Universidad de San Andrés

Departamento de Matemática y Ciencias

Maestría en Ciencia de Datos

Revelando a los presuntos responsables "desconocidos" de asesinatos selectivos en el conflicto armado colombiano

Autora: Valentina Rozo Ángel

Director: Lucas Fernández Piana

2023

Maestría en Ciencia de Datos
Departamento de Matemática y Ciencias

Revelando a los presuntos responsables “desconocidos” de asesinatos selectivos en el conflicto armado colombiano

Valentina Rozo Ángel **2023**

Director: Lucas Fernández Piana

Universidad de SanAndrés

Universidad de
SanAndrés

Resumen

El conflicto armado colombiano es uno de los más largos del hemisferio occidental. Sin embargo, conocer la responsabilidad de los distintos presuntos responsables resulta difícil, pues esta variable suele estar incompleta. La literatura establece que existen tres tipos de datos faltantes: completamente aleatorios, aleatorios y no aleatorios. Pero ¿qué tipo de datos faltantes es el presunto responsable?

En este artículo utilizo los datos de asesinatos selectivos del Centro Nacional de Memoria Histórica y el algoritmo de missForest para estudiar el tipo de datos faltantes. Por medio de dos estrategias, dejando los NAs originales y eliminándolos, y por medio de un puntaje global propuesto, muestro que el algoritmo tiene un mejor desempeño con los NAs originales. Este es un indicio de que, tal y como indica la experiencia cualitativa, los campos faltantes de esta variable no son completamente aleatorios. Además, muestro que la responsabilidad de los paramilitares sería mayor a la actualmente documentada, siendo estos los principales responsables de asesinatos selectivos en Colombia.

Universidad de SanAndrés

Agradecimientos

Mi cambio de carrera hacia la ciencia de datos no habría sido posible sin mi equipo base, que siempre cree en mi y me impulsa a seguir mis sueños. Gracias a mis papás por apoyarme y acompañarme. Gracias a Alejandro por cambiar nuestra vida en una conversación nocturna, irnos a vivir a Buenos Aires, enseñarme sobre conflicto armado y DIH, e impulsarme a volar en Yale. Y a Tumbao y Pub porque desde lejos siempre están cerca.

Gracias a Lucas, quien aceptó mi proyecto y lo acompañó con todos sus cambios: inició como aprendizaje no supervisado y terminó como supervisado. Sin toda su retroalimentación esta tesis no existiría. A UDESA, que me enseñó y me dio un lugar feliz en CABA. Y a mis amigas de Argentina, con quienes fuimos felices haciendo TPs los fines de semana y seguimos riéndonos hasta hoy.

Por último, gracias a la Fox International Fellowship por llevarme a Yale, donde escribí esta tesis. A quienes la leyeron y retroalimentaron, gracias. A Elisabeth que se interesó desde un principio, gracias. A mis compañeros, amigas y amigos que hicieron que fuera un año feliz mientras la escribía, gracias.

SanAndres

Índice general

QUAERERE VERUM
Universidad de SanAndrés

Universidad de
SanAndrés

Introducción

En 2023 la Corte Interamericana de Derechos Humanos determinó que Colombia era responsable del exterminio del partido político de izquierda Unión Patriótica (UP). Durante el litigio, el Estado colombiano argumentó que de las más de 6.000 víctimas identificadas por los demandantes, no era posible determinar la responsabilidad estatal en 3.719 casos porque los representantes de las víctimas no demostraron suficientes elementos fácticos y probatorios de su participación. En respuesta, la Corte Interamericana estableció que existían causas razonables para justificar las posibles inconsistencias y falta de detalle de los demandantes. Esta Corte expresó que era razonable no exigir precisión de todas las circunstancias de los homicidios y desapariciones forzadas ocurridas a lo largo de dos décadas, ya que el Estado no hizo una investigación adecuada (Corte Interamericana de Derechos Humanos, 2022). El caso de la UP ilustra la dificultad de asignar responsabilidades por violaciones masivas de derechos humanos, pues muchas veces la información no está completa.

Si bien es clave estudiar los patrones de violencia política, definidos como "configuraciones de repertorios, objetivo, frecuencia y técnica en la que una organización armada se involucra regularmente" (Gutiérrez-Sanín y Wood, 2019), un requisito para ello es conocer al perpetrador. Sin embargo, es de esperar que no todas las víctimas tengan la información sobre quién cometió el hecho victimizante. Por lo tanto, los patrones de violencia podrían basarse en datos incompletos, lo que afecta el derecho de las víctimas y de la sociedad a la verdad, pone en riesgo el derecho a la no repetición y refuerza la impunidad.

Existen diferentes razones para explicar la falta de información sobre el perpetrador o los "datos faltantes" en una base de datos. En primer lugar, las víctimas pueden tener miedo a las represalias si denuncian lo ocurrido. En segundo lugar, puede ser que el suceso se produjera en situaciones confusas, en las que no es posible identificar al autor. Además, la violación podría haber tenido lugar sin testigos que la denunciaran, entre otras. Las comisiones de la verdad de diferentes países se han enfrentado al reto de los datos faltantes en la variable de perpetrador o presunto responsable. Por ejemplo, la Comisión de la Verdad de Perú estimó que alrededor del 20 % de los homicidios durante el conflicto armado fueron cometidos por un perpetrador no identificado o por uno distinto al Estado o al Sendero Luminoso (Ball, et al. 2003). Mientras que, en el caso de Colombia, tras combinar 112 bases de datos de violaciones de derechos humanos, la Comisión de la Verdad colombiana determinó que el

perpetrador no fue identificado en el 66 % de los homicidios ocurridos durante el conflicto armado entre 1985 y 2018 (CEV, JEP, HRDAG, 2022).

La comprensión que las sociedades tienen de las violaciones masivas de los derechos humanos puede variar drásticamente con base en la distribución de los perpetradores "desconocidos". Aunque el campo de la estadística lleva décadas analizando los datos faltantes, comenzando por Rubin (1976), el análisis de los perpetradores en contextos de violencia política suele tratarse de dos formas diferentes. La primera es estudiar la categoría "desconocido" y dar estadísticas descriptivas basadas en ella, como la Comisión de la Verdad peruana. La segunda consiste en excluir a los autores "desconocidos" del análisis. Este método se conoce en estadística como "eliminación de datos" (*listwise deletion* en inglés).

En este trabajo utilizo los datos del Centro Nacional de Memoria Histórica (CNMH) de Colombia para estudiar el mecanismo detrás de los datos faltantes de la variable "presunto responsable". En esta introducción presento los datos del CNMH y las responsabilidades documentadas. En el primer capítulo ahondo en los datos faltantes. Para comenzar, expongo las razones para creer que los datos faltantes en violaciones a derechos humanos no se dan de forma completamente aleatoria y presento la literatura de datos faltantes con sus tres categorías: completamente aleatorios (*missing completely at random* - MCAR), aleatorios (*missing at random* - MAR) y no aleatorios (*missing not at random* - MNAR). Después, describo el algoritmo elegido para imputar al presunto responsable o perpetrador -missForest-, así como la estrategia que utilizo para evaluar la imputación. En el segundo capítulo presento los resultados y las limitaciones relacionadas con los datos y el método. Por último, en el tercer capítulo, concluyo en dos maneras. Primero, muestro que la responsabilidad de los paramilitares cuando se imputa al actor "desconocido" es superior a la que se ha documentado (71 % en lugar de 44 %). Segundo, reflexiono sobre la mejora del rendimiento del algoritmo con los datos faltantes originales, lo que sugiere que la ausencia del perpetrador no es MCAR. Por el contrario, los resultados indican que en los datos faltantes siguen una distribución, que puede ser identificada por el algoritmo. Por lo tanto, las variables observables, como "año" o "departamento" podrían explicar por qué falta el responsable.

Panorama de los homicidios en el conflicto armado colombiano con base en la información recopilada por el Centro Nacional de Memoria Histórica

Colombia ha tenido un conflicto armado activo por lo menos desde 1964 con la fundación de los grupos guerrilleros marxista-leninistas Fuerzas Armadas Revolucionarias de Colombia (FARC) y el Ejército de Liberación Nacional (ELN). Un año después, el gobierno publicó el Decreto 3398 de 1965, que se convertiría en la Ley 48 de 1968. Esta ley permitió la creación de fuerzas de seguridad privadas con entrenamiento militar y armamento, que serían conocidas como "paramilitares". Desde entonces, durante el conflicto armado colombiano, múltiples grupos armados organizados, así como el Estado, han cometido violaciones masivas de los derechos humanos.

Décadas después del inicio del conflicto se promulgó la Ley de Víctimas (Ley 1448 de 2011) con el objetivo de establecer un conjunto de medidas para beneficiar a las víctimas del conflicto armado y garantizar sus derechos a la verdad, la justicia y la reparación (Congreso de Colombia, 2011). Como parte de estas medidas, se creó el CNMH. La misión del CNMH es "la recepción, recuperación, conservación, compilación y análisis de todo el material documental, testimonios orales y por cualquier otro medio, relativo a las violaciones ocurridas con ocasión del conflicto armado interno colombiano" (Decreto Número 4803 2011). Como parte de su mandato, el CNMH creó el Observatorio de Memoria y Conflicto (OMC) con la tarea de unificar e integrar múltiples conjuntos de datos sobre el conflicto armado a partir de 11 tipos diferentes de actos violentos entre 1958 y la actualidad (Centro Nacional de Memoria Histórica, sin fecha-a). Uno de ellos son los "asesinatos selectivos", en los que me centro y a los que me refiero a lo largo del documento indistintamente como homicidios o asesinatos. Cabe señalar que el OMC también documenta "masacres", entendidas como cuatro o más asesinatos con gran exposición pública. No incluyo a las víctimas de masacres en este análisis porque, dada su magnitud, suelen tener más visibilidad. Por lo tanto, la probabilidad de que se documente a un autor "conocido" podría ser diferente a la de los asesinatos selectivos. Entonces, el comportamiento de los datos faltantes pueden ser de distintos tipos entre los dos hechos.

Según los datos de la OMC, entre 1958 y marzo de 2022 se habían cometido en total 178.243 homicidios en el marco del conflicto armado en Colombia[1]. El OMC tiene 10 categorías para el autor conocido: agente del Estado, agente del Estado y grupo paramilitar, agente del Estado y grupo posdesmovilización, agente extranjero, bandolerismo, crimen organizado, grupo paramilitar, grupo posdesmovilización, guerrilla y otro. Tres de ellos son responsables de más del 90 % de los homicidios con autores conocidos: grupos paramilitares, guerrilla y agentes del Estado. Con el objetivo de reducir la cantidad de categorías, decido excluir aquellas con menos de 1.000 víctimas documentadas (0,48 % del total) -agente extranjero, crimen organizado, Estado-paramilitares y Estado-posdesmovilización-.

Además de los autores conocidos, el OMC tiene tres categorías diferentes para los casos en los que no se conoce al presunto responsable: "desconocido", "grupo armado no identificado" y "grupo armado no dirimido". "Desconocido" se refiere a los casos en los que no hay información sobre el autor, pero la víctima forma parte de una población vulnerable. "Grupo armado no identificado" se aplica a los casos en los que la víctima menciona características relacionadas con un grupo armado, pero no sabe cuál es. Por ejemplo, el hecho de que el agresor llevara uniforme. "Grupo armado no dirimido" es una categoría para los casos en los que dos o más fuentes tienen información contradictoria sobre el autor (Centro Nacional de Memoria Histórica, s.f.). En este artículo integro "desconocido" y "no identificado" como "desconocido". Esto se debe a que la

[1] Esta cifra no debe considerarse el número "final" de homicidios ocurridos durante el conflicto armado. Corresponde a lo que el CNMH está en capacidad de documentar. La Comisión de la Verdad de Colombia utilizó estimación por sistemas múltiples para calcular las víctimas de homicidio en Colombia. Según las estimaciones, entre 1985 y 2018 habría habido entre 777.852 y 852.756 víctimas con un intervalo de credibilidad del 95 % (CEV, JEP, HRDAG, 2022).

implicación práctica de los dos es la misma: no se conoce al responsable. En cambio, en el caso de "no dirimido" sí hay información sobre el grupo. Entonces, excluyo a las víctimas de un presunto responsable "no dirimido".

