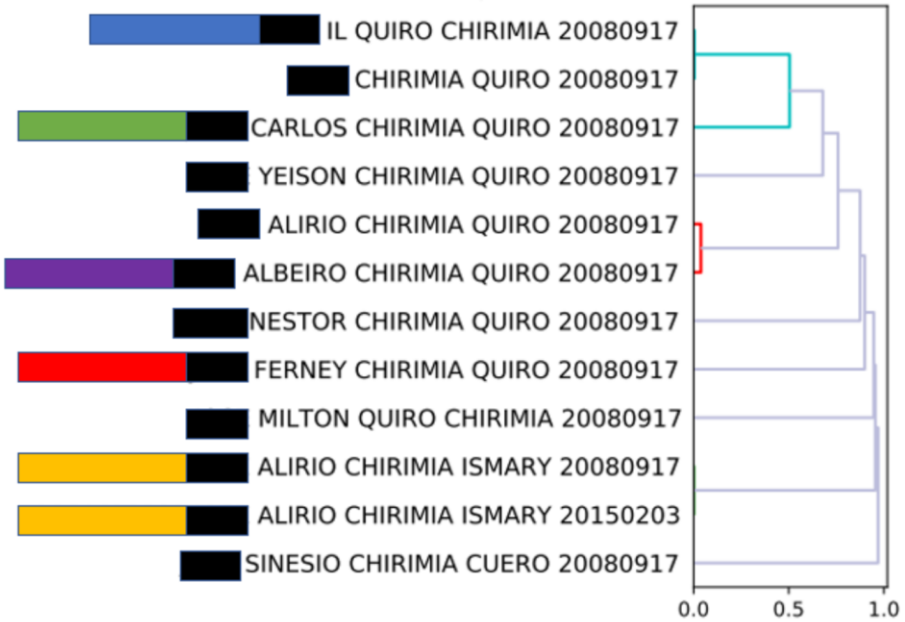


Figura 8: Dendrograma

HRDAG
Statisticians
for human rights

Vinculación de registros

Resultados del modelo

Medida	Conteo
Falso Negativo	714
Falso Positivo	962
Verdadero Negativo	1.716.239
Verdadero Positivo	1.081.765

HRDAG
Statisticians
for human rights

Vinculación de registros

Resultados de la deduplicación

- Comenzamos con **12.863.977** registros.
- Concluimos con **8.775.884** víctimas únicas.

HRDAG
Statisticians
for human rights

Imputación estadística de campos faltantes

HRDAG
Statisticians
for human rights

JEP
JURISDICCIÓN ESPECIAL PARA LA PAZ
COMISIÓN DE LA VERDAD
HRDAG
Statisticians
for human rights

Informe metodológico del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística*

18 de agosto de 2022**

Campos faltantes:

Hecho de violencia	Edad	Etnia	Responsable	Sexo
Desaparición	0,06	**0,59**	**0,7**	0
Homicidio	0,1	0,43	0,66	0,04
Reclutamiento	0,1	0,39	0,23	**0,19**
Secuestro	**0,23**	0,44	0,33	0,01

a) Esta tabla incluye los registros de todas las bases, sin limitarse al conflicto armado.
b) En todas las variables a imputar hay campos faltantes. Los dos decimales pueden ser aproximados a cero.

Fuente: Informe metodológico - HRDAG (2023)

No tenemos (ni tendremos) **certeza** sobre la característica real de la víctima…

El campo es **faltante**, así que tenemos **incertidumbre** en lo que debe ser el **valor verdadero**

HRDAG
Statisticians for human rights

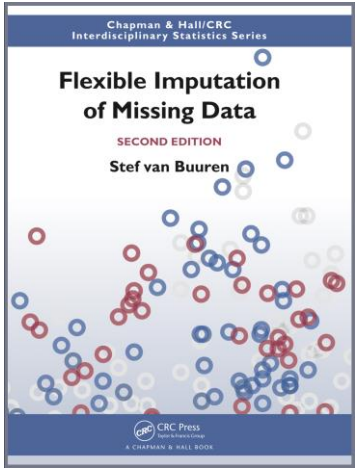
Campos faltantes - Imputación múltiple

- ***MAR: Missing at Random* - Faltantes Aleatorios:** Una hipótesis razonable sobre valores faltantes de una variable es que estos siguen un patrón similar a los valores observados de la variable condicionada al resto de variables. No *MCAR*, no *MNAR*.

- Usar **la información** de lo que sí observamos para predecir o estimar un valor probable en un campo faltante: ***Especificación completamente condicionada.***

- Método: *Predictive mean matching* del paquete `mice` en R.

- Variables de la base y las **variables de soporte.**



HRDAG
Statisticians
for human rights

Campos faltantes - Variables de soporte

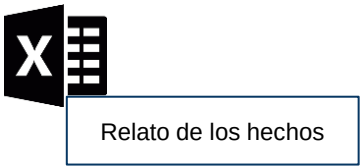
- Características necesarias:

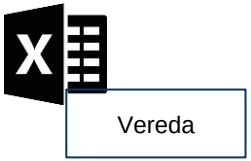
  - Están fuertemente correlacionadas con las variables a imputar.
  - Sin valores faltantes - creadas para cada registro.

Después de estandarizar, se entrena un **modelo tipo memoria a largo-corto plazo** (LSTM por sus siglas en inglés) utilizando `Tensorflow` y `Keras`.

- Utilizando la información de los registros que no faltan este campo.

Resultado es un ***score***. Registros que tienen un *score* parecido en varios campos probablemente son parecidos, así que esta información puede dar pistas adicionales al modelo de la imputación.





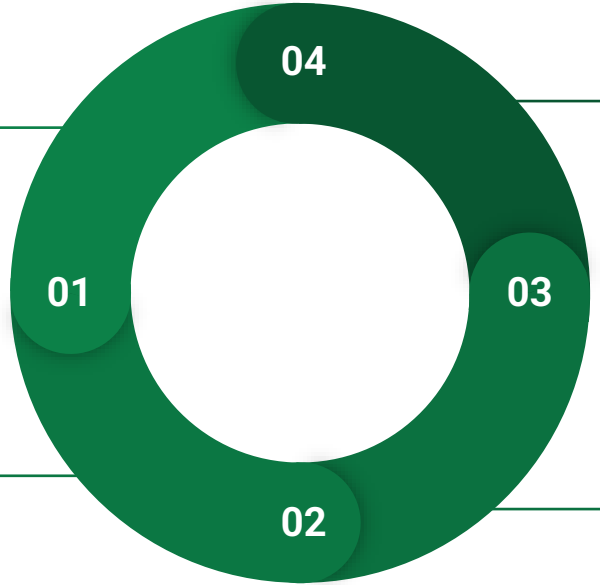
HRDAG
Statisticians
for human rights

Registros en su lema y limpios como cadenas de caracteres

Palabras del registro	is_conflict
economia informal cocinero civil placer hormiga sicariato grupo paramilitar autodefensa unir colombia auc domingo agostar esposar salir buscar plata remesar finca cultivo cocar trabajar cocinar espaciar mesar trabajar matar epoca sujetar trabajador vocero autodefensa amenazar delatar hurtar veneno tipo esperar esposar apunalar llevar hormiga morir mirar dario vestir camuflar matar querer deshacer comandante bajo sujetar masacrar	1
hombre etapa grupo personar identificar miembro frente farc portar armar cortar alcanzar sacar personar vivienda llevar potrero mano atar interrogar disparar paramilitar terminar vida concursar delito personar bien protegido personar protegido calda manizales pensilvania calda samaria direccion especializar violacion derecho humano bogota fiscalia frente cuarenta blocar	NA
conductor inml violento homicidio bello suizo informacion delincuencia comun soltero necropsia escena vehicular mestizar rango armablanca proyectil de arma de fuego violento transitar fechahecho inml	0
hombre informar callar carrera barrer habian herir personar causar impacto armar fuego personar hecho residenciar esquinera funcionar almacen venta repuesto establecimiento comercial razon social auto centrar espiral hallar elementar material probatorio evidenciar fisica lago hematico diligenciar fijar ceder hallar cuerpo vida personar mencionar posicion cubito dorsal camilla metalica cubrir orificio cadena custodio informar patrullar rspondiente victimario vestir chaqueta oscuro portar casco blanco movilizar motocicleta eco deluxe color egra cometidfo hecho huir rumbo desnocido	NA

Campos faltantes - Variables de soporte

Entrenar una variable que indica **si esta idea es cierta para esta variable.** Una variable de soporte para cada categoría.



75% entrenamiento, 25% de prueba.

Seleccionamos registros que tenían valores conocidos (cuadros azules) y tomar aleatoriamente registros positivos y negativos.

Aplicamos el modelo a los datos que tienen NA's. **La expectativa es que esta variable (su *score*) sea cercano al valor observado.**

HRDAG
Statisticians for human rights

Variables de soporte - Coeficiente de correlación de Pearson con los datos de prueba

conflict = 0.994
farc = 0.992
forced_dis = 0.893
indigena = 0.999
masc = 0.991
narp = 0.994
nino = 0.975
para = 0.993

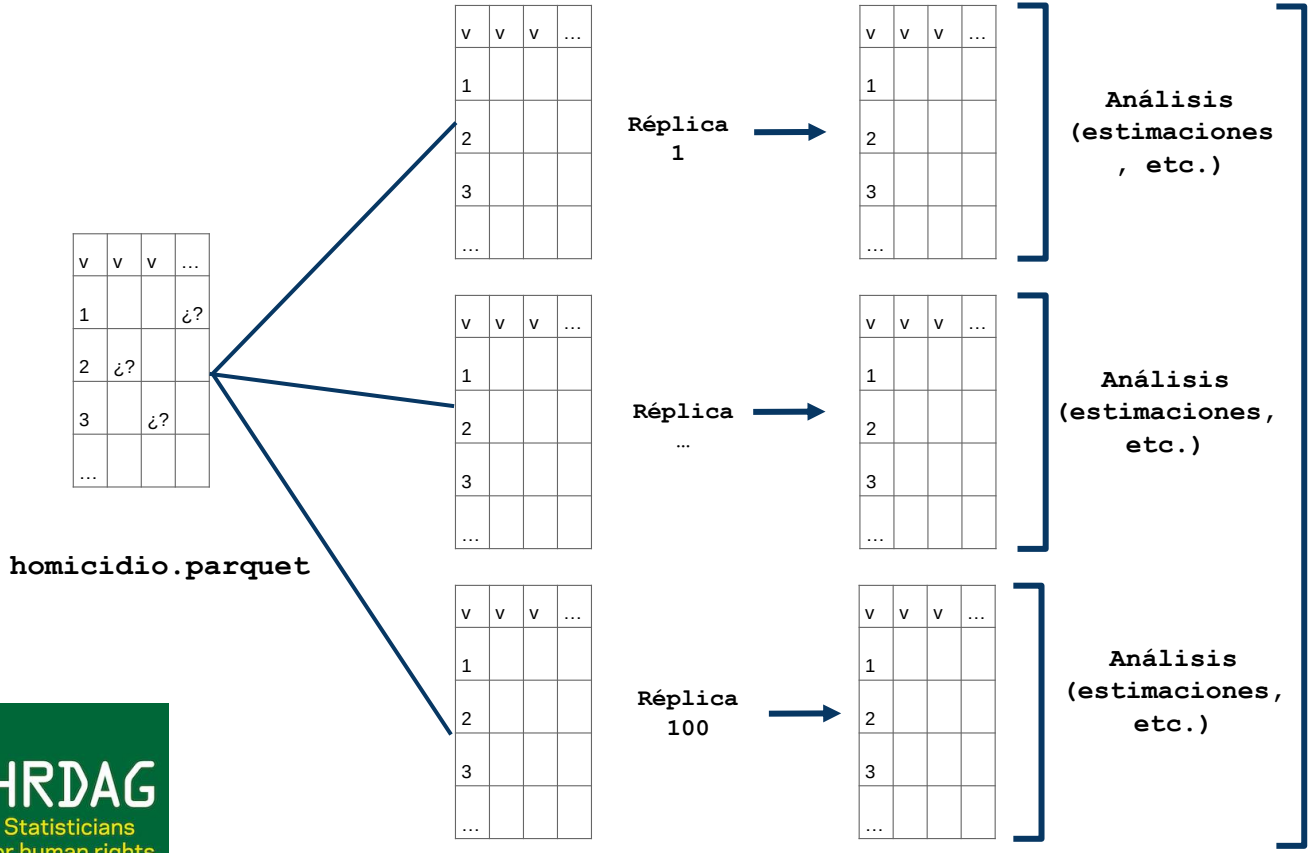
Alta correlación de las variables de soporte con la información conocida que no faltaba en los valores que estábamos prediciendo y un AUC de más del 93% de precisión.

HRDAG
Statisticians
for human rights

Campos faltantes - Variables de soporte

Palabras del registro	is_conflict	Soporte
economia informal cocinero civil placer hormiga sicariato grupo paramilitar autodefensa unir colombia auc domingo agostar esposar salir buscar plata remesar finca cultivo cocar trabajar cocinar espaciar mesar trabajar matar epoca sujetar trabajador vocero autodefensa amenazar delatar hurtar veneno tipo esperar esposar apunalar llevar hormiga morir mirar dario vestir camuflar matar querer deshacer comandante bajo sujetar masacrar	1	0,99
hombre etapa grupo personar identificar miembro frente farc portar armar cortar alcanzar sacar personar vivienda llevar potrero mano atar interrogar disparar paramilitar terminar vida concursar delito personar bien protegido personar protegido calda manizales pensilvania calda samaria direccion especializar violacion derecho humano bogota fiscalia frente cuarenta blocar	NA	0,99
conductor inml violento homicidio bello suizo informacion delincuencia comun soltero necropsia escena vehicular mestizar rango armablanca proyectil de arma de fuego violento transitar fechahecho inml	0	0,001
hombre informar callar carrera barrer habian herir personar causar impacto armar fuego personar hecho residenciar esquinera funcionar almacen venta repuesto establecimiento comercial razon social auto centrar espiral hallar elementar material probatorio evidenciar fisica lago hematico diligenciar fijar ceder hallar cuerpo vida personar mencionar posicion cubito dorsal camilla metalica cubrir orificio cadena custodio informar patrullar rspondiente victimario vestir chaqueta oscuro portar casco blanco movilizar motocicleta eco deluxe color egra cometidfo hecho huir rumbo desnocido	NA	0,02

Imputar 100 veces: 100 versiones de los datos (réplicas)



1. Partimos de una base con víctimas **únicas y campos faltantes**.

1. Aplicamos la imputación múltiple una y otra vez. Tenemos en cuenta la incertidumbre.

1. **Estimar** los parámetros de interés de cada conjunto de datos imputado. Versiones distintas, resultados distintos. Incertidumbre.

1. **Combinamos** nuestras imputaciones para tener la media de la estimación puntual. **Reglas de Rubin**.

Ver más en:
https://bookdown.org/mwheymans/bookmi/

HRDAG
Statisticians for human rights

Estimación del subregistro de víctimas

HRDAG
Statisticians
for human rights

JEP
JURISDICCIÓN ESPECIAL PARA LA PAZ
COMISIÓN DE LA VERDAD
HRDAG
Statisticians
for human rights

Informe metodológico del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística*

18 de agosto de 2022**

Datos faltantes - subregistro

Víctima	Año hecho	Etnia	Edad	Departamento	Responsable	Pertenece al conflicto	Sexo
Persona 1	1999	INDIGENA	14	GUAVIARE	FARC-EP	1	MUJER
Persona 2	2002	INDIGENA	17	GUAVIARE	FARC-EP	1	MUJER
Persona 3	1997	MESTIZO	24	BOLÍVAR	Múltiple	¿?	¿?
Persona 4	1998	INDIGENA	¿?	GUAVIARE	¿?	¿?	¿?
Persona 5	2008	ROM	¿?	NORTE DE SANTANDER	GUE-OTRO	1	HOMBRE
Persona 6	2002	¿?	2	BOYACÁ	¿?	0	MUJER
Persona 7	2005	AFROCOLOMBIANO	19	VALLE DEL CAUCA	FARC-EP	1	HOMBRE
¿?	¿?	¿?	¿?	¿?	¿?	¿?	¿?
¿?	¿?	¿?	¿?	¿?	¿?	¿?	¿?

HRDAG
Statisticians for human rights

¿Qué es el subregistro?

- Las bases de datos tienen limitaciones y son reflejos imperfectos de la realidad.
- No hay ninguna base de datos que cuente a **todas** las víctimas.
- Una base de datos nos habla del número de víctimas **registradas**, no real.

HRDAG
Statisticians
for human rights

¿Por qué hay subregistro?

- Las víctimas tienen razones lógicas para no denunciar.
- Las entidades que registran muchas veces no tienen las **capacidades** suficientes.
- Hay **limitaciones** geográficas, o estructurales que dificultan la documentación.

HRDAG
Statisticians
for human rights

¿Por qué estimar?

- No hacer estimaciones implica un supuesto: que los **datos faltantes faltan por azar**.
- Eso implicaría que la **distribución de los datos** sin estimaciones es igual a la distribución con estimaciones.
- Hay **factores estructurales** que hacen que la probabilidad de registro sea diferente para distintas víctimas.

HRDAG
Statisticians
for human rights

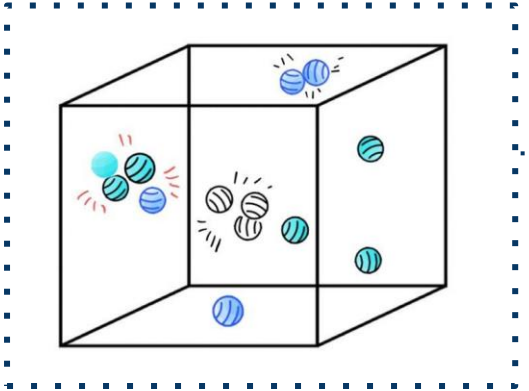
Estimación por Sistemas Múltiples (ESM)

Chao, Anne (2001). “An Overview of Closed Capture-Recapture Models.” *Journal of Agricultural, Biological, and Environmental Statistics* 6 (2): 158–75.”

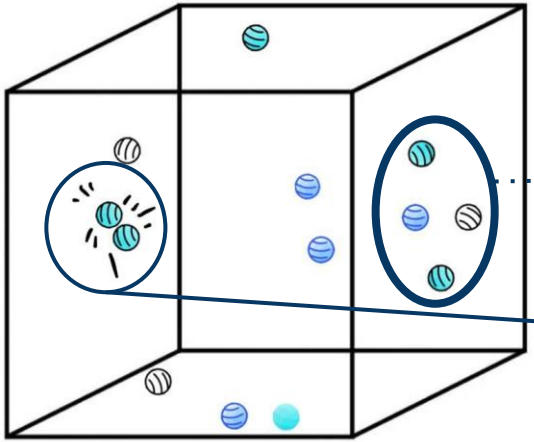
Bird, Sheila M, and Ruth King (2018). “Multiple Systems Estimation (or Capture-Recapture Estimation) to Inform Public Policy.” *Annual Review of Statistics and Its Application* 5: 95–118.

HRDAG
Statisticians
for human rights

Estimación por Sistemas Múltiples



Tamaño de la habitación: la población total.



Fuentes de datos o listas

Patrones de superposición para estimar el tamaño total de víctimas, incluído el subregistro

Estimación por Sistemas Múltiples

- Implementada por **primera vez en 1783** por el matemático Pierre-Simon Laplace
- Se implementó con 2 fuentes de datos y **4 supuestos**
- La aplicación moderna de ESM no necesariamente requiere de estos supuestos.

Amorós, Jaume (2014). “Recapturing Laplace.” *Significance*. https://rss.onlinelibrary.wiley.com/doi/full/10. 1111/j.1740-9713.2014.00754.x.

Scheuren, Fritz (2004). “History Corner.” *The American Statistician*. https://www.tandfonline.com/doi/abs/10.1198/0003130042926.

HRDAG
Statisticians for human rights

Estimación por Sistemas Múltiples

1. **La población estimada es "cerrada,"** es decir, miembros de la población no son creados ni eliminados durante el periodo de documentación:

    - No hay forma de salir de la población de víctimas una vez ingresada.

HRDAG
Statisticians
for human rights

Estimación por Sistemas Múltiples

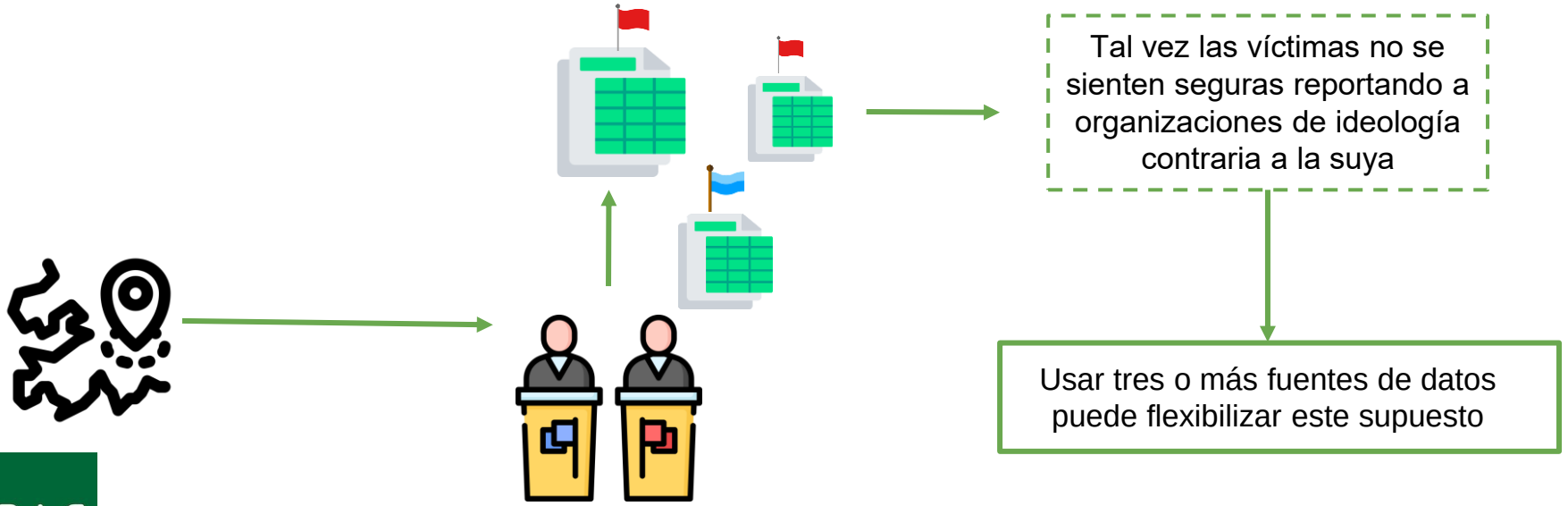
2. **El vínculo de los registros es preciso:**

- La ESM requiere una base de datos consolidada que incluya a cada una de las víctimas e indique la fuente que la registró.
- El proceso de vinculación no siempre es perfecto, el proceso semi-supervisado permite la evaluación del vínculo.
- Se ha demostrado que incluso si la vinculación no es buena, los resultados de ESM son confiables (J. Johndrow, Lum, and Dunson 2018)

HRDAG
Statisticians
for human rights

Estimación por Sistemas Múltiples

3. Estar documentado en un sistema o fuente no afecta la probabilidad de estar documentado en cualquier otro **(independencia de listas o fuentes)**:



HRDAG
Statisticians for human rights

Bishop, Yvonne M, Stephen E Fienberg, and Paul W Holland. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Springer.

Chao, Anne. (1987). "Estimating the Population Size for Capture-Recapture Data with Unequal Capturability." *Biometrics* 43: 783–91.

Estimación por Sistemas Múltiples

4. La probabilidad de ser documentado en un sistema o fuente particular es igual para todos los miembros de la población **(homogeneidad de captura)**.

- Esto probablemente no es cierto

- Para controlar esto, se agrupan registros similares: **estratificación**
  - En el proyecto **se estratifica por años** y otras variables, dependiendo del análisis.

HRDAG
Statisticians for human rights

Chao, Anne. (1987). "Estimating the Population Size for Capture-Recapture Data with Unequal Capturability." *Biometrics* 43: 783–91.

Sekar, C Chandra, and W Edwards Deming (1949). "On a Method of Estimating Birth and Death Rates and the Extent of Registration." *Journal of the American Statistical Association* 44 (245).

Insumos para la ESM con LCMCR

réplica	id_víctima	año	in_fuenteA	in_fuenteB	...	in_fuenteZ
R1	1	...	1	0	0	0
R1	2	...	1	1	0	0
R1	3	...	0	1	1	1
R1	.	...	...	...	...	...
R1	500	...	0	0	1	1

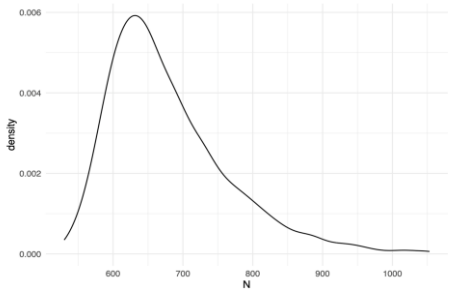
1 = la víctima aparece en la fuente
0 = la víctima **no** aparece en la fuente

Esta información no cambia para cada réplica

HRDAG
Statisticians
for human rights

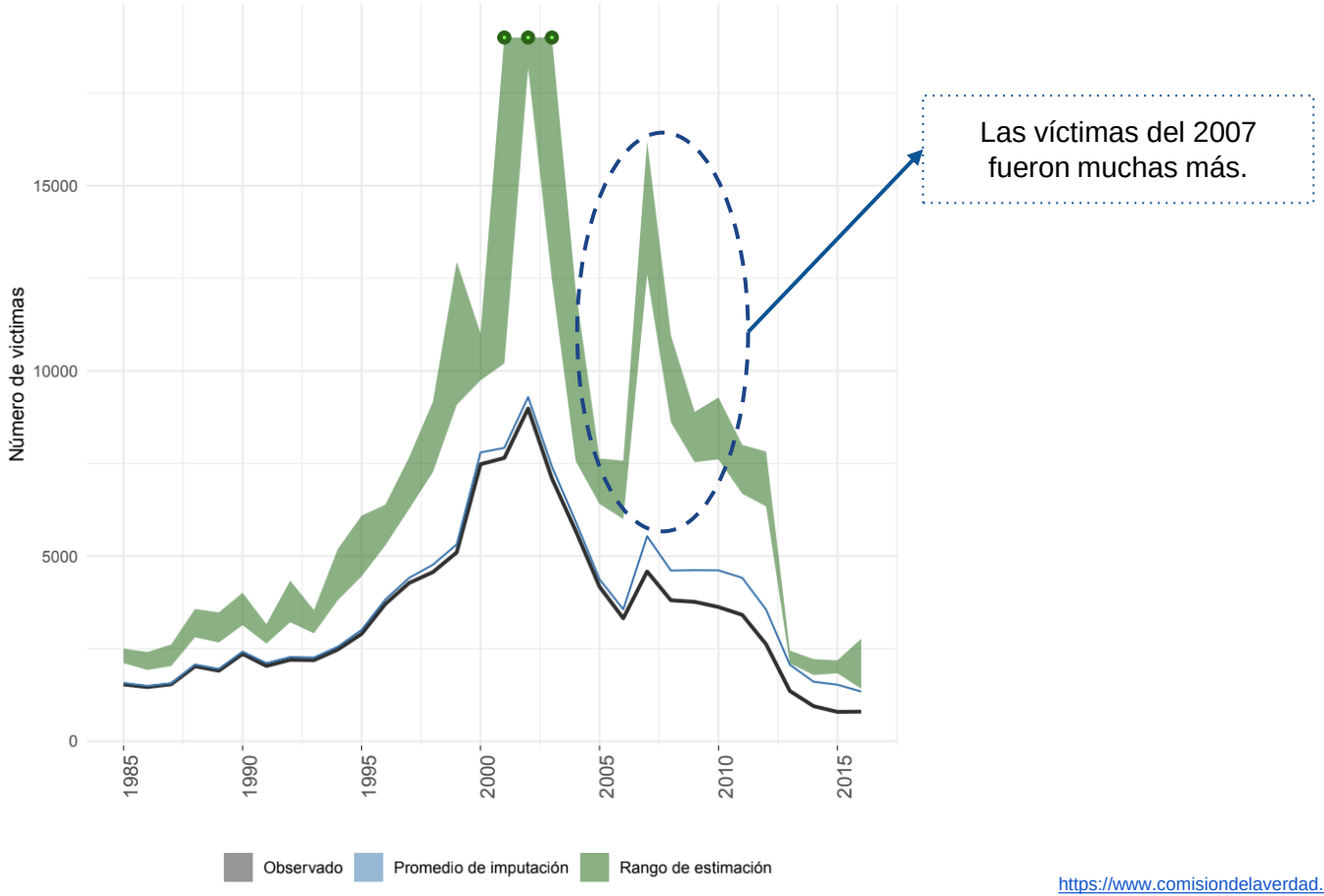
¿Cómo sabemos que la estratificación es adecuada?

- Contar con un caso de estudio o **una hipótesis de patrones de violencia.**
- **Estratos con datos “suficientes”**. Al menos 3 fuentes con al menos una víctima.
- Estratos que no presenten una distribución **multimodal o bimodal**, es decir, distribuciones con más de un pico (después de la estimación).



HRDAG
Statisticians for human rights

Víctimas de desaparición entre 1985 y 2016



HRDAG
Statisticians for human rights

Tomado de: Informe metodológico del proyecto JEP-CEV-HRDAG. https://www.comisiondelaverdad.co/sites/default/files/descargables/2022-08/04_Anexo_Proyecto_JEP_CEV_HRDAG_08022022.pdf

https://www.comisiondelaverdad.co/violaciones-de-derechos-humanos-infracciones-al-derecho-internacional-humanitario-y/desaparicion

Links

- Página web de la Comisión de la Verdad

https://www.comisiondelaverdad.co/analitica-de-datos-informacion-y-recursos#c3

- Repositorio del paquete de R en GitHub (en desarrollo)

https://github.com/HRDAG/verdata

- Repositorio de ejemplos del uso de los datos (en desarrollo)

https://github.com/HRDAG/verdata-examples

- Informe metodológico CEV-JEP-HRDAG

https://hrdag.org/wp-content/uploads/2022/08/20220818-fase4-informe-corrected.pdf

HRDAG
Statisticians
for human rights

Gracias.

HRDAG
Statisticians
for human rights

JEP
JURISDICCIÓN ESPECIAL PARA LA PAZ
COMISIÓN DE LA VERDAD
HRDAG
Statisticians
for human rights

Informe metodológico del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística*

18 de agosto de 2022**