DANE
INFORMACIÓN PARA TODOS

# DCD - DIRPEN

# CONCEPTO TÉCNICO
# INFORME COMISIÓN DE LA VERDAD

**JULIO de 2022**

El futuro es de todos
Gobierno de Colombia

# CONTENIDO

## Lista de tablas



## Lista de gráficos

# 1. Introducción

El Departamento Administrativo Nacional de Estadística (DANE), como ente rector del Sistema Estadístico Nacional (SEN), tiene como función definir lineamientos, estándares y normas técnicas para la producción de estadísticas oficiales de sus miembros, así como para un mejor aprovechamiento de los registros administrativos y otras fuentes de datos para la producción de información estadística.

En línea con lo anterior, el DANE ha puesto a disposición tres documentos pilares para la producción estadística, estos son el Código Nacional de Buenas Prácticas, una herramienta para establecer el deber ser de la producción estadística en el país; los Lineamientos para el Proceso Estadístico en el Sistema Estadístico Nacional en su versión 2.0, documento creado para orientar a las dependencias del DANE y las entidades del SEN en prácticas estandarizadas dentro de la producción estadística; y la Norma Técnica de Calidad del Proceso estadístico NTC PE 1000 2020, que establece los requisitos que se evalúan para garantizar la calidad del proceso estadístico.

El presente documento contiene el análisis de los soportes presentados en el marco del "proyecto conjunto JEP[1]-CEV-HRDAG[2] de integración de datos y estimación estadística" por la Comisión de la Verdad (CEV). Estos fueron revisados por el DANE, y si bien se reconoce que de momento no aplican los marcos de calidad establecidos para las estadísticas oficiales del SEN, pueden ser tenidos en cuenta como un referente de buenas prácticas. Esto permite al DANE generar una opinión técnica frente al ejercicio presentado y documentado por la CEV.

De esta forma, se pretende aprovechar la experiencia del DANE en la producción de estadísticas, la cual permite articular en diferentes fases los procedimientos que llevan a la obtención de análisis y resultados de estas. Esto facilita la toma de decisiones informadas por parte de las entidades encargadas del diseño y seguimiento a la política pública. Por consiguiente, el objetivo del presente documento es emitir un concepto técnico, sobre la calidad, metodología y los resultados de las estimaciones sobre hechos victimizantes que la CEV ha implementado y desarrollado durante el período de su gestión.

# 2. Principios del proyecto y garantía de la calidad

[1] Jurisdicción Especial para la Paz
[2] Human Rights Data Analysis Group, Statisticians For Human Rights

A continuación, se valoran tanto el objetivo del proyecto, como los principios que orientaron el desarrollo metodológico y contribuyeron a la garantía de calidad en el proceso y sus resultados.

Los cuatro principios implementados (transparencia, auditabilidad, reproducibilidad, escalabilidad) garantizan la estandarización y replicabilidad de las actividades, así como la revisión de pares en el desarrollo y en los resultados de estas actividades. En un proyecto que implica la integración de un número considerable de fuentes secundarias, estos principios son claves para garantizar la trazabilidad y minimizar el riesgo de error, a partir de la automatización de todas las acciones. Así mismo, se evidencia la aplicación de buenas prácticas de documentación y autorregulación. Es importante anotar que, a partir de la documentación compartida no es posible para el DANE emitir una opinión técnica frente a la aplicación total de estos principios en los algoritmos del proyecto, o validar la reproducibilidad a partir de la realización de cálculos espejo. Sin embargo, si es claro que estos orientan el detalle de la documentación presentada en los informes de las fases 3 y 4.

A partir de la documentación recibida y el horizonte temporal de la revisión, se emite una opinión técnica soportada en la vinculación de más de 111 bases de datos relacionadas con el conflicto, teniendo como referente las orientaciones brindadas por el DANE para el aprovechamiento de registros administrativos y la revisión de calidad de los datos. La metodología aplicada para las estimaciones de víctimas (integración, tratamiento de duplicados, generación de bases de entrenamiento a partir de conocimiento experto, imputación, criterios para toma de decisiones frente a los resultados de las estimaciones) tiene como referente el estándar para la producción de información estadística GSBPM[3], por sus siglas en inglés. Finalmente los resultados se revisan en contraste con las Estadísticas Vitales.

Teniendo en cuenta lo anterior, a continuación, se presentan los datos generales del proyecto:

> Nombre del proyecto: Integración de datos y estimaciones estadísticas de víctimas en el marco del conflicto armado.
>
> Propósito: "Dotar a las entidades del Sistema Integral de Verdad, Justicia, Reparación y No Repetición (SIVJRNR) de argumentos científicos sólidos para develar la magnitud de la violencia, mediante la estimación del subregistro y la identificación de patrones de victimización de homicidio, desaparición forzada, secuestro, desplazamiento forzado y reclutamiento ilícito."[4]

[3] Generic Statistical Business Process Model
[4] Tomado de "Informe metodológico del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística"

### 2.1. Principios expresados dentro del documento entregado por la CEV al equipo técnico del DANE con nombre "CO-Apr2022-panorama"

- **Transparencia**

"el código debe ser legible por seres humanos, escrito de acuerdo con una estricta guía de estilo (según el idioma) y organizado en directorios bien nombrados y estandarizados y guiones breves."

- **Auditabilidad**

"El código debe organizarse para permitir que los programadores dentro del equipo y las personas externas realicen pruebas. Cada paso debe estar separado y claro. Toda transformación de los datos debe documentarse en código; Los pasos interactivos están prohibidos."

- **Reproducibilidad**

"El código debe ejecutarse en cualquier computadora con los idiomas, sistemas de archivos, sistema operativo y variables de entorno especificados. No debe haber dependencias en la configuración de ningún usuario específico. Cada ejecución en la computadora de cualquier usuario debe producir salidas binarias idénticas. Todas las dependencias y parámetros de cada tarea deben registrarse en un Makefile."

- **Escalabilidad**

"El código debe organizarse para adaptarse a cualquier idioma compatible con canalización, cualquier número de entradas, cualquier número de actualizaciones, de forma totalmente automatizada."

- **Documentación soporte recibida**

En el marco del proceso de revisión, el DANE recibió documentación de referencia para el proyecto: reportes sintéticos de las bases de datos consideradas, documentos específicos para actividades del flujo del proceso (como el record linkage), imputación de valores faltantes, MSE, así como las versiones de las fases 3 y 4 del documento *'Informe metodológico del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística'*, ciento once (111) reportes y ejemplos de información de cada fuente empleada, y por último, los resultados preliminares de los cinco ámbitos objetivo homicidio, desaparición forzada, secuestro, desplazamiento forzado y reclutamiento ilícito.