También excluyo a las personas que fueron identificadas como combatientes. Esto se debe a que me interesa estudiar las responsabilidades por la victimización de civiles durante los conflictos armados. Una "víctima", para este propósito, es una persona que sufrió violaciones al Derecho Internacional Humanitario por parte de diferentes grupos armados o del Estado. Los asesinatos de miembros de los grupos armados o de las fuerzas de seguridad y militares del Estado quedarían fuera del ámbito de esta definición, dado que los combatientes suelen ser objetivos militares legítimos[2].

Siguiendo la definición de un patrón de violencia política por Gutiérrez-Sanín y Woods (2019), en el Cuadro 1 presento las variables que pueden ser informativas dentro de cada categoría. Hay otras variables que relacionadas con el objetivo que serían útiles, pero que tienen más de un 40 % de datos faltantes, como la edad, la etnia, el tipo de población vulnerable y la afiliación política. Debido al alto porcentaje de datos que faltan, opto por no utilizarlas.

Cuadro 1: Estadística descriptiva

Variable	**Tipo**	**Niveles**	**% datos faltantes**
Tiempo y lugar			
Municipio	Categórica	1070	0.34
Año	Categórica	64	0.85
Mes	Categórica	12	1.78
Objetivo			
Sexo de la víctima	Categórica	2	1.75
Técnica			
Técnica[3]	Categórica	20	32.99
Arma			
Arma cortopuntante	Categórica	2	0
Arma de fuego	Categórica	2	0
Artefacto explosivo	Categórica	2	0
Asfixia mecánica	Categórica	2	0
Material incendiario	Categórica	2	0
Motosierrra	Categórica	2	0
Objeto contundente	Categórica	2	0
Otro	Categórica	2	0
Uso de químicos tóxicos	Categórica	2	0
Presunto responsable			
Presunto responsable	Categórica	6	33.1

[2]Hay excepciones en las cuales los combatientes no son objetivos militares legítimos. Por ejemplo, los prisioneros de guerra.

[3]i) Asalto, ii) asesinato circunstancial, iii) ataque a propiedad, iv) ataque indiscriminado, v) atentado, vi) citación, vii) engaño, viii) falso positivo, ix) incursión, x) interceptación, xi) operación militar, xii) otra, xiii) persecución, xiv) resistencia a la retención, xv) retén, xvi) retención/ejecución, xvii) reunión pública, xviii) ruta, xix) sicariato, xx) NA.

Al estudiar el perpetrador (Figura 1), los paramilitares serían responsables del 44 % de los homicidios documentados, seguidos por la guerrilla con 16 %, grupos posdesmovilización y el Estado con 3 % cada uno y bandoleros con 1 %. Sin embargo, para el 33 % de las víctimas el presunto responsable es un dato faltante.

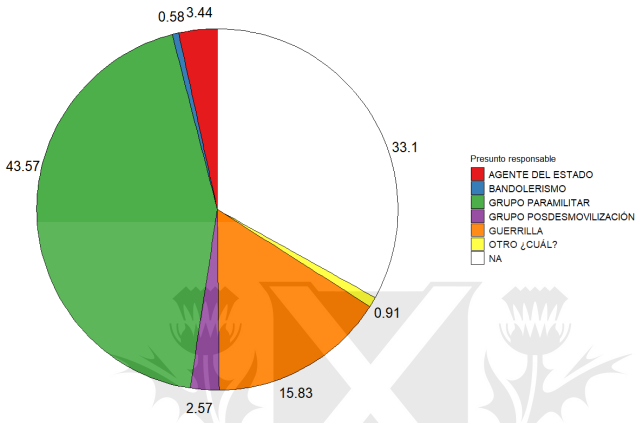

Figura 1: Responsabilidad en los asesinatos selectivos documentados por el CNMH.
Nota: NA equivale a "no disponible" ("*not available*" en inglés).

Al observar el comportamiento de los presuntos responsables a lo largo de los años en la Figura 2 (arriba) se observa que la violencia homicida empezó a crecer en los años 80, alcanzando su máximo en 2002. Mientras que al analizar el perpetrador por año en la Figura 2 (abajo), es claro que la proporción de NAs no es constante a través de los años. Esta proporción creció entre 1965 y 1975, descendió durante un par de años y volvió a aumentar en la década de 1990. Luego, en 1995 empezó a bajar hasta alrededor del 25 % y aumentó en 2014 por encima del 75 %. Cabe señalar que la proporción de NA no alcanza su punto más alto cuando lo hace la violencia. De hecho, en los últimos años la proporción de NA ha crecido mientras que la violencia se ha reducido. También es interesante encontrar que en 1968 no se registró ningún agresor.

Hay algunas características de los distintos agresores que merece la pena destacar en el Figura 2 (abajo), según los responsables "conocidos". Para empezar, hay registros de bandoleros después de 1964 aunque se considera que estos grupos dejaron de estar activos en ese año (Centro Nacional de Memoria Histórica, s.f-b). Por su parte, los paramilitares empiezan a ser actores principales en 1982 y alcanzan el mayor porcentaje de responsabilidades con su expansión como Autodefensas Unidas de Colombia (AUC), una coalición de grupos paramilitares creada a finales de los 90s. Las AUC se desmovilizaron en 2006 y desde ese año hay presencia de grupos posdesmovilización. Por último, la responsabilidad de la guerrilla es mayor en los años 70.

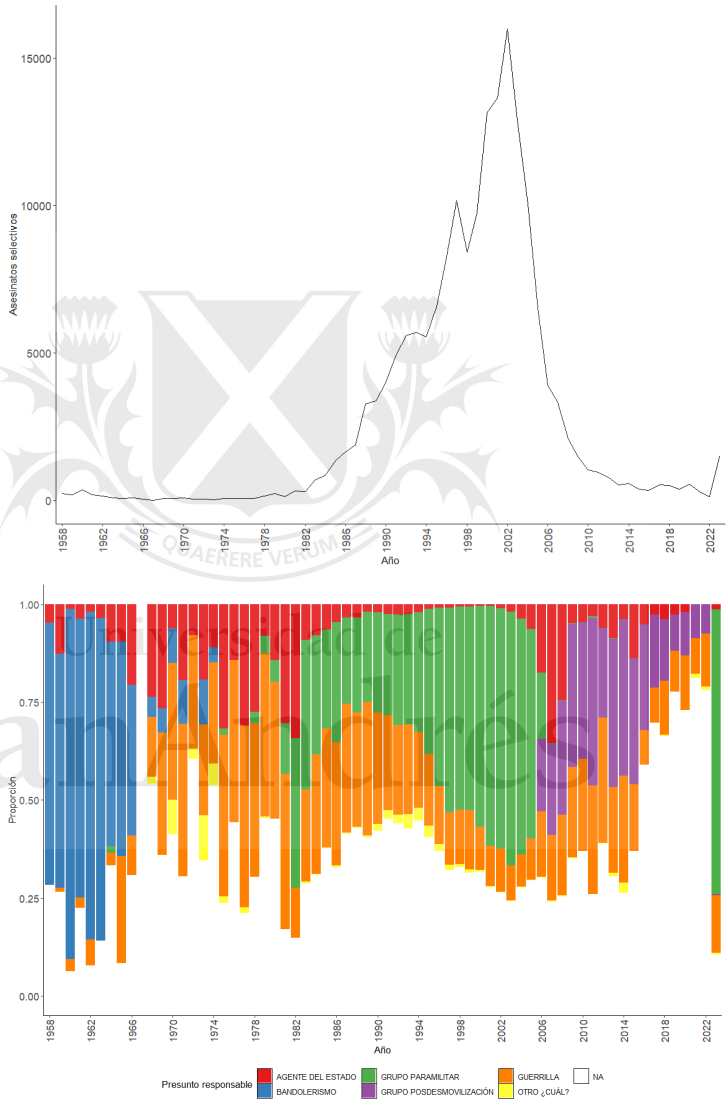

Figura 2: Asesinatos selectivos por año (arriba) y Proporción de responsabilidades por actor por año (abajo).

CAPÍTULO 1

Datos faltantes

Los datos faltantes pueden clasificarse en tres categorías según el mecanismo que genera su ausencia (Rubin, 1976). En primer lugar, los datos pueden ser faltantes completamente al azar (*missing completely at random* - MCAR) cuando el hecho de no tenerlos no está relacionado con datos observados o no observados. En segundo lugar, los datos pueden ser faltantes de forma aleatoria (*missing at random* - MAR) cuando su ausencia está relacionada con los datos observados. Por último, los datos pueden ser faltantes de forma no aleatoria (*missing not at random* - MNAR) si hay razones desconocidas detrás del mecanismo de omisión (van Buuren, 2018).

Al estudiar las violaciones de los derechos humanos, como los asesinatos, intuitivamente tiene sentido pensar que los datos faltantes no son completamente aleatorios. Esto se debe a que una víctima necesita ser documentada por una organización o institución. La probabilidad de que una víctima sea registrada en una base de datos se conoce como "probabilidad de captura", y probablemente esté relacionada con variables observables. Por ejemplo, el lugar en el que se produjo el homicidio, así como el año, ya que ambos están relacionados con la capacidad de la organización. La probabilidad de captura podría estar relacionada con las características de la víctima, pues algunas podrían tener más visibilidad que otras. Además, podría estar relacionada con la técnica utilizada por el perpetrador porque algunas técnicas pueden ser más difíciles de documentar. Por ejemplo, dos técnicas diferentes son el sicariato y la retención. El primero es un ataque perpetrado contra una víctima sin oportunidad de defenderse en un espacio público. La segunda es el ataque para retener a la víctima y llevarla a cometer un hecho victimizante en un lugar diferente, que termina con la ejecución. El autor de un sicariato podría ser más fácil de documentar, ya que, por definición, la victimización se produce en un espacio público.

1.1. Métodos de imputación

MissForest -imputación no paramétrica de datos faltantes para datos de tipo mixto- se propuso como un método no paramétrico para tratar variables continuas y categóricas simultáneamente (Stekhoven y Bühlmann, 2012). La motivación detrás del algoritmo es, en primer lugar, manejar tipos de datos mixtos y, en segundo lugar, hacer el menor número posible de suposiciones sobre los datos.

A diferencia de otros métodos, como k vecinos más cercanos, este algoritmo no requiere parámetros de ajuste. Además, no hace ninguna suposición sobre la distribución de los datos. Si se compara con Multiple Imputation by Chained Equation -MICE- (van Buuren, 1999) , por ejemplo, no es necesario asumir que existe una distribución multivariada completa que cubra el conjunto de datos (Stekhoven y Bühlmann, 2012).

Dado que el algoritmo de random forest requiere que las variables de respuesta no tengan datos faltantes para el entrenamiento del modelo, Stekhoven y Bühlmann (2012) proponen predecir los datos faltantes utilizando un random forest entrenado con los datos observados. Se asume que $\mathbf{X} = (X_1, X_2, ..., X_p)$ es una matrix de $n \times p$. Entonces, para una variable arbitraria $\mathbf{X}_s$ con datos faltantes $i^{(s)}_{\text{mis}} \subseteq \{1, ..., n\}$ proponen separar la base de datos en cuatro:

(I) $y^{(s)}_{\text{obs}}$, que representa los valores observados de la variable $\mathbf{X}_s$;

(II) $y^{(s)}_{\text{mis}}$, que representa los datos faltantes de la variable $\mathbf{X}_s$;

(III) $x^{(s)}_{\text{obs}}$, que representa las variables distintas a $\mathbf{X}_s$ con observaciones $i^{(s)}_{\text{obs}} = \{1, ..., n\} \setminus i^{(s)}_{\text{mis}}$;

(IV) $x^{(s)}_{\text{mis}}$, que representa a las variables distintas a $\mathbf{X}_s$ con observaciones $i^{(s)}_{\text{mis}}$.

Universidad de los Andes QUAERERE VERUM

El pseudoalgoritmo puede verse en el Algoritmo 1. Este consiste en un proceso iterativo que comienza utilizando la media/moda para completar los valores que faltan en $\mathbf{X}$. Una vez que se dispone de un conjunto de datos completo, las variables $\mathbf{X}_s$, $s = 1, ..., p$ se ordenan según su porcentaje original de datos faltantes, iniciando con el menor porcentaje. El algoritmo comienza ajustando un random forest a las variables observadas (y_{obs} y x_{obs}) para cada $\mathbf{X}_s$. Este modelo entrenado luego se utiliza en x_{mis} para predecir los datos faltantes y_{mis} . El proceso de imputación se repite hasta que se cumple un criterio de finalización γ, que consiste en llegar al máximo de iteraciones permitidas o al momento en el que aumente la diferencia entre la matriz más reciente y la matriz imputada aumenta por primera vez para las variables categóricas y continuas (Stekhoven y Bühlmann, 2012). La diferencia entre las matrices para el set de variables continuas $\mathbf{N}$ está dada por:

$$\Delta_N = \frac{\sum_{j \epsilon \mathbf{N}} (X^{imp}_{\text{new}} - X^{imp}_{\text{old}})^2}{\sum_{j \epsilon \mathbf{N}} (X^{imp}_{\text{new}})^2}$$

y para el set de variables categóricas $\mathbf{F}$ por:

$$\Delta_F = \frac{\sum_{j \epsilon \mathbf{F}} \sum_{i=1}^{n} I^2_{(X^{imp}_{\text{new}} \neq X^{imp}_{\text{old}})}}{\#NA}$$

donde $\#NA$ es el número de datos faltantes en las variables categóricas.

En este caso, se toma la penúltima matriz imputada, pues la última suele ser menos precisa que la anterior (Stekhoven, 2022).

Dados X, una matriz $n \times p$, y un criterio de finalización γ

1. Inicie el algoritmo utilizando como valores iniciales de los datos faltantes la media/moda de cada variable (X_{ini}^{imp});
2. **k** $\leftarrow$ vector de los índices ordenados de las columnas en **X** respecto a la cantidad de valores faltantes en aumento;
3. **mientras** (*while*) no se alcance γ **hacer (*do*)**
4. $X_{\text{old}}^{imp} \leftarrow$ guarde la matriz previamente imputada;
5. **para** (*for*) s **en** k **hacer (*do*)**
6. Ajuste un random forest: $y_{\text{obs}}^{(s)} \sim x_{\text{obs}}^{(s)}$;
7. Prediga $y_{\text{mis}}^{(s)}$ usando $x_{\text{mis}}^{(s)}$;
8. $x_{\text{new}} \leftarrow$ actualice la matriz imputada usando $y_{\text{mis}}^{(s)}$ predicha;
9. **termine el *for***
10. Actualice γ según el número actual de iteraciones y la diferencia entre las matrices.
11. **termine el *while***
12. Devuelva la matriz imputada X^{imp}.

Figura 1.1: Algoritmo 1: Pseudo algoritmo de MissForest. Traducción propia. Tomado de Stekhoven y Bühlmann (2012)

Universidad de San Andrés

En el artículo original, missForest se comparó con cuatro métodos en 10 conjuntos de datos diferentes con distintos tipos de datos (Stekhoven y Bühlmann, 2012). Para datos continuos, se comparó con KNN y mostró una reducción del error cuadrático medio normalizado medio (NRMSE) incluso superior al 50 % (Stekhoven y Bühlmann, 2012). Para variables categóricas, se comparó con MICE y KNN demostrando una reducción de la proporción de entradas clasificadas erróneamente (PFC) de hasta el 60 % (Stekhoven y Bühlmann, 2012). Por último, al utilizar los mismos métodos para comparar su rendimiento con datos de tipo mixto, se demostró que mejora el error de imputación (NRMSE/PFC) en algunos casos en más del 50 % (Stekhoven y Bühlmann, 2012).

Para estudiar el tipo de datos faltantes del presunto responsable utilizo los siguientes tres pasos en R:

1. Amputación: Eliminar las etiquetas de "perpetrador" en el conjunto de datos original.

2. Imputación: Imputar los datos faltantes eliminados en el paso 1 y los NA originales (estrategia "con NA originales") y sólo los datos faltantes generados (estrategia "sin NA originales").

3. Análisis: Calcular un puntaje global para evaluar el rendimiento de la imputación.

Amputación

Utilizo el algoritmo missForest para imputar el presunto responsable faltantes en la base de datos de asesinatos selectivos del CNMH. Para medir su rendimiento, elimino ("amputo") el 10 %, el 20 % y el 30 % de las etiquetas conocidas dentro de cada categoría de presunto responsable. Dado que los resultados serían sensibles a las etiquetas eliminadas, utilizo 100 semillas diferentes para eliminar distintos grupos de etiquetas. Hay que tener en cuenta que R mantiene las identificaciones eliminadas y sigue aumentando en función del porcentaje. Por ejemplo, con la semilla 1 y el 10 % de eliminación si los ID eliminados son 1, 27, 135, 4000 al eliminar el 20 % los ID eliminados son 1, 27, 135, 4000, y otros cuatro. Esto resulta relevante porque, en caso de que haya diferencia entre los porcentajes de amputación, habrá certeza de que los tres porcentajes incluyen al mismo grupo de IDs.

Imputación

El paquete missRanger, una implementación del algoritmo missForest hecha por Mayer (2023), se utiliza para imputar los valores que faltan en cada conjunto de datos. Utilizo dos estrategias diferentes para la imputación. En primer lugar, la estrategia "con NA originales" le entrega al modelo con todas las observaciones, incluidos los registros de los datos faltantes originales de los presuntos responsables, aunque no puedan utilizarse para medir el rendimiento del modelo. En segundo lugar, la estrategia "sin NA originales" elimina los registros con datos faltantes en el presunto responsable e imputa sólo las etiquetas amputadas.

