Comparación del Libro Amarillo con bases de datos de derechos humanos históricas en El Salvador

Patrick Ball

20 de agosto de 2014

Human Rights Data Analysis Group
everybody counts.

1. Introducción

Este proyecto fue realizado en base a los 1857 registros identificados por nombre en el Libro Amarillo. Hicimos una comparación de los nombres del Libro Amarillo con cuatro bases de datos creadas en El Salvador: una de la organización no-gubernamental Comisión de Derechos Humanos de El Salvador (CDHES); una de los informes publicados por Tutela Legal del Arzobispado sistematizados por El Rescate (Rescate); una base de datos que reúne las denuncias recibidas por la Comisión de la Verdad por medio de testimonios directos (CV1); y la base de datos de denuncias recibidas por la Comisión de Verdad por medio de "fuentes indirectas" (CV2).

El cruce del Libro Amarillo con las demás bases de datos se presenta y se explica en la Sección 1.1 a continuación, y se muestra visualmente en la Tabla 1.

Este informe no llega a conclusiones sustanciales sobre patrones. Únicamente describe los patrones cuantitativos de los vínculos observados entre el Libro Amarillo y los demás listados.

Todos los números y cifras incluidos en este informe están formateados según los estándares de la Organización Internacional de Normalización 31-0. El separador de miles se indica con un espacio pequeño, y el decimal se indica con un punto. Los términos técnicos que provienen del inglés están escritos en *cursiva*; los nombres de programas de computadora específicos, las variables y los campos utilizados en el análisis se indican en letra `typewriter`. Los enlaces que hacen referencia a este mismo documento se indican en rojo y los enlaces a sitios web externos se indican en color azul.

1.1. Hallazgos

El Cuadro 1 muestra el número de nombres encontrados en el Libro Amarillo que coinciden con los nombres de los demás listados.

Esto significa, por ejemplo, que 127 de los nombres del Libro Amarillo coincidieron con un nombre registrado como víctima de asesinato por las fuentes directas de la Comisión de Verdad (CV1), mientras que 21 nombres del Libro Amarillo coincidieron con nombres de desaparecidos según las fuentes directas de la Comisión

Cuadro 1: Conteo de registros y coincidencia entre bases de datos históricas y el Libro Amarillo

	assinato	desaparición	tortura	detención	otro
CDHES	35	75	157	337	179
Rescate	93	65	43	397	66
CV1	127	21	7	NA	7
CV2	124	151	123	NA	6
Libro Amarillo	273	233	274	538	242
Libro Amarillo %	15	13	15	29	13

de Verdad. (Cabe destacar que ninguno de los listados de la Comisión de Verdad registró denuncias de detención.)

Es posible que la misma víctima haya sufrido violaciones de derechos humanos fatales y no-fatales. Los números del Cuadro 1 representan la cantidad de violaciones, no de víctimas. Una misma victima puede haber sufrido una o varias violaciones de sus derechos humanos, por lo cual no pueden sumarse los números por fila.

De la misma manera, un nombre del Libro Amarillo puede aparecer en una o en varias de las bases de datos, por lo cual no pueden sumarse las columnas del Cuadro 1. El conteo de los nombres del Libro Amarillo sumado de todas las bases de datos se encuentra en la penúltima fila. Por ejemplo, el conteo de nombres que coinciden con denuncias de asesinatos en uno o más de los otros listados es 273, y el total de nombres del Libro Amarillo que coinciden con desapariciones es 233. El porcentaje de nombres del Libro Amarillo que coinciden con uno o más de los registros de los otros listados se encuentra en la última fila: por ejemplo, el 29 % de los nombres del Libro Amarillo coincide por lo menos con una denuncia de detención.

Por último, 1063 registros del Libro Amarillo no fueron encontrados en los demás listados.

Los nombres del Libro Amarillo vinculados con violaciones de derechos humanos documentadas en los otros listados (las dos últimas filas del Cuadro 1) pueden ser subdivididos por el perpetrador a quien se atribuye la violación según la documentación original. El Cuadro 2 describe las violaciones de derechos humanos sufridas por personas nombradas en el Libro Amarillo con los supuestos perpetradores de las violaciones. Por ejemplo, la primera celda del Cuadro 2 muestra

Cuadro 2: Conteo de violaciones de derechos humanos de personas nombradas en el Libro Amarillo atribuidas a actores armados en las bases de datos históricas

	asesinato	desaparición	tortura	detención	otro
Ejército	177	71	109	208	81
Fuerzas de seguridad[a]	70	73	160	344	68
Escuadrones de la Muerte	128	43	32	81	37
FMLN	19	12	0	1	2
Libro Amarillo	273	233	274	538	242
Libro Amarillo %	15	13	15	29	13

[a]Policía Nacional, Guardia Nacional, y/o Policía de Hacienda

que según los otros listados, 177 personas nombradas en el Libro Amarillo fueron asesinadas por el Ejército de El Salvador, mientras que 128 de las personas nombradas en el Libro Amarillo fueron asesinadas por escuadrones de la muerte. La misma víctima puede haber sido asesinada en un operativo en donde participaron unidades del ejército y paramilitares, por lo cual el asesinato estaría registrado en ambas filas del Cuadro 2. De la misma manera, la misma víctima puede haber sido torturada y asesinada, por lo cual no pueden sumarse las columnas. Las dos últimas filas del Cuadro 2 se reproducen en el Cuadro 1.

Para calcular el porcentaje de personas nombradas en el Libro Amarillo que sufrieron alguna violación dada por algún perpetrador dado, es necesario dividir la cantidad en la celda relevante del Cuadro 2 por el total de personas nombradas en el Libro Amarillo, que son consideradas en este informe (1857). Así, el 10 % de las personas nombradas en el Libro Amarillo fueron asesinadas por el Ejército, mientras que el 19 % de las personas nombradas en el Libro Amarillo fueron detenidas por las fuerzas de seguridad (Policía Nacional, Guardia Nacional, y/o Policía de Hacienda).

2. Métodos

Emparejamos cada nombre del Libro Amarillo con los ocho nombres de mayor *proximidad* del conjunto de los nombres de las *bases de datos históricas*. Definimos las ideas en cursiva en orden inverso:

Las *bases de datos históricas* incluyen la CDHES;[1] la base de datos del proyecto de El Rescate/Tutela Legal;[2] el listado de fuentes directas de la Comisión de Verdad (CV1); y el listado de fuentes indirectas de la Comisión de Verdad (CV2).[3] De manera conjunta, nos referimos a estos últimos como los otros listados, y éstos contienen 62 292 registros.[4] Cabe destacar que un nombre del Libro Amarillo puede coincidir con uno, dos, tres, o cuatro de los otros listados.

Por el término *proximidad*, entendemos la similitud que existe entre los nombres. Una manera de imaginar la similitud entre nombres es preguntarse cuántos caracteres tendrían que insertarse o borrarse para que un nombre se convierta en otro. Esta medida es denominada la distancia Levenshtein. Para efectos de este informe, utilizamos una medida relacionada denominada medida Jaro (Winkler, 2006).

La medida Jaro está optimizada para comprobar la similitud de nombres, y fue desarrollada por estadísticos de la Oficina del Censo de los EE.UU. Tiene un rango estandarizado entre 0 y 1. Probé cada uno de los 1857 nombres del Libro Amarillo con cada uno de los 37 520 nombres únicos encontrados dentro de uno o más de los otros listados.[5]

Es decir, calculamos una medida Jaro para cada par de nombres entre los 1857 nombres del Libro Amarillo y los 37 520 nombres únicos de los otros listados. Para cada nombre del Libro Amarillo, seleccionamos los ocho nombres más parecidos, es decir, ocho nombres con la medida Jaro más grande; luego uns persona revisó manualmente los ocho nombres comparados. En algunos casos, hubo menos de ocho nombres remotamente plausibles de los otros listados para comparar con el nombre del Libro Amarillo ("remotamente plausible" significa una medida Jaro >0.7). En total, hubo 14 833 pares de nombres revisados manualmente.

Antes de realizar las comparaciones, hicimos un sorteo de identificadores de nombres. Por "identificador" entendemos una parte de un nombre más largo. Esto significa que si un nombre fue escrito como "CACERES JOAQUIN" y el otro como "JOAQUIN CACERES CERES" sorteamos cada cadena de caracteres para

[1] Ver Ball (2000).
[2] Ver Howland (2008).
[3] Ver United Nations (1993).
[4] De los 62 292 registros, 37 520 son nombres únicos, porque algunos registros tienen los mismos nombres que otros.
[5] Muchos nombres estaban registrados varias veces en los otros listados; los nombres idénticos de los diferentes registros de los otros listados fueron combinados para ser comparados con los nombres del Libro Amarillo.

luego comparar "CACERES JOAQUIN" con "CACERES CERES JOAQUIN." La resultante medida Jaro es 0.905, la cual indica que existe proximidad entre los nombres.

Con el término *emparejar*, queremos decir que una persona revisó manualmente 14 833 combinaciones de nombres. Muchas de las comparaciones fueron realizadas dos veces por dos personas diferentes, y la comparación de sus hallazgos se describe en la Sección 4.1.

3. Anexo de Datos

El Cuadro 3 muestra el total de registros dentro de los otros listados.

Cuadro 3: Conteo de registros de las bases de datos históricas de violaciones de derechos humanos en El Salvador

	asesinato	desaparición	tortura	detención	total
CDHES	1870	2555	3708	8745	24 550
Rescate	5030	3040	950	6010	16 929
CV1	5604	994	434	NA	7258
CV2	7376	3682	2230	NA	13 555

4. Anexo de Medición

En el fondo, el proceso de emparejamiento de los nombres implica decidir si un par de registros refieren a la misma persona o no. Enfrentados a un mismo par de nombres, las personas que realizan la comparación pueden tomar la misma decisión, o bien decisiones diferentes. Las medidas descritas en esta sección explican cómo evaluamos la cercanía entre las decisiones de dos analistas distintos.[6]

En el contexto de este informe, un analista fue designado arbitrariamente como A y el otro como B. Con estas designaciones, definimos lo siguiente

- un *verdadero positivo* (VP) es un par identificado como emparejado por las dos personas;

[6]Estas medidas son empleadas ampliamente en la literatura estadística y de las ciencias informáticas sobre el aprendizaje de máquinas. Ver, por ejemplo, Witten et al. (2011), especialmente capítulo 5.

- un *falso positivo* (FP) es un par identificado por la persona A como no emparejado pero identificado como emparejado por la persona B;
- un *falso negativo* (FN) es un par identificado por la persona A como emparejado pero identificado como no emparejado por la persona B;
- un *verdadero negativo* (VN) es un par identificado como no emparejado por las dos personas.

Dadas esas categorías, podemos definir tres medidas adicionales. La primera es *recall*, que es la proporción de los pares clasificados correctamente. La medida recall describe la tasa de respuestas que la persona B identifica como no emparejado cuando la persona A identifica un emparejado. Un valor de recall más alto significa que existe menos desacuerdo de este tipo.

$recall = \frac{VP}{VP+FN}$

La segunda medida es la *precisión*, una descripción de qué tan seguido la persona B identifica un emparejado cuando la persona A también identifica un emparejado. Una medida de precisión más alta significa que existe menos desacuerdo de este tipo.

$precision = \frac{VP}{VP+FP}$

La medida F es el balance entre precisión y recall, y se calcula así:

$F = \frac{2 \times recall \times precision}{recall + precision}$

F es una medida general de la calidad de clasificación. La medida F se usa para describir la confiabilidad entre los analistas en la Sección 4.1.

4.1. Confiabilidad entre analistas

La confiabilidad de los resultados del emparejamiento es cuestión de si son correctas o no las decisiones tomadas por los analistas sobre si los dos nombres representan o no a la misma persona. No es posible evaluar la verdadera exactitud—no sabemos exactamente quiénes son las personas registradas en el Libro Amarillo, y tampoco las de los otros listados. En lugar de eso, consideramos la frecuencia del acuerdo versus el desacuerdo entre los dos analistas en cuanto a sus decisiones de si los nombres refieren a la misma persona o no. Los dos analistas

revisaron un total de 1604 pares en común. La medida de su acuerdo o desacuerdo se llama confiabilidad entre analistas (*inter-rater reliability*, o *IRR*). Hay varias maneras de calcular el IRR, pero en este informe, utilizamos la medida F definida en la Sección 4 ubicada más arriba.

Para los nombres emparejados, $F^+ = 0.79$, mientras que para los nombres no emparejados, $F^- = 0.91$. En esencia, los analistas estuvieron de acuerdo sobre los nombres no emparejados, pero estuvieron menos de acuerdo sobre los nombres no emparejados positivamente.

Las ideas de precisión y recall nos pueden revelar mayor información sobre el tipo de desacuerdos entre los analistas. Para los nombres emparejados positivamente, $precision^+ = 0.66$, $recall^+ = 0.99$. La medida más baja para la precisión sugiere que los analistas tendieron a estar en desacuerdo de manera que uno decidió más frecuentemente que un par de nombres estaba emparejado positivamente. La mediada relativamente alta de recall sugiere que el desacuerdo fue de un lado solamente: uno de los analistas fue más renuente a identificar a los pares como emparejados positivamente. Sin embargo, cuando el analista más renuente decidió que un par de nombres estaba emparejado, el analista más agresivo casi siempre estuvo de acuerdo.

Para fines de este informe, se han utilizado las decisiones del analista más renuente, específicamente para estimar un nivel mínimo de superposición entre el Libro Amarillo y los otros listados.

Referencias

Ball, P. (2000). The Salvadoran Human Rights Commission: Data Processing, Data Representation, and Generating Analytical Reports. In Ball, P., Spirer, H. F., and Spirer, L., editors, *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis*. AAAS, Washington, DC.

Howland, T. (2008). How El Rescate, a Small Nongovernmental Organization, Contributed to the Transformation of the Human Rights Situation in El Salvador. *Human Rights Quarterly*, 30(3):703–757.

United Nations (1993). *From Madness to Hope: The 12-Year War in El Salvador: Report of the Commission on the Truth for El Salvador*.

Winkler, W. E. (2006). Overview of Record Linkage and Current Research Directions. Technical Report RRS2006/02, Statistical Research Division, U.S. Census Bureau.

Witten, I., Frank, E., and Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufman.

Sobre HRDAG

El Grupo de Análisis de Datos de Derechos Humanos (HRDAG) es una organización sin fines de lucro,[7] no-partidaria, que aplica métodos científicos al análisis de violaciones de derechos humanos alrededor del mundo. Este trabajo comenzó en 1991 cuando Patrick Ball empezó a desarrollar bases de datos para organizaciones de derechos humanos en El Salvador. HRDAG creció con la Asociación Americana para el Avance de la Ciencia durante 1994-2003, y con la Iniciativa Benetech durante 2003–2013. En febrero de 2013, HRDAG se independizó, y hoy tiene sus oficinas en San Francisco, California; se pueden encontrar detalles de contacto y mayor información en el sitio web de HRDAG y en Facebook.

HRDAG está compuesto por estadísticos aplicados y matemáticos, cientistas en computación, demógrafos, y cientistas sociales. HRDAG apoya las garantías

[7] Formalmente, HRDAG es un proyecto patrocinado financieramente por Community Partners.

establecidas en la Declaración Universal de Derechos Humanos, el Convenio Internacional sobre Derechos Civiles y Políticos, y otros tratados e instrumentos internacionales de derechos humanos. Los científicos de HRDAG proveen resultados no sesgados, basados en la ciencia, a activistas de derechos humanos para esclarecer violaciones de derechos humanos. HRDAG cree que las estadísticas sobre violencia deben ser lo más veraces posible, utilizando los mejores datos y métodos científicos.

Este proyecto fue diseñado por Patrick Ball, quien también escribió el software utilizado para el análisis y supervisó todos los procesos asociados al proyecto. Las ideas técnicas fueron desarrolladas en colaboración con el Dr. Scott Weikart y Dra. Megan Price. El análisis manual fue realizado por la consultora de manejo de datos de HRDAG, Michelle Dukich, y por Phil Neff, Coordinador del Proyecto de Oraciones Incompletas del Centro de Derechos Humanos de la Universidad de Washington. El informe fue traducido al castellano por Ursula Mosqueira del Centro de Derechos Humanos de la misma universidad.

El contenido de este informe representa las opiniones de los autores y editores y no debiera ser interpretado como el punto de vista de HRDAG, de las contrapartes de HRDAG, de la junta de asesores de HRDAG, de los donantes de HRDAG, o de este proyecto en su conjunto. El documento "Comparación del Libro Amarillo con bases de datos de derechos humanos históricos en El Salvador" por Patrick Ball está bajo una licencia de Creative Commons Attribution-NonCommercial 4.0 International License. Es posible que pueda obtenerse permiso fuera del alcance de esta licencia en https://hrdag.org.