A partir de la documentación tipo parámetro (metodología), tipo registro (reportes de base de datos) y tipo código (sintaxis de los algoritmos usados) del proyecto se elabora el concepto técnico por parte del DANE, resaltando que la entidad no tuvo acceso a las bases de datos generadas en cada fase del proyecto por lo que no fue posible realizar ejercicios de cálculo tipo espejo para validar y replicar los resultados obtenidos.

# 3. Consideraciones de calidad desde el aprovechamiento de registros administrativos

En este apartado se contrastan los elementos presentados en el "Informe metodológico del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística" con los parámetros establecidos en la Metodología de Diagnóstico de Registros Administrativos para su Aprovechamiento Estadístico propuesta por el DANE en 2018[5]. Este contraste busca identificar si en el proyecto se consideraron aspectos básicos para el aprovechamiento de los registros administrativos, también se listan las evidencias documentales de soporte que fueron compartidas, y por último, se menciona qué elementos o aspectos de análisis quedarían pendientes por aclaración por parte del equipo investigador.

La metodología de diagnóstico de registros administrativos para su aprovechamiento estadístico se construyó con el objetivo de "Suministrar a las entidades del SEN una herramienta de diagnóstico que les permita determinar las condiciones y las características de los registros administrativos para su aprovechamiento, frente a un potencial uso estadístico identificado". Si bien el objetivo del proyecto realizado por la JEP-CEV-HRDAG no tiene como alcance directo la realización de un diagnóstico exhaustivo de cada una de las fuentes de información empleadas, dicho proyecto si tiene como finalidad el aprovechamiento estadístico de diferentes registros administrativos. A su vez, busca el fortalecimiento de las actividades realizadas para la captación de información, para que respondan a los criterios de calidad, completitud y cobertura propuestos por la metodología de diagnóstico del DANE.

De acuerdo con los requerimientos que deben cumplir estas fuentes de información, se realiza el diagnóstico de registros administrativos, considerando los elementos y procedimientos llevados a cabo de por JEP-CEV-HRDAG, y verificando su abordaje por el equipo investigador. Es importante aclarar que no se contempla ni se validan otros aspectos metodológicos relacionados con:

- la implementación de modelos de clasificación.
- la de duplicación de datos.
- la imputación de datos.
- los modelos de estimaciones por sistemas múltiples.

Por lo que la actividad se enfoca principalmente en la caracterización de los registros administrativos a los que el proyecto tuvo acceso, los aspectos mencionados anteriormente harán parte de esta sección.

[5] Ver https://www.sen.gov.co/files/RegistrosAdministrativos/Metodologia_Diagnostico_Registros_Administrativos.pdf

Entre los usos más comunes identificados por la Dirección de Regulación, Planificación y Estandarización- DIRPEN – del DANE, para los registros administrativos, se encuentran la construcción de marcos o directorios estadísticos, el soporte de operaciones estadísticas y realización de contrastes. Esta utilidad podrá precisarse en los 9 aspectos relacionados en la Tabla 1.

**Tabla 1. Utilidad estadística**

| Aspectos para la utilidad estadística | Evidencia a partir de la documentación compartida | Observaciones frente a la documentación compartida |
|---|---|---|
| **1.** Fundamento legal que garantiza la producción continua del registro. | Las fuentes empleadas cuentan con un sustento legal que las faculta en el levantamiento de información relacionada con el proyecto. En el anexo A. se presentan cada una de las entidades consultadas y las bases de datos empleadas, con algunos datos estadísticos que dan cuenta de: la cobertura temporal, el corte de la información, el número registros presentes en la base, el número de registros procesados y la clasificación temática asignada a la fuente de datos (conflicto armado, Violencia sociopolítica). | |
| **2.** Las variables temáticas, de identificación y de ubicación están contenidas en el listado de las variables del registro. | Se tuvo acceso a 111 archivos planos con la descripción general de cada archivo de datos empleado en el proyecto. Se destaca que en estos archivos se encuentra la lista de las variables presentes en cada uno, a partir de estas se puede identificar la presencia de variables de ubicación general (departamento y municipio) y otras variables de ubicación de mayor detalle (corregimiento, vereda, barrio, zona, etc.). Se aprecia también la inclusión del conteo de registros no nulos, lo que permite tener una aproximación a la calidad de las variables, ya que, en algunos casos una variable puede estar presente en la base, pero todos sus valores pueden ser nulos. La evidencia entregada permite comprobar que estos chequeos fueron considerados. | Si bien uno de los objetivos del proyecto es la integración de las bases de datos, no se encontró dentro de los soportes compartidos la correlativa de variables construida para el proceso de integración.<br><br>Incorporar como parte de los anexos esta correlativa daría cuenta de la estandarización previa que se requiere para la vinculación exitosa de las bases y va en línea con la aplicación del principio de reproducibilidad. |
| **3.** La definición de las variables concuerda con los conceptos requeridos para satisfacer la necesidad de información. | En el proyecto se pueden identificar variables clave para la implementación de los diferentes procesos de analítica planteados (vinculación de registros, la estimación de los modelos de duplicación, reclasificación de eventos, imputación de datos y estimación de sistemas múltiples).<br><br>En este sentido, las variables requirieron de revisión previa a partir de los diccionarios de datos o contacto con las fuentes para la verificación de los conceptos clave abordados. Por ejemplo: nombre de la víctima, apellido de la víctima, departamento del hecho victimizante, año del hecho victimizante, tipo de violaciones (homicidio, | En la introducción del documento se afirma lo siguiente: "P*or tanto, cuando se podía obtener completa claridad de la relación de la información entregada con el conflicto, por parte de la entidad u organización social, se estableció relación de este conjunto de datos con el conflicto armado. En casos donde existía ambigüedad en la relación de la información con el conflicto interno, se implementó un esquema de* |

| | | | |
|---|---|---|---|
| | | desaparición forzada, secuestro, reclutamiento ilícito, desplazamiento forzado, exilio). | *imputación de esta relación con base en la información de esas bases."*<br><br>Bajo esta afirmación se asume un proceso de validación de las fuentes y las diferentes variables involucradas, sin embargo, no es explicito en el documento el resultado de dicha validación. |
| **4.** | Las variables de interés cumplen con estándares nacionales e internacionales. | En el documento se mencionan variables que pueden estar bajo revisión de clasificaciones nacionales e internacionales, básicamente nos referimos al manejo de la Divipola (municipio, departamento) y las tipologías de violencia. | No se hace mención de las verificaciones realizadas en cuanto al manejo de la Divipola en el documento, y tampoco sobre un estándar internacional para la clasificación de violencia (homicidio, desaparición forzada, secuestro, reclutamiento ilícito, desplazamiento forzado, exilio) |
| **5.** | De acuerdo con las validaciones establecidas por la entidad, los resultados de los indicadores de consistencia y completitud de las variables son adecuados para su uso estadístico. | Respecto a los indicadores de completitud dentro del informe y los anexos compartidos se evidencia un análisis realizado en términos de revisión de valores nulos presentes en la base de datos.<br><br>En cuanto a consistencia, no encontramos indicadores explícitos relacionados con el chequeo de los diferentes dominios asociados a cada variable, sin embargo, dentro del documento se menciona la realización de análisis de frecuencia previo para las bases de datos. Para el caso de algunas variables como: nombre y apellido, se observan, en algunas de las presentaciones anexas, los chequeos realizados a la longitud y frecuencia de las letras presentes en dichas variables. Esto se encuentra asociado con el levantamiento de indicadores de consistencia parcial sobre las bases de datos.<br><br>Otros indicadores de completitud que se deberían abordar en las etapas previas corresponden a los chequeos en términos de: completitud relacionada a los hechos victimizantes y la consistencia temporal. No encuentran indicadores en el documento sobre estos chequeos. | |
| **6.** | La cobertura del registro cumple con los requerimientos para la | Para cumplir el objetivo del proyecto se realizó la vinculación de 115 registros administrativos relacionados con conflicto, evidenciando la exhaustividad del proceso de recolección de datos relacionados con el fenómeno de estudio. A partir de esta base se llevan a cabo los procesos | |

| | | | |
|---|---|---|---|
| | producción estadística. | de gestión de duplicados con el fin de consolidar la base de victimas para el periodo de referencia.<br><br>Así mismo, se reconocen como parte de los desafíos del proyecto los datos faltantes y el subregistro. Sin embargo se establecen chequeos de calidad como la construcción de una matriz que presenta el listado de víctimas e identifica las fuentes (listas) en las que aparece.<br><br>En línea con lo anterior se aplica el método de estimación por sistemas múltiples (ESM) o "captura-recaptura", para determinar la población total de victimas y aproximarse al subregistro de la base integrada de víctimas resultante del proyecto. | |
| **7.** | La periodicidad del registro satisface la necesidad de información. | En el Anexo A. Datos, se presenta información para cada una de las bases de datos en términos de la fecha de corte de la base y los periodos que la fuente está documentando, estos cubren el periodo de interés del estudio (1990-2020). | |
| **8.** | La población objetivo del registro coincide con la población que se quiere estudiar. | La población capturada en las diferentes fuentes de datos es exhaustiva y evidencia la orientación del proyecto a garantizar el mayor número de registros asociados con el conflicto posible, que según se reportó en las presentaciones, permitió conformar una base de datos inicial de 26 millones de registros.<br><br>En el documento se hace especial énfasis en la revisión realizada con las entidades u organizaciones sociales para identificar si los datos suministrados se podían considerar como "relacionados con el conflicto armado". En caso de existir ambigüedad se plantea una metodología de imputación de datos que subsane los vacíos encontrados. | |
| **9.** | Las unidades de observación (o fuentes de información) del registro son de las que se requiere información. | Se considera que la unidad de observación del proyecto corresponde a las víctimas y los hechos victimizantes[6], sobre estos se obtiene la información a partir de la vinculación de 115 fuentes. | |

6 Una persona puede ser víctima de varios hechos violentos

Fuente: **DANE (2018**). DIRPEN a partir del proceso de diagnóstico de registros administrativos

En la revisión de la Base de Datos, se tiene en cuenta la consistencia de la información y se abarcan los siguientes aspectos:

1) Verificar que la definición de las variables incluidas en el diccionario de la base de datos coincida para todas las variables contenidas en la misma; teniendo en cuenta las variables: nombre, tipo, longitud, valores de dominio, descripción, entre otros.
2) Revisar la consistencia de las variables del registro, a partir del cálculo de indicadores de calidad para cada una de las variables de la base de datos.

Estos indicadores ayudan a identificar si la base de datos del registro tiene en cuenta aspectos como: completitud de la información, consistencia entre variables, coherencia entre el diccionario de la base de datos y la base de datos, y finalmente, el cumplimiento de estándares y clasificaciones nacionales e internacionales. La tabla 2 muestra los indicadores requeridos en la revisión de bases de datos, sin embargo, como se ha mencionado anteriormente no se tuvo acceso a las bases de datos usadas por JEP-CEV-HRDAG por lo cual no es posible replicar los valores de los indicadores mencionados.

**Tabla 2 Utilidad estadística, ficha de revisión bases de datos**

| Aspectos para la utilidad estadística | | Evidencia encontrada | Observaciones |
|---|---|---|---|
| **1.** | Porcentaje de campos cuyo tipo de variable corresponde al reportado en el diccionario de base de datos | No se cuenta con información. | En las presentaciones de resultados provistas se resaltó la carencia de diccionarios de datos en la mayoría de los registros administrativos empleados en el proyecto de medición. |
| **2.** | Porcentaje de campos cuya longitud es menor a la longitud máxima reportada en el diccionario de base de datos. | No se cuenta con información. | En las presentaciones de resultados se resaltó la carencia de diccionarios de datos en la mayoría de los registros administrativos empleados en el proyecto de medición. |
| **3.** | Porcentaje de campos que están dentro de los valores permitidos en la clasificación o nomenclatura definidos | No se cuenta con información. | |
| **4.** | Porcentaje de completitud en las variables de respuesta obligatoria | No se cuenta con información. | De los 111 anexos se puede realizar una inferencia de estos porcentajes por variable, tomando como insumos los conteos de registros no nulos frente al total de registros presentes en cada base de datos. En este punto se requiere validar si todos los campos son obligatorios en la base. |

| | | | |
|---|---|---|---|
| **5.** | Porcentaje de campos con valores dentro del dominio | No se cuenta con información. | |
| **6.** | Porcentaje de campos que cumple las reglas de validación temáticas | No se cuenta con información. | |

Fuente: DIRPEN a partir del proceso de diagnóstico de registros administrativos

Con respecto al referente de buenas prácticas en la gestión de la calidad de datos de registros administrativos, se destaca la articulación y gestión con las entidades a cargo de registros administrativos relacionados con conflicto en el país, que ha permitido identificar y vincular 115 bases de datos, según se reporta en la documentación entregada. En este punto es importante anotar que se enviaron 111 reportes de calidad de bases de datos, esto evidencia la exhaustividad de las bases compiladas y permite gestionar uno de los principales desafíos del proyecto, como lo es el subregistro.

En línea con lo anterior, la generación de reportes para cada una de las fuentes con la información general de las variables de la base y el número de registros no nulos confirma la realización de chequeos generales de calidad.

Así mismo, se reconoce el desafío que implica la falta de estandarización de variables clave para la vinculación de los registros administrativos (nombres, apellidos, tipo y número de documento, etc). Sobre este punto, se podría enriquecer la información documental anexa con las correlativas entre las variables de las fuentes originales y las de la base integrada de víctimas integrada.

Teniendo en cuenta la importancia de las variables de ubicación para la desagregación geográfica de la información, se podría precisar cómo se conformo esta variable a nivel de municipios, departamentos y regiones con el fin de facilitar la aplicación del principio de replicabilidad.

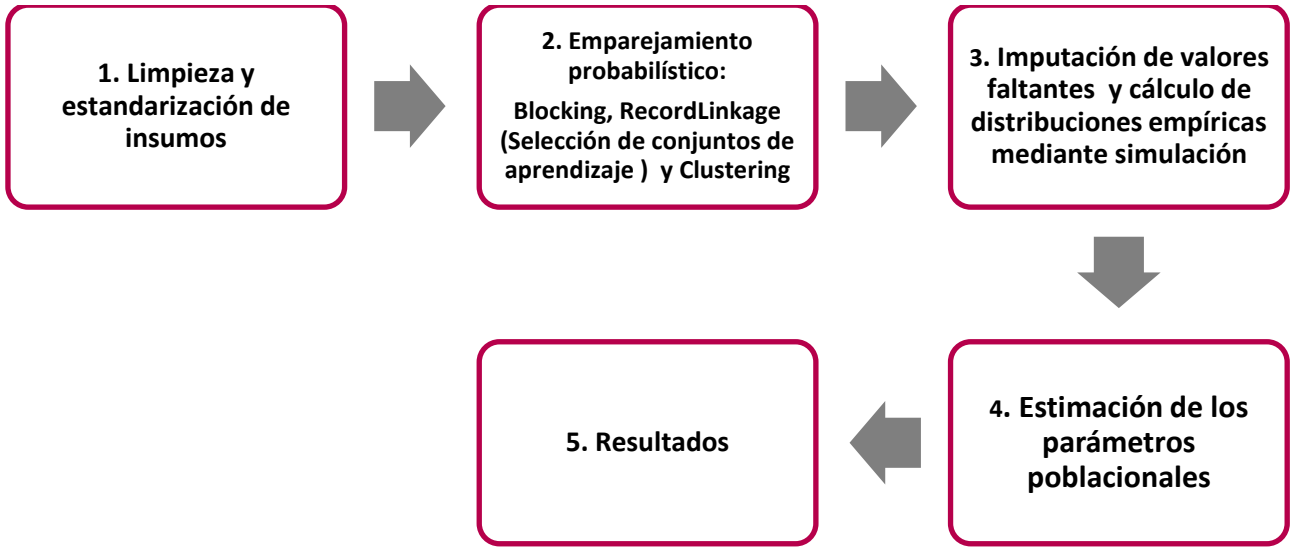
Como lección para el DANE, a partir de las limitaciones expresadas en el informe de la Comisión de la Verdad, se hace evidente la necesidad de seguir realizando esfuerzos con las entidades del SEN, para incentivar la implementación de las herramientas del Programa de Fortalecimiento de Registros Administrativos, lo que puede generar una mayor documentación de los registros administrativos, tanto a nivel de metadatos como de microdatos, así como la adopción de los estándares, nomenclaturas y codificaciones dispuestas por el DANE.

## 4. Aspectos metodológicos analizados

## Análisis de la metodología

En el reporte metodológico, junto con los documentos de investigación y soporte provistos a los profesionales y técnicos del DANE, se muestra un panorama sobre los principales aspectos metodológicos aplicados y resultados obtenidos. Lo anterior permitió la identificación del flujo del proceso como se muestra en la ilustración 1.

**Ilustración 1: Flujo proceso de estimación**




Fuente: Elaboración propia DANE

A continuación, se presenta una breve descripción de cada uno de los pasos ilustrados en el diagrama de flujo:

1. Limpieza y estandarización de insumos: Es una de las tareas básicas en cualquier proceso de producción estadística basado en registros administrativos, resulta imprescindible, en particular cuando los datos no son generados por una única fuente. Su implementación no cambia la información contenida en los conjuntos de datos ya que su objetivo es generar estructuras que puedan ser automatizadas, remover palabras vacías[7], y remover acentos y signos de puntuación. La aplicación de estas técnicas está ampliamente documentada y recomendada[8] .

[7] Palabras que carecen de sentido por si solas, pueden ser artículos, pronombres, preposiciones, etc.
[8] Ver Christen, Peter. 2012. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. New York: Springer.

2. Emparejamiento probabilístico: Al disponer de 111 fuentes de información provenientes de distintas organizaciones, es natural que se presenten errores de digitación o declarativos en las variables de identificación, siendo uno de los casos extremos la ausencia de dichas variables. Por lo tanto, la simple tarea de individualizar una persona dentro de las fuentes de información resulta en uno de los mayores desafíos del proyecto. Adicionalmente el impacto de los duplicados en la estimación de los fenómenos estudiados podría invalidar cualquier resultado, por tanto, la depuración de duplicados es fundamental para garantizar la calidad de las estimaciones. Resulta inviable computacionalmente comparar a todas las parejas posibles contenidas en los registros para determinar si son la misma persona, por lo que agrupar los individuos para realizar comparaciones resulta ser la única alternativa.

   A diferencia del proceso de limpieza y estandarización, el emparejamiento tiene un impacto directo en los resultados (la selección del modelo y conjunto de aprendizaje adiciona incertidumbre a las estimaciones finales, por lo que aplicar modelos de emparejamiento y selección de conjuntos de entrenamiento diferentes podrían arrojar estimaciones diferentes), ya que no es posible afirmar con toda certeza que un par de registros correspondan a la misma persona, ya sea por la agrupación o por la escogencia del modelo que vincula los registros.

3. Imputación de valores faltantes y cálculo de distribuciones empíricas mediante simulación: Dada la sensibilidad del tema del conflicto armado en Colombia, es de esperarse que la recolección de información relacionada con el fenómeno presente faltantes en muchas de las variables. Por lo tanto, para lograr una estimación con mayor robustez estadística que se aproxime al valor poblacional real, resulta relevante completar la información ausente dejando de manifiesto que al igual que proceso de emparejamiento, la imputación de valores faltantes repercute de forma directa en los resultados.

4. Selección de conjuntos de aprendizaje y estimación de los parámetros poblacionales: El resultado de los pasos anteriores constituye un insumo para los modelos de predicción de los parámetros poblacionales. Las técnicas implementadas están bajo la supervisión humana, por lo que la selección del conjunto de aprendizaje y prueba afectan el resultado final (cabe resaltar que los investigadores documentan las pruebas estadísticas para validar los resultados). Por lo tanto, la selección del mejor conjunto de aprendizaje y los resultados idóneos se realizan bajo el precepto de minimizar el error[9] en el conjunto de prueba, este procedimiento se logra mediante el uso de simulación para el cálculo de las medidas de tendencia central y variabilidad, como lo son la media y la varianza.

5. Obtención y contraste de resultados como producto de ejecutar los pasos anteriores.

[9] La métrica de desempeño usado fue el error cuadrático medio

### Buenas prácticas para el desarrollo de Software

Desde la perspectiva de la ingeniería aplicada en el desarrollo del software de integración y estimación, se resalta que, dentro de la investigación realizada, se establecieron y cumplieron los principios que los mismos autores trazaron para el desarrollo y la constitución del proceso. Esto último, en razón a la revisión del proceso desde su ejecución, implementación y documentación, la cual fue gestionada desde un proyecto GitHub[10]. Los principios de transparencia, auditabilidad, reproducibilidad y escalabilidad tienen correspondencia en la filosofía de algunas de las buenas prácticas trazadas en la Ingeniería de Software, así como en el software de Código Abierto y en la Ciencia Abierta[11]. La transmisión de conocimiento al equipo técnico del DANE ha sido transparente, a la luz de los principios guía de Mark D. Wilkinson[12] para la gestión y administración del software para investigación. A continuación, se señala la revisión de cada uno de estos principios:

- *Localizable*: El proyecto GitHub implementado.
- *Accesible*: Con excepción de las bases de datos del proyecto, se brindó acceso de lectura y transmisión de conocimiento del proyecto GitHub a las dependencias del DANE, en el que se desarrolla la implementación de la metodología de Ciencia de Datos que se aplicó en la investigación.
- *Interoperable*: El software desarrollado es ejecutable sin importar el sistema operativo y sus diferentes scripts interoperan con archivos en formato de almacenamiento de datos .parquet de Apache, .npy de Python, .xlsx de Excel y archivos de texto en formato .csv y txt.
- *Reusable*: Los scripts de la implementación de las metodologías están disponibles al DANE y son reusables desde los softwares de ciencia de datos R, Python y Julia.

Se verificó, además, la propiedad del proyecto GitHub para salvaguardar la seguridad y confidencialidad de los datos que gestiona el proceso, pues estos son administrados por, y de acceso exclusivo de, los miembros del correspondiente equipo de investigación.

### Análisis demográfico de los resultados

El proyecto tiene como propósito: "dotar a las entidades del Sistema Integral de Verdad, Justicia, Reparación y No Repetición (SIVJRNR) de argumentos científicos sólidos para develar la magnitud de la

[10] Ver https://github.com
[11] Ver European Commission, Directorate-General for Research and Innovation, Open innovation, open science, open to the world : a vision for Europe, Publications Office, 2016, https://data.europa.eu/doi/10.2777/061652
[12] Ver The FAIR Guiding Principles for scientific data management and stewardship

violencia, mediante la estimación del subregistro y la identificación de patrones de victimización de homicidio, desaparición forzada, secuestro, desplazamiento forzado y reclutamiento ilícito"[13]. Por tanto, uno de los principales focos de atención en la revisión que desarrolló el DANE fue el proceso utilizado para el desarrollo de la integración de datos y la elaboración de estimaciones.

Como primera medida, se considera que la investigación aborda un tema de gran interés para el DANE, dado que disponer de los resultados del ejercicio de estimación de cifras relacionadas con fenómenos como las desapariciones, el desplazamiento forzado, los reclutamientos y los secuestros en el país, permite enriquecer los análisis que desarrolla la Dirección de Censos y Demografía (DCD) en relación con la estimación de niveles, patrones y tendencias de la mortalidad y del desplazamiento interno en Colombia. En este sentido, lo que se ha podido identificar es que, tanto en los Censos de Población y Vivienda, como en encuestas por muestreo y en los registros administrativos disponibles, existe una omisión diferencial por edad y sexo, que se refleja en menores niveles de cobertura en los hombres de edades intermedias; es decir estas operaciones estadísticas tienen menores niveles de cobertura en un grupo poblacional que está fuertemente afectado por la violencia. Este patrón identificado es consistente en términos generales con los hallazgos presentados por los informes del proyecto conjunto de la CEV, JEP y HRDAG. No obstante, la intensidad estimada de los fenómenos depende de las técnicas implementadas para tal fin, las cuales se consideran idóneas[14]; razón por la que todos los resultados son válidos, especialmente al articular o integrar diferentes fuentes de observación, lo cual robustece las estimaciones contando con parámetros estadísticos descriptivos de seguimiento imparciales, suficientes y verosímiles.

En este sentido, los datos generados guardan coherencia con investigaciones destacadas en esta materia, y generan un potencial de aprovechamiento para la comprensión de las características y determinantes de aspectos claves del conflicto armado en el territorio nacional, que reflejan las marcadas brechas del impacto de estas de manera diferencial entre hombres y mujeres, por regiones y departamentos, así como para minorías étnicas[15].

[13] Tomado de "Informe metodológico del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística"
[14] Antecedentes de aplicación de las mismas pueden ser consultadas en las cifras disponibles del World Population Prospect (WPP): Raftery, A. E., Alkema, L., & Gerland, P. (2014). Bayesian population projections for the United Nations. Statistical science: a review journal of the Institute of Mathematical Statistics, 29(1), 58.
[15] Loja, J., Giraldo, F. U., & Carabali, B. Título: Un abordaje para el Análisis de Mortalidad en Colombia según Grupo Étnico; 2005-2010.
Fabio, S. Ã. (2005). Los efectos del conflicto armado en el desarrollo social colombiano, 1990-2002 (No. 003167). Universidad de los Andes-CEDE.
Segura-Cardona, A., & Cardona-Arango, D. (2018). Mortalidad y años potenciales de vida perdidos por causas externas: Colombia 1998-2015. Universidad y Salud, 20(2), 149-159.

## Análisis de las técnicas estadísticas y de aprendizaje maquinal utilizadas

Es clara la exposición de los aspectos conocidos sobre metodologías de estimación. Los antecedentes son relevantes y están a la vanguardia en relación con los retos que la Inteligencia Artificial enfrenta en la actualidad en cuanto a los problemas de procesamiento e interpretación del lenguaje natural para la clasificación y emparejamiento de registros. No obstante, como bien se señala en el informe, y como es usual en este tipo de investigaciones, se tomaron una serie de decisiones metodológicas para el desarrollo de este proyecto que pueden generar un impacto no mesurable en los resultados de esta.

En cuanto a las técnicas estadísticas y de aprendizaje maquinal utilizadas, más puntualmente, para la integración de individuos de diferentes fuentes (*Record Linkage*) y para la imputación de datos faltantes y estimaciones de número de víctimas; se presentan las siguientes observaciones, las cuales hacen énfasis en la intervención del investigador en una probable incidencia en los resultados finales:

1. Se revisaron las reglas de decisión para la generación de posibles emparejamientos a partir de los valores de los diferentes campos. Se verificó que todas estas reglas tengan un sentido lógico de la interpretación humana. Cabe mencionar que estos listados de reglas son en parte provenientes de la experticia en emparejamientos de datos que los investigadores han tenido en ejercicios de estimación de víctimas de Comisiones de la Verdad de otros países.
2. Se especificó sobre el insumo de construcción de reglas de emparejamiento para entrenamiento del modelo, el cual fue elaborado de forma manual por analistas expertos en etiquetado de estos datos, llamados "Orácula" y "Oráculos". Este es un punto de decisión en el que interviene la interpretación humana y que se encuentra fuera del alcance de comprender o contar la generación de todas las reglas. Se reconoce, además, que el etiquetado de estos datos solo puede ser realizado mediante interpretación de un humano que reconozca el lenguaje en que están los datos. Sin embargo, la fiabilidad de estas reglas fue validada estadísticamente con desempeños notablemente altos y por encima del 99% en las métricas reportadas: *F-score*, *recall* y *accuracy*[16].
3. La técnica de Pares Coincidentes Ocultos (citada como *aka pairing*) resulta pertinente en la reducción del universo de búsqueda de pares coincidentes entre bases de datos, dada la imposibilidad computacional de realizar comparaciones a todas las parejas posibles.
4. Se verificaron las variables, o más exactamente, las características o *features* generadas a partir de las diferentes bases de datos, con las cuales se comparan uno a uno los pares obtenidos.

Chaparro-Narváez, P., Cotes-Cantillo, K., León-Quevedo, W., & Castañeda-Orjuela, C. (2016). Mortalidad por homicidios en Colombia, 1998-2012. Biomédica, 36(4), 572-582.

[16] Ver Christen, Peter. 2012. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection - PP 165 a 171.

Todas las características se encuentran basadas en la interpretación humana, tienen un sentido lógico y son reconocibles en la literatura sobre Procesamiento del Lenguaje Natural[17].

5. Se especificó sobre el insumo de la clasificación de pares positivos para entrenamiento, el cual también fue elaborado de forma manual por analistas expertos en etiquetado de estos datos, a llamados "Orácula" y "Oráculos". Se resalta que, este es un punto de decisión en que interviene la interpretación humana y que se encuentra fuera del alcance de comprender una parte importante de todos los pares probables, pero se reconoce, además, que el etiquetado de estos datos solo puede ser realizado mediante interpretación de un humano que reconozca el lenguaje en que están los datos. Con este insumo se sustentó que la validación de la técnica de *clustering* implementada obtuvo un *Adjusted Rand Index*[18] o ARI de 0.9, lo que los mismos autores calificaron como optimo.
6. Se verificó que la imputación basada en la metodología *Hot-deck*[19] para los datos obtenidos de homicidios con bajas proporciones de datos faltantes, es pertinente y adecuada debido a esta última condición señalada.
7. Se verificó que la imputación basada en la arquitectura de red neuronal (LSTM) para la predicción y posterior imputación de datos de homicidios con altas proporciones de datos faltantes, y basados en las descripciones de los hechos del homicidio, es pertinente y adecuada en razón a que se hace uso de datos no estructurados. Se verificó puntualmente que, en el caso de la imputación de datos sobre el actor perpetrador del homicidio, el conjunto de datos de entrenamiento no tuviese intervención humana que sesgue la culpabilidad puntual hacia un perpetrador. Esto dado que se evidenció que el conjunto de datos de entrenamiento tuvo que ser construido mediante un muestreo aleatorio que permitiese balancear los conjuntos de datos soporte, debido a que existen perpetradores que tienen una menor participación relativa en el conflicto armado y que pueden estar siendo subestimados en los reportes finales. Esta es una condición natural e inevitable de las técnicas de aprendizaje maquinal e incluso de los métodos estadísticos.
8. Los autores reportan sobre la proporción de víctimas por perpetrador y la etnia de la víctima, sobre el conjunto de datos sin imputación, pero también sobre el conjunto de datos con imputación. Se resalta que los valores se pueden considerar coincidentes y sin evidencia a una posible intervención o sesgo de los mismos autores para incidir en el señalamiento de algún actor al momento de reportar estadísticas que puedan servir como proxi de culpabilidad en el conflicto armado.
9. Se valida el método de estimación bayesiana de poblaciones por captura-recaptura como el método más adecuado para estimar la población de víctimas no registrada y en el marco del conflicto armado, en comparación con métodos como *Dual systems* y los modelos Log-lineales,

[17] Ver Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. MIT press.
[18] Ver W. M. Rand (1971). "Objective criteria for the evaluation of clustering methods". Journal of the American Statistical Association. American Statistical Association. 66 (336): 846–850
[19] Ver Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. International statistical review, 78(1), 40-64.

y pese a los supuestos sobre población cerrada, homogeneidad de captura e independencia de registros o fuentes, claramente advertidos por los mismos autores.

10. Los autores estratificaron el conjunto de datos de víctimas capturadas y recapturadas, principalmente por año y el lugar del hecho; con el fin de dividir el procesamiento de la información y homogeneizar los datos para que el método reciba datos con menor variación y por ende le sean más fáciles de modelar. Los autores señalan que el procesamiento a partir de estos métodos es demandante en recursos computacionales y tiempo, por lo que redujeron el número de replicas[20] a analizar a un total de 10. Esto último incide de forma no mesurable en la variación de la estimación del número de víctimas. También, puede ser una posible fuente de introducción de sesgo por parte de los autores. A pesar de esto, se recalca que estas decisiones se ven obligadas por la selección y construcción de una metodología eficaz en relación con los recursos presupuestales y computacionales disponibles.
11. Los autores destacan textualmente la propagación de la varianza y de los errores de predicción y estimación, generados por el encadenamiento de métodos: "El efecto de las imputaciones múltiples es que se propaga el error de la imputación en el error reportado por las estimaciones nacionales. Si la imputación es incierta debido a que la variable imputada es débilmente correlacionada con las otras variables, el error de esta será mayor y, por lo tanto, el error reportado en las estimaciones nacionales también lo será. En caso inverso, si las variables imputadas están fuertemente correlacionadas con las otras variables del conjunto de datos, el error de la imputación será menor, lo cual implica que el error de las estimaciones nacionales también sea menor."

## 5. Análisis de resultados

Un primer aspecto que se considera bien manejado dentro de los resultados aportados es el uso de visualizaciones del tipo *Upset plots* para mostrar el nivel de coincidencia entre las múltiples fuentes de información que se trabajaron. Esto resulta útil para identificar fuentes que aportan unidades que no están presentes en otras fuentes, y para dimensionar la magnitud de los conjuntos de individuos coincidentes entre dos o más conjuntos de datos.

Otro punto para resaltar es la reconstrucción de series históricas a partir de estimaciones derivadas de múltiples fuentes de información, con gran potencial para el desarrollo de investigación aplicada y para enriquecer la discusión sobre los patrones y tendencias históricas que se han presentado en los fenómenos estudiados. El disponer de datos que dan cuenta de las diferencias geográficas y de los cambios temporales brindan posibilidades de reutilización de los resultados para la construcción de

[20] Existen 10 replicas para cada uno de los estratos escogidos por los investigadores.

modelos espacio temporales, así como para la identificación de patrones y tendencias de los fenómenos estudiados por parte de otros investigadores. También se considera que los contrastes planteados en gráficos y tablas son relevantes para el cumplimiento de los objetivos de la investigación y aportan información importante que de otra manera no podría visualizarse de manera integral.

En conjunto, resulta evidente el cuidado en el manejo de los microdatos y la construcción de indicadores. Las estimaciones de homicidios y su evolución, realizadas por la CEV reflejan la necesidad de complementar los registros oficiales de la mortalidad de manera consistente con la proyección demográfica, que permite construir las expectativas de vida de la población en Colombia (las cuales contemplan una medición del subregistro de los hechos vitales de forma indirecta de acuerdo con los resultados históricos de los censos de población). No obstante, dicho subregistro está supeditado a diferentes factores (relacionados con el contexto de reporte del hecho vital), de los cuales solo las aproximaciones estadísticas que incorporan la perspectiva demográfica logran evidenciar el volumen e intensidad de este fenómeno.

La evolución de los homicidios brutos reportados entre 1979 y 2019 en el registro de Estadísticas Vitales (EEVV) del DANE (ver gráfico 1)[21], sirve para realizar un contraste con el ejercicio realizado en el marco de la Comisión de la Verdad. Es de anotar que este número calculado de estadísticas vitales incluye tanto los homicidios asociados a conflicto armado como los que no están relacionados, sino que están asociados a otros factores determinantes como por ejemplo la violencia intrafamiliar o interpersonal. No obstante, también se hace necesario señalar que el registro de defunciones[22] tiene problemas de falta de cobertura diferencial a nivel subnacional, y no se cuenta con una estimación del nivel de omisión para esta causa específica, sino que se tiene una estimación de la omisión de la mortalidad por todas las causas.

Por consiguiente, la interpretación de las cifras brutas presenta distorsiones para la política pública y su seguimiento dado factores como los cambios en el nivel de cobertura y calidad a través del tiempo. Diferenciales asociados a la territorialidad y acceso a la recolección de información desagregada para áreas urbanas y rurales; así como, en características de los individuos según sexo, edad, nivel educativo, área de residencia, entre otras. Especialmente, no se cuenta con elementos técnicos para minimizar el subregistro de información como consecuencia del conflicto armado, teniendo en cuenta que existe un

[21] Las bases de defunciones del DANE contienen información sobre documento de identidad desde 1998, y a partir de 2008 se cuenta con las variables de nombres y apellidos, esto permite realizar cruces de información de individuo a individuo con cualquier otra base de datos disponible, a través del emparejamiento directo de registros.

[22] La información sobre defunciones que presenta el DANE en Estadísticas Vitales toma como insumo la información consignada en los certificados antecedentes del registro de defunción, que están a cargo del Sector Salud y Medicina Legal. El DANE en esta operación estadística, publica información bruta, es decir sin ajuste o imputaciones asociadas a la cobertura territorial de los hechos vitales.

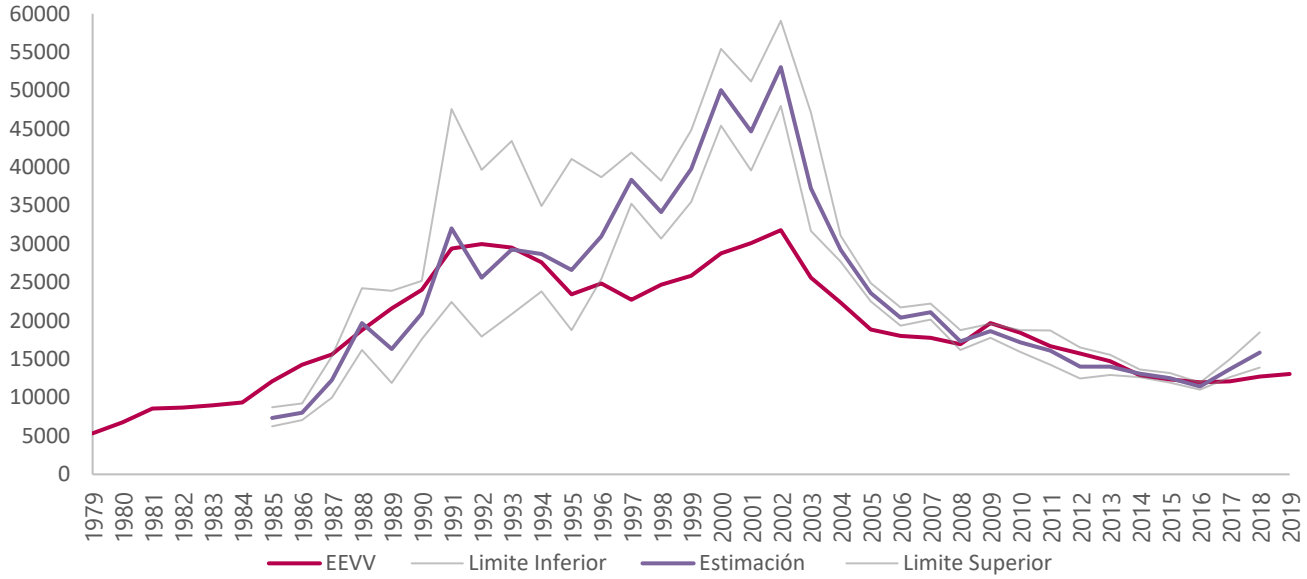
número significativo de desaparecidos y fosas comunes, que con la información disponible en el DANE no es posible identificar una estimación robusta de las cifras faltantes de la mortalidad por hechos violentos, para contar con un referente de comparación directa frente a los ejercicios validados por HRDAG y que permitan fortalecer el análisis de las tendencias de éste fenómeno, dichos ejercicios de estimación suelen aplicarse a partir de la aplicación de los métodos de sistema dual o múltiple[23].

La comparación entre los homicidios registrados en las estadísticas vitales del DANE y las estimaciones de la Comisión de la Verdad muestra resultados que pueden ser clasificados en cuatro periodos:

1. **1979-1987:** La tendencia de homicidios brutos reportados por EEVV y CEV es similar, siendo mayor la intensidad en los reportes de oficiales que la estimación realizada.
2. **1988-1996:** Es el periodo de mayor incertidumbre de la estimación, sin embargo, se destaca como el reporte oficial se encuentre dentro de los límites de confianza de la estimación.
3. **1997-2008:** Aunque las tendencias de crecimiento (1997-2002) y caída (2003-2008) son similares, este periodo resulta ser el de mayor discrepancia donde las estimaciones presentan niveles mucho mayores que lo reportado en EEVV.
4. **2009-2019:** Este periodo es de mayor similaridad tanto en tendencia como en volumen.

**Gráfico 1. Evolución histórica del número de homicidios en el registro de defunciones (DANE) 1979-2019 en contraste con las Estimaciones Comisión de la Verdad - HRDAG**

[23] CEPAL, N. (1974). Informe de la Reunión del Comité de Expertos para el Mejoramiento de las Fuentes de Estadísticas Demográficas= Report of the Meeting of the Expert Committee for the Improvement of Sources of Demographic Statistics.




Fuente: DANE - Estadísticas Vitales y Comisión de la Verdad - HRDAG

Por otra parte, para estimaciones de población inferior a 20 mil habitantes, la interpretación y el análisis de resultados es restringida (especialmente en el caso de los valores de reclutamiento ilícito de menores), dado que cualquier transformación estadística en poblaciones con un tamaño o volumen inferior o reducido presentan intervalos de incertidumbre muy amplios para ser concluyentes.

## 6. Concepto

El equipo directivo del Departamento Administrativo Nacional de Estadística (DANE), con el apoyo del personal técnico especializado, en el marco de sus funciones como ente rector de la producción estadística en el país, observa la suficiencia de las técnicas de estadística aplicada en el marco del proyecto conjunto JEP-CEV-HRDAG de integración de datos y estimación estadística, el cual fue desarrollado por el equipo liderado por el Dr. Patrick Ball del Human Rights Data Analysis Group (HRDAG), con la colaboración de Paula Andrea Amado y Alejandro Castro, y cuyos resultados cumplen con las características necesarias en materia de calidad. Entre estas, se ha valorado metodológicamente la complejidad de la implementación de los procesos de diseño, estandarización, imputación, organización de insumos, capacidad de personal, herramientas estadísticas, procesamientos y consolidación de resultados, realizada bajo altos estándares (División de Estadística de Naciones Unidas

- UNSD[24]) y buenas prácticas[25] en la ciencia estadística. Frente al panorama y visibilidad estadística de los fenómenos de desaparición, desplazamiento forzado, homicidios, reclutamiento a menores y secuestro asociados al conflicto armado colombiano, en el periodo 1985 – 2020, se asegura la calidad de las estimaciones derivadas por las cuales se destacan los siguientes criterios de validación técnica y temática:

- **Principios de buenas prácticas en estadística y ciencia de datos aplicada:** se observa coherencia entre los procedimientos aplicados de manera transparente, auditable, reproducible y escalable entre otras consideraciones técnicas de calidad frente a los propósitos de las estimaciones aplicadas.
- En los procesos de compilación e integración de datos de registros administrativos se observa la **aplicación de actividades y reportes orientados a validar la calidad de los datos**, esto garantiza que la base integrada sobre la que se realizan los ejercicios y análisis posteriores no genere sesgos adicionales para la identificación de la completitud de las variables o los registros duplicados.
- **Principios de buenas prácticas en desarrollo de software:** se establecieron y cumplieron los principios acordes a las buenas prácticas trazadas por la Ingeniería de Software, así como por los lineamientos del software de Código Abierto y de la Ciencia Abierta.
- **Metodología:** es loable el esfuerzo en el desarrollo y seguimiento del flujo de procesos implementados para la elaboración de las estimaciones en cada uno de los ámbitos temáticos sobre el conflicto armado del país. Si bien el DANE ha adelantado ejercicios similares con la integración de diferentes fuentes de información, la diversidad de estas para su estandarización, conceptualización e imputación de datos demuestra un alto desempeño y calificación del equipo de trabajo vinculado al proyecto. Así mismo, esto expresa un alto grado de robustez técnica que es respaldado consistentemente por las metodologías utilizadas ampliamente en el contexto internacional[26], así como en los problemas que actualmente aborda la Inteligencia Artificial para el procesamiento e interpretación automática del lenguaje natural.
- **Resultados:** se identifica un gran potencial en las estimaciones aplicadas, siendo congruentes con las estadísticas espejo en el contexto nacional. Especialmente, la información de homicidios muestra tendencias similares a las identificadas con otras fuentes como las estadísticas vitales (DANE), que corresponden a las cifras oficiales para el Estado colombiano. Es así, que los datos acumulados para el periodo histórico 1985-2020 tanto los observados, como los imputados y estimados, son referentes clave que generaran a mediano y largo plazo un conjunto de información relevante para el estudio histórico de las características propias de la realidad

[24] https://unstats.un.org/unsd/methodology/dataquality/references/UNNQAFManual-WEB-ESP-UNIFICADO-(final-manuscript)-April-2021.pdf
[25] Código Nacional disponible en: https://unstats.un.org/unsd/dnss/docs-nqaf/Codigo%20Nal%20Buenas%20Practicas.pdf
[26] Ver Civilian Deaths in the Syrian Arab Republic - Report of the United Nations High Commissioner for Human Rights

colombiana y sus implicaciones en diferentes ámbitos sociales. Sin embargo, se reitera la recomendación dada para el uso de estimaciones de poblaciones inferiores a 20 mil habitantes, ya que la interpretación y análisis de estos resultados presentan intervalos de incertidumbre muy amplios para ser concluyentes (en especial en lo referente a reclutamiento ilícito de menores) dado el tamaño o volumen significativamente bajo en términos de población.

## 7. Bibliografía

Christen, Peter. 2012. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. New York: Springer.

European Commission, Directorate-General for Research and Innovation, Open innovation, open science, open to the world: a vision for Europe, Publications Office, 2016, https://data.europa.eu/doi/10.2777/061652

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Raftery, A. E., Alkema, L., & Gerland, P. (2014). Bayesian population projections for the United Nations. Statistical science: a review journal of the Institute of Mathematical Statistics, 29(1), 58.

Loja, J., Giraldo, F. U., & Carabali, B. Título: Un abordaje para el Análisis de Mortalidad en Colombia según Grupo Étnico; 2005-2010.

Fabio, S. Ã. (2005). Los efectos del conflicto armado en el desarrollo social colombiano, 1990-2002 (No. 003167). Universidad de los Andes-CEDE.

Segura-Cardona, A., & Cardona-Arango, D. (2018). Mortalidad y años potenciales de vida perdidos por causas externas: Colombia 1998-2015. Universidad y Salud, 20(2), 149-159.

Chaparro-Narváez, P., Cotes-Cantillo, K., León-Quevedo, W., & Castañeda-Orjuela, C. (2016). Mortalidad por homicidios en Colombia, 1998-2012. Biomédica, 36(4), 572-582.

Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. MIT press.

W. M. Rand (1971). "Objective criteria for the evaluation of clustering methods". Journal of the American Statistical Association. American Statistical Association. 66 (336): 846–850

Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. International statistical review, 78(1), 40-64.