Análisis

Las etiquetas imputadas se comparan con las originales y, a partir de ahí, se calcula la tasa de verdaderos positivos (*True Positive* en inglés - TP). Para comparar estas dos estrategias propongo el siguiente puntaje global

$$PG = \sum_{perp=1}^{6} w_{perp} * TP_{perp}$$

siendo w_{perp} el peso de cada categoría de perpetrador en la base de datos original y TP_{perp} el porcentaje de verdaderos positivos de cada categoría. Este puntaje toma el valor de 1 si las predicciones de todas las categorías corresponden a las etiquetas originales y 0 de lo contrario.

El PG está ponderado porque los datos están desbalanceados: hay más registros para los paramilitares que para otras categorías. Dado el desbalance, es importante que el algoritmo funcione mejor en las categorías mayoritarias.

CAPÍTULO 2

Resultados y limitaciones

2.1. Resultados

En el Figura 3 se presentan los puntajes globales de las estrategias con y sin NA originales. Las barras de error muestran una desviación estándar. Hay dos resultados principales. En primer lugar, como muestran las barras, el modelo predice mejor al perpetrador cuando se incluyen los registros con NA originales. La estrategia "con NA originales" tiene una media de 0,77, 0,77 y 0,76 para un 10 %, 20 % y 30 % de amputación, respectivamente. Mientras que, para los mismos porcentajes de amputación, los valores de la estrategia "sin NA originales" son 0,49, 0,5 y 0,51. Como puede verse, los intervalos entre ambas estrategias no se solapan, lo que significa que son estadísticamente diferentes. En segundo lugar, esto no está relacionado con el hecho de tener más datos, ya que el puntaje global no se reduce a medida que se eliminan más porcentajes de etiquetas: no hay una reducción del PG cuando se elimina el 30 % de las etiquetas conocidas frente al 10 %.

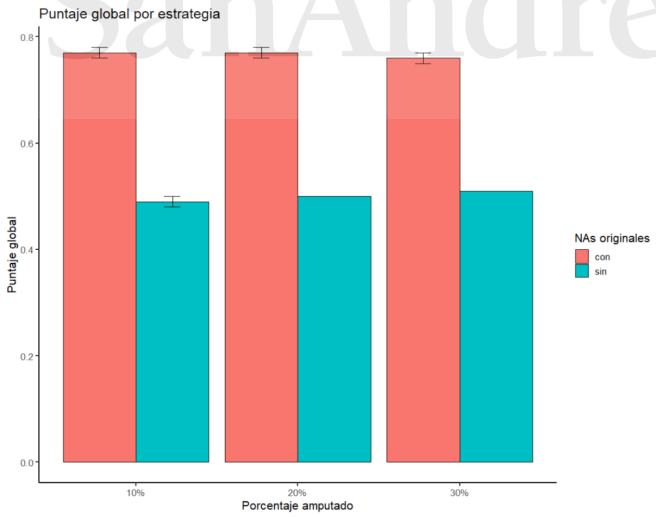

Figura 3: Puntajes globales de las estrategias "con NA originales" y "sin NA originales" para la variable de presunto responsable.

Universidad de San Andrés

Estos dos resultados tienen dos implicaciones. En primer lugar, demuestran que el método es coherente con diferentes proporciones de datos faltantes. En otras palabras, el algoritmo es estable. En segundo lugar, sugieren que el mecanismo de datos faltantes que subyace a los NA originales y a los NA creados completamente al azar es diferente. El hecho de que el puntaje global sea diferente en los datos con y sin las NA originales no está relacionado con la cantidad de datos: no hay diferencias estadísticamente significativas entre el 30 %, el 20 % y el 10 % de los datos amputados. Por lo tanto, debería estar relacionada con la distribución de las NA. En este sentido, el modelo sería capaz de identificar una distribución de los datos faltantes al recibir los NA originales. Por eso, esta estrategia tiene una puntuación más alta: los NA originales dan información adicional al modelo, que puede utilizar para predecir el autor. Por lo tanto, estaríamos ante un escenario en el que los datos faltantes del perpetrador son no son MCAR.

Para analizar las diferencias entre las estrategias con y sin NA construí una matriz para cada una de las imputaciones. Los Figuras 4 y 5 muestran las matrices como mapas de calor para facilitar su interpretación. El eje y muestra las etiquetas originales, mientras que el eje x presenta las imputadas. Si el algoritmo predijera correctamente el 100 % de las etiquetas amputadas, habría solamente una diagonal oscura con 100 en cada celda. Las figuras muestran los resultados para una misma semilla para las dos estrategias diferentes. Como puede verse, cuando se eliminan los registros con NA originales en el presunto responsable hay un sesgo hacia paramilitares, que es la categoría mayoritaria. Pero, cuando el modelo recibe los registros con NA originales el sesgo se reduce. Aunque en la vida real no habría razón para amputar a los autores conocidos, esto demuestra que el método missForest está sesgado hacia la categoría mayoritaria, un problema común con los árboles de decisión. Sin embargo, el sesgo se reduce cuando el algoritmo puede aprender de la distribución original de los NA.

Al estudiar las responsabilidades que tendrían los diferentes perpetradores según el algoritmo missRanger (la implementación de missForest), queda clara la importancia de utilizar métodos de imputación (Figura 6). Para el principal perpetrador, los paramilitares, hay grandes cambios. Si el porcentaje se calcula incluyendo los NA, su responsabilidad sería del 43,56 %, mientras que con la eliminación por lista sería del 65,12 %, y por imputación por missForest sería del 71,66 %. En cuanto a la guerrilla, sus responsabilidades serían del 15,83 %, 23,66 % y 18,18 %, respectivamente. Al estudiar el Estado, los porcentajes son 3,44 %, 5,14 % y 4,44 %, respectivamente. Mientras que para los grupos posdesmovilización sus responsabilidades serían del 2,57 %, 3,84 % y 4,13 %. Las responsabilidades de "otros" grupos son del 0,91 %, 1,36 % y 0,91 %. Por último, para los bandoleros las responsabilidades son del 0,58 %, 0,87 % y 0,67 %.

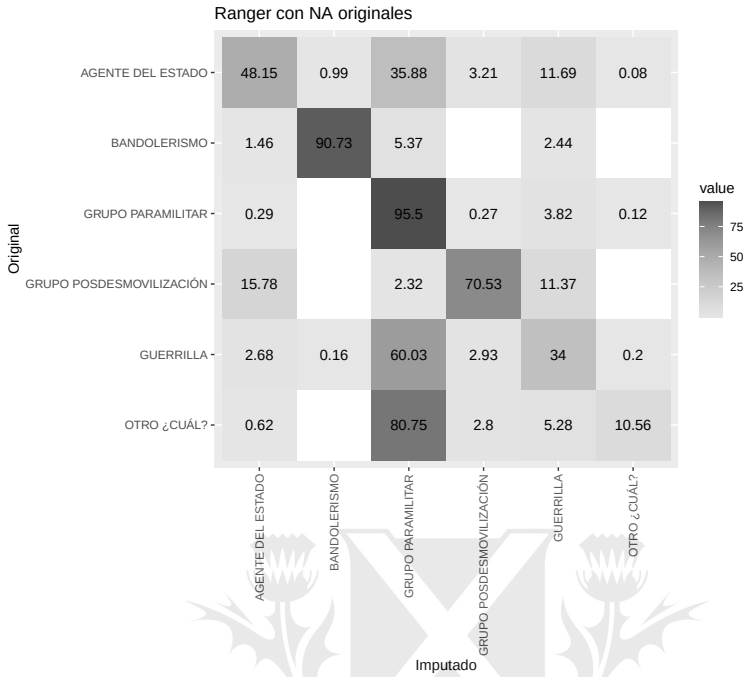

Figura 4: Mapas de calor de etiquetas originales vs. imputadas para semilla 1234 con el 20 % de amputación con NA originales.

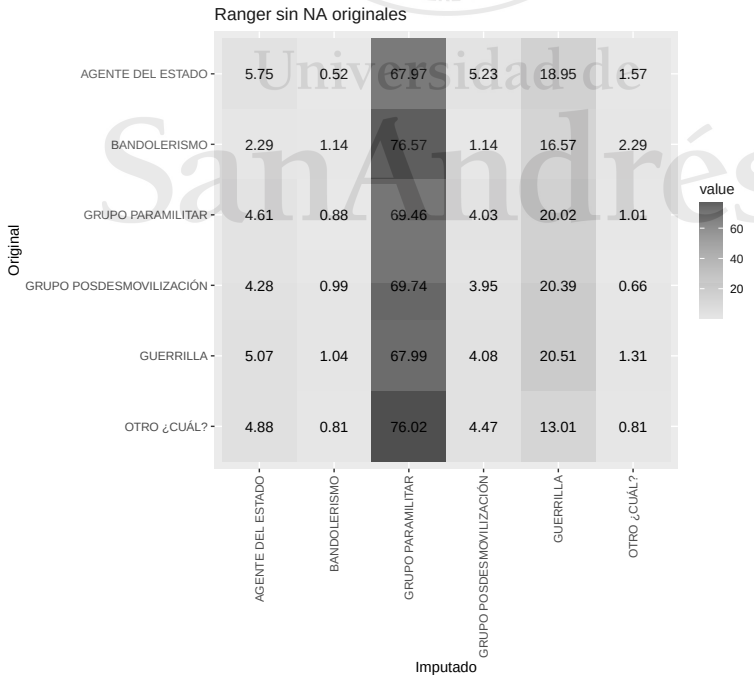

Figura 5: Mapas de calor de etiquetas originales vs. imputadas para semilla 1234 con el 20 % de amputación sin NA originales.

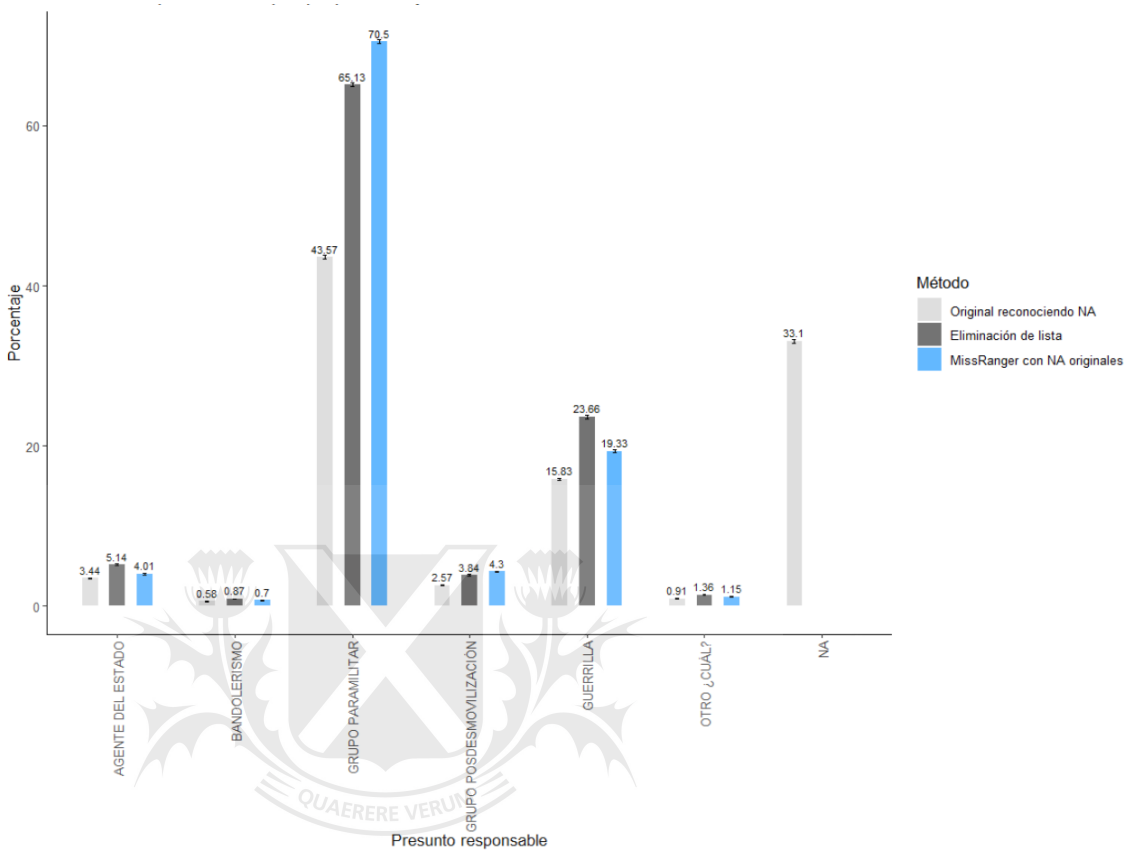

Figura 6: Presuntas responsabilidades por responsable y método.

2.2. Limitaciones

Existen dos tipos de limitaciones en esta investigación. En primer lugar, las relacionadas con el método. En segundo lugar, las relacionadas con los datos.

En cuanto a las relacionadas con el método, la estrategia que utilicé para evaluar la imputación consiste en introducir las omisiones de forma totalmente aleatoria. Sin embargo, probablemente en la vida real los presuntos responsables no sean de este tipo. Aunque el mecanismo no está claro, podría ser un escenario de MAR y estar relacionado con el municipio y el año en que tuvo lugar el homicidio. Si este fuera el caso, los datos faltantes podrían explicarse por la capacidad de las diferentes instituciones para documentar la violencia en el territorio a lo largo de los años o por variables como el sexo de la víctima. O también podría estar relacionado con la técnica o el arma. También podría ser una combinación de todo lo anterior o de algunas variables específicas. En un trabajo futuro, sería recomendable incluir diferentes mecanismos de MAR y replicar el análisis.

En cuanto a los datos, la probabilidad de captura es heterogénea cuando se estudian las violaciones de los derechos humanos. Aunque se intente documentar

a todas las víctimas, hay algunas que son más visibles que otras. Los resultados aquí encontrados no deben entenderse como la responsabilidad de los perpetradores en el contexto del conflicto armado en Colombia. Por el contrario, deben tomarse como la responsabilidad de los perpetradores en el contexto del conflicto armado de Colombia con base en los datos del OMC y el algoritmo missForest.

Para futuras investigaciones, sería recomendable estudiar otros conjuntos de datos como el RUV. También, se sugeriría comparar otros métodos de imputación y discutir los rendimientos con base también en la capacidad de incluir la incertidumbre, como lo hace el algoritmo de MICE.

QUAERERE VERUM
Universidad de
SanAndrés

CAPÍTULO 3

Conclusiones

Los resultados del algoritmo missForest con y sin NAs sugieren que los datos faltantes del perpetrador no son MCAR. Este resultado complementa el conocimiento que los académicos y defensores de derechos humanos han generado con base en el trabajo con víctimas, en el que se ha explicado cómo la falta de información corresponde a lógicas como la presencia del actor en el territorio o el miedo de las víctimas.

Cuando se utiliza missForest para imputar el presunto responsable de homicidios en Colombia según los datos del CNMH, las responsabilidades varían en comparación con el uso del método tradicional de eliminación de la lista. Utilizando este último, el principal responsable serían los paramilitares con un 65 %, seguidos de la guerrilla con un 18 %, los grupos posdesmovilización con un 8 %, el Estado con un 7 % y los bandoleros con un 2 %. Cuando se compara con el método de eliminación de la lista, la responsabilidad de los paramilitares basada en missForest aumenta del 65 % al 71 %, mientras que la de la guerrilla disminuiría del 23 % al 18 %. Sin embargo, estos resultados no son definitivos. Podrían variar si se utilizara una fuente diferente. Además, no consideran todo el universo de víctimas, sólo las documentadas por el CNMH.

Universidad de SanAndrés

Bibliografía

Ball, Patrick, Jana Asher, David Sulmont, and Daniel Manrique. 2003. *How many Peruvians have died? An estimate of the total number of victims killed or disappeared in the armed internal conflict between 1980 and 2000.* New York: American Association for the Advancement of Science.

Buuren, Stef van. 2018. *Flexible Imputation of Missing Data*, Second Edition. Chapman Hall.

CEV, JEP, HRDAG. 2022. *Informe metodológico del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística.* Consultado el 13 de diciembre de 2022. https://hrdag.org/wp-content/uploads/2022/08/20220818-fase4-informe-corrected.pdf

Centro Nacional de Memoria Histórica. sin fecha - a. Observatorio de Memoria y Conflicto - Nosotros. Consultado el 1 de marzo de 2023. https://micrositios.centrodememoriahistorica.gov.co/observatorio/nosotros/preguntas-frecuentes/.

Centro Nacional de Memoria Histórica. sin fecha - b. Observatorio de Memoria y Conflicto - Preguntas frecuentes. Consultado el 10 de marzo de 2023. https://micrositios.centrodememoriahistorica.gov.co/observatorio/nosotros/.

Congreso de Colombia. 2011. "Ley 1448 de 2011." *Decreto Número 4803.* 2011. (Ministerio de Justicia y del Derecho de la República de Colombia).

Gutiérrez-Sanín, F., Wood, E.J. 2019. *Cómo debemos entender el concepto de "patrón de violencia política": repertorio, objetivo, frecuencia y técnica.* Revista Estudios Sociojurídicos, 22(1), 13-65. Doi: http://dx.doi.org/10.12804/revistas.urosario.edu.co/sociojuridicos/a.8211

Mayer, Michael. 2023. 'Package 'missRanger'. Version 2.2.1." CRAN. URL: https://cran.r-project.org/web/packages/missRanger/missRanger.pdf

Rubin, Donald B. 1976. "Inference and Missing Data." Biometrika, Vol. 63, No. 3 581 - 592.

Universidad de San Andrés

Stekhoven, Daniel J., and Peter Bühlmann. 2012. "MissForest—non-parametric missing value imputation for mixed-type data." Bioinformatics, Volume 28, Issue 1 112-118.

Stekhoven, Daniel J. 2022. "MissForest—non-parametric missing value imputation for mixed-type data." Version 1.5" CRAN. URL: https://cran.r-project.org/web/packages/missForest/missForest.pdf

van Buuren, Stef. 1999. "Flexible Multivariate Imputation by MICE." TNO Prevention and Health, Volume PG/VGZ/99.054.

QUAERERE VERUM
Universidad de
SanAndrés