PÚBLICA

INSTITUTO COLOMBIANO DE BIENESTAR FAMILIAR

GUÍA METODOLÓGICA PARA LA ANONIMIZACIÓN DE REGISTROS

TABLA DE CONTENIDO

1. Introducción 3
2. Objetivo 3
3. Alcance 3
4. Definiciones 4
5. Antecedentes 5
6. Proceso de Anonimización 7
6.1. Etapas del proceso de anonimización. 7
6.1.1. Etapa I. Revisiones previas. 8
6.1.2. Etapa II. Análisis de riesgos de identificación de las fuentes de información. 9
6.1.3. Etapa III. Identificación y selección de técnicas de anonimización. 10
6.1.3.1. Métodos de reducción o generalización. 10
6.1.3.2 Métodos de aleatorización o perturbación. 11
6.1.4 Etapa IV. Análisis de viabilidad del proceso. 13
6.1.5 Etapa V. Aplicación de técnicas de anonimización. 13
6.1.6 Etapa VI. Evaluación de resultados del proceso. 14
7 DOCUMENTOS DE REFERENCIA 15
8 CONTROL DE CAMBIOS 16

PÚBLICA

1. Introducción

En concordancia con el propósito del Plan de Analítica Institucional en el fortalecimiento de la cultura de datos, organizacional y tecnológica del Instituto que permita mejorar la calidad de los procesos, tratamiento y difusión de las estadísticas oficiales generadas, es evidente la necesidad de adoptar las recomendaciones que, en materia de tratamiento de datos, plantea el *Código Nacional de Buenas Prácticas del Sistema Estadístico Nacional* (SEN) del Departamento Administrativo Nacional de Estadística – DANE.

Siguiendo los principios del Código para el procesamiento estadístico, es necesario establecer protocolos de seguridad y confidencialidad que protejan la privacidad de las fuentes en el proceso estadístico o en el intercambio de microdatos. Para esto se recomienda el uso de técnicas para la anonimización de microdatos que garanticen la protección de la identificación o localización geográfica de las fuentes empleadas en el proceso estadístico (DANE, 2017: 12). El objetivo de este proceso es facilitar la accesibilidad de la información, permitiendo el acceso de las estadísticas y los microdatos asociados a todo tipo de usuarios con el máximo detalle posible y en diferentes formatos y medios que faciliten la consulta, visualización y uso (DANE, 2017: 11).

Bajo estos principios, el DANE (2018) desarrolló la Guía para la Anonimización de Bases de Datos en el Sistema Estadístico Nacional, cuyo propósito se centra en orientar a los integrantes del Sistema sobre el proceso de anonimización de bases de datos que provienen de registros administrativos y de operaciones estadísticas. Este documento identifica las buenas prácticas, herramientas e instrumentos cuando se implementen procesos de anonimización para la producción de estadísticas, así como para otros usos de la información anonimizada.

Tomando como base el documento anteriormente mencionado se construye, a partir de las recomendaciones del DANE, esta guía con el fin de establecer los lineamientos para la anonimización de las distintas bases de datos de la entidad.

2. Objetivo

Establecer los lineamientos metodológicos para realizar procesos de anonimización de las bases de datos, microdatos que resultan de los registros administrativos del Instituto Colombiano de Bienestar Familiar.

3. Alcance

La presente guía aplica a los procesos de gestión y procesamiento de bases de datos del nivel nacional.

PÚBLICA

4. Definiciones

Las siguientes definiciones son tomadas de la Guía para la Anonimización de Bases de Datos en el Sistema Estadístico Nacional.

Anonimización: Se define la anonimización de microdatos como un proceso técnico que consiste en transformar los datos individuales de las unidades de observación, de tal modo que no sea posible identificar sujetos o características individuales de la fuente de información, preservando así las propiedades estadísticas en los resultados" (Decreto 1743 de 2016: Art. 2.2.3.1.1). La finalidad de la anonimización es impedir que, a partir de una información o de una combinación de informaciones, se logren identificar sujetos individuales ya sean individuos, empresas o establecimientos, u otro tipo de unidades de observación en un archivo de microdatos (DANE, 2018: 9).

Microdato: Cada uno de los datos sobre las características de las unidades de estudio de una población (individuos, hogares, establecimientos, entre otras) que se encuentran consolidados en una base de datos.

Base de Datos: Conjunto o colección de datos interrelacionados entre sí, que se utilizan para la obtención de información de acuerdo con su contexto y que son almacenados sistemáticamente para su posterior uso.

Datos personales: Es toda información numérica, alfabética, grafica, fotográfica, acústica o de cualquier otro tipo, susceptible de recogida, registro, tratamiento y transmisión, concerniente a personas físicas identificadas o identificables (tales como nombre, apellidos, estado civil, sexo, edad, domicilio, número de la seguridad social, número de matrícula del empleado, identificación personal, número de teléfono, etc.)

información sensible: Es la información considerada como estrictamente confidencial. información y características referentes a la edad, procedencia, salud, raza, religión, ideología, afiliación, finanzas, etc., se consideran de carácter sensible y requieren de una protección especial.

Tratamiento de los datos personales: Hace referencia a cualquier operación o procedimiento técnico, sea o no automatizado, que permita la recogida, grabación, conservación, elaboración, modificación, consulta, utilización, cancelación, bloqueo o supresión, así como las cesiones de datos que resulten de comunicaciones, consultas, interconexiones y transferencias.

5. Antecedentes

El principal propósito de la implementación de técnicas de anonimización es "*incrementar la desagregación de la información, manteniendo los niveles de confidencialidad, así como generar un mayor aprovechamiento estadístico de la misma."* (DANE, 2018: 5). En este sentido, es necesario garantizar los mecanismos para dar cumplimiento a la legislación en materia de protección, privacidad y confidencialidad de la información.

A nivel internacional, se pueden resaltar los esfuerzos de entidades como la División de Estadísticas de las Naciones Unidas, la cual establece como principio que "*los datos que reúnan los organismos de estadística para la compilación estadística, ya sea que se refieran a personas naturales o jurídicas, deben ser estrictamente confidenciales y utilizarse exclusivamente para fines estadísticos*" (MINSALUD, 2016: 8). Así mismo, la Comisión Económica para Europa de las Naciones Unidas publicó el manual sobre el control de divulgación estadística, el cual provee lineamientos técnicos para el control de la divulgación de la información, describiendo los métodos aplicables para la protección de la privacidad de la información y explicando detalladamente el programa de anonimización ARGUS[1], una iniciativa que viene liderando la implementación de mecanismos de anonimización en los productores de estadística.

De igual forma, en el contexto jurídico internacional, la Unión Europea desde su Directiva 95/46/CE relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos, en su considerando 26, establece que *"los principios de la protección deberán aplicarse a cualquier información relativa a una persona identificada o identificable… los principios de la protección no se aplicarán a aquellos datos hechos anónimos de manera tal que ya no sea posible identificar al interesado… los datos pueden hacerse anónimos y conservarse de forma tal que impida identificar al interesado*". Lo que excluye los datos anonimizados del alcance de la legislación sobre protección de datos personales siempre y cuando cumpla con los estándares establecidos.

En cuanto a la legislación colombiana, la Constitución Política en su artículo 15 establece que todas las personas tienen derecho a su intimidad personal, e indica que en lo referente a la recolección, tratamiento y circulación de datos, se respetarán la libertad y demás garantías consagradas en la Constitución. Por otro lado, en su artículo 20 se establece el derecho a recibir información veraz e imparcial. En este sentido, en Colombia se decretó la Ley Estatutaria 1266 de 2008 por la cual se dictan las disposiciones generales del hábeas data y se regula el manejo de la información contenida en bases de datos personales, en especial la financiera, crediticia, comercial, de servicios y la proveniente de terceros países.

[1] ARGUS es un programa interactivo y libre, cuyas funcionalidades permiten identificar los datos de una base (metadatos), seleccionar y calcular las tablas de frecuencia, establecer la base de los métodos de anonimización y aplicar las diferentes técnicas a las variables relevantes. El programa es compatible con otros programas estadísticos.

A su vez, la Ley 1581 de 2012, reglamentada por el decreto 1377 de 2013, dicta las disposiciones generales para la protección de datos personales. De esta forma, dispone en sus principios referentes al acceso y circulación restringida de seguridad y de confidencialidad que el acceso a los datos se debe restringir y la información debe estar sujeta a tratamiento por parte del responsable*"* (MINSALUD, 2016: 6).

Posteriormente, la Ley 1712 de 2014 crea la Ley de Transparencia y establece el derecho de acceso a la información pública nacional, los procedimientos para el ejercicio y garantía del derecho y las excepciones a la publicidad de información, "*establece la divulgación de la información para promover y generar una cultura de la transparencia"* (MINSALUD, 2016: 6), "*haciendo énfasis en el establecimiento de una política de datos abiertos por parte de las entidades públicas."* (DANE, 2018: 8).

Por otro lado, el Instituto Colombiano de Bienestar Familiar (ICBF), como responsable del tratamiento de datos personales obtenidos en desarrollo de su objeto y funciones legales, y en ejercicio de los deberes contenidos en la Ley 1581 de 2012 y el capítulo 25 del Título 2 de la Parte 2 del Libro 2 del Decreto 1074 de 2015, estableció su política de tratamiento de datos personales, buscando garantizar la protección de los derechos fundamentales en su tratamiento, la cual empezó a regir a partir del 25 de noviembre del año 2015, fue actualizada el 7 de diciembre del año 2017 y posteriormente ajustada en el año 2019.

En este sentido, en el desarrollo de la política de tratamiento de datos personales, se aplican, de manera armónica e integral, los principios contenidos en la Ley 1581 de 2012: a) legalidad en materia de tratamiento de datos, b) finalidad, c) libertad, d) veracidad o calidad, e) transparencia, f) acceso y circulación restringida, g) seguridad y h) confidencialidad. Estos principios y disposiciones contenidas en la política de tratamiento son aplicables a los datos personales registrados en cualquier base de datos que los haga susceptibles de tratamiento, debiendo ser aplicada por los servidores públicos y contratistas del ICBF, así como de todos aquellos que actúen como encargados de datos personales cuyo responsable sea el ICBF.

Finalmente, desde el punto de vista del índice de madurez de Big Data del año 2019, el DNP adelantó una consultoría con Economía Urbana y Galileo para construir un índice que permitiera definir la capacidad de cada entidad para la explotación de datos (Índice de Maduración de Big Data (Imbigda)). El índice de madurez mide la profundidad de la implementación de aspectos relacionados con el ámbito Organizacional, cultural, tecnológico, jurídico y ético. En este sentido, estas dimensiones comprenden el entendimiento y aplicación por parte de la entidad tanto de los preceptos normativos relacionados con la protección de datos personales, la transparencia y el gobierno abierto, tanto de las consideraciones éticas que pueden traer consigo la aplicación de algoritmos o de otras técnicas de análisis.

6. Proceso de Anonimización

Este proceso debe procurar controlar el riesgo de identificación de las personas que acceden a entregar su información con fines estadísticos según los medios disponibles, teniendo en cuenta que se debe preservar la utilidad y aprovechamiento de los datos que necesitan los directos responsables o terceros.

"*La desagregación de datos de manera muy detallada puede revelar información sensible o confidencial que conlleve a la identificación de una persona en particular… Las herramientas o metodologías utilizadas para llevar a cabo la anonimizarían de un conjunto de datos deben ser constantemente evaluadas y actualizas según la tecnología disponible y los elementos contextuales, teniendo en cuenta que este proceso lleva implícito un factor de riesgo*." (MINSALUD, 2016: 9).

El ICBF es responsable del tratamiento de datos personales obtenidos en desarrollo de su objeto y funciones legales, teniendo en cuenta que la misión del ICBF consiste en trabajar con calidad y transparencia por el desarrollo y la protección integral de la primera infancia, la niñez, la adolescencia y el bienestar de las familias colombianas.

Por lo anterior, y ante la necesidad de asegurar una adecuada y eficiente gestión institucional, en ejercicio de los deberes contenidos en la información sensible puede ser utilizada con fines inapropiados, resulta primordial entregar directrices a los servidores públicos y contratistas del ICBF, así como de todos aquellos que actúen como encargados de datos personales cuyo responsable sea el ICBF, con el fin de salvaguardar la integridad de la información hasta que sea trasmitida al ICBF para su custodia.

6.1. Etapas del proceso de anonimización.

Para lograr un proceso efectivo de anonimización de los datos es necesario eliminar de ellos los elementos suficientes para que no se pueda identificar a una persona mediante el conjunto de medios que puedan ser razonablemente utilizados por el responsable del tratamiento o por terceros. Lo óptimo es decidirlo caso por caso y puede requerir la aplicación de una o varias técnicas, sopesando siempre el menor daño residual en la calidad de los datos. Es conveniente tener en cuenta que el proceso en todas sus etapas debe quedar documentado (MINSALUD, 2016: 9).

El proceso de anonimización de una base de datos se encuentra compuesto por seis etapas que son: i) Revisiones previas; ii) Análisis de riesgos de identificación de las fuentes de información; iii) Identificación y selección de técnicas de anonimización; iv) Análisis de viabilidad del proceso, v) Aplicación de técnicas de anonimización, y vi) Evaluación de resultados del proceso.

6.1.1. Etapa I. Revisiones previas.

La disponibilidad de los requerimientos previos empezará con la primera etapa del proceso de anonimización. Para esto es necesario contar con los siguientes insumos: 1) Un equipo de trabajo que conozca temáticamente el contenido de la base de datos a anonimizar y que maneje herramientas que permitan el análisis exploratorio de datos (R, SAS, SPSS, Stata, Python, entre otros); 2) Disponer de una base de datos; 3) El diccionario de datos de la base a anonimizar; 4) Infraestructura tecnológica (paquetes estadísticos, equipos de cómputo que permitan el manejo de datos, y en general tecnología que se encuentre acorde con el volumen de información a anonimizar); 5) Definir mecanismos de seguridad sobre la base de datos a anonimizar.

En esta etapa se busca que el equipo encargado realice una revisión de los insumos disponibles para la ejecución del proceso. La etapa se compone de tres subprocesos, así: 1) Análisis exploratorio de la base de datos. 2) Revisión normativa sobre protección de datos e identificación de usuarios de la información (restricciones de publicación de la información y necesidades de los usuarios de la información). 3) Definición de las propiedades estadísticas a conservar en la base de datos.

Al finalizar con esta etapa se contará con los siguientes productos:

Producto Etapa I:
- Base de datos a anonimizar caracterizada
- Propiedades globales de las variables
- Revisión temática de la base
- Revisión de restricciones de publicación de la información
- Identificación de usuarios de la información
- Propiedades estadísticas a conservar en la base de datos anonimizada

Fuente: DANE, 2018: 25.

PÚBLICA

6.1.2. Etapa II. Análisis de riesgos de identificación de las fuentes de información.

Un riesgo de identificación de las unidades de observación de la base de datos es aquel en el cual existe una posibilidad en la cual, mediante la combinación de variables en la base de datos, se logre identificar características de las unidades de observación que deben ser protegidas por el tipo de información que contienen. Esta etapa se compone de cuatro subprocesos:

1) Clasificación de variables por su nivel de sensibilidad y su nivel de contenido de información privada de la unidad de observación de la base de datos.
    a. Identificadores directos: Contienen información sensible como de identificación o ubicación (cédula, NIT o direcciones, entre otras)
    b. Pseudo-identificadores: Combinadas con otras variables conllevan a la identificación de las unidades de observación (comúnmente coinciden con variables temáticas o de ubicación en el análisis exploratorio de la base de datos, por ejemplo nivel de escolaridad, ocupación e ingreso promedio mensual en determinado municipio).
    c. No confidenciales: Estas variables no permiten la identificación de las unidades de observación de la base de datos, ni siendo combinadas con pseudo-identificadores (habilidades, gustos, enfermedades en una base de datos de mercado laboral)

2) Planteamiento de riesgos de la base de datos. Los riesgos se pueden entender como todas las posibles combinaciones de las variables (entre identificadores directos y pseudo-identificadores) y sus niveles de desagregación (geográfica o temática), que pueden aumentar la probabilidad de que una o varias unidades de observación sean identificadas por los usuarios de la información. Una vez definidos los riesgos de la base de datos, el equipo de trabajo los organizará en un listado y posteriormente los priorizará teniendo en cuenta la frecuencia en que pueden ser identificadas las unidades de observación.

3) Identificación de unidades de observación riesgosas. Las unidades de observación riesgosas son aquellas que cumplen con al menos una de las condiciones planteadas por el equipo de trabajo para ser susceptibles a identificación. Una unidad de observación puede ser riesgosa por sólo un riesgo, o por todos los riesgos planteados por el equipo de trabajo.

4) Creación del informe de riesgos. Será utilizado en la identificación y aplicación de técnicas de anonimización (Etapa III), y describirá cómo se clasifican las variables según tipo de sensibilidad. Los criterios utilizados para la definición de riesgos de identificación y las unidades de observación que son riesgosas a la hora de publicar la base de datos, contiene: Criterios y aspectos considerados en la definición de los riesgos, listado de riesgos definitivos priorizados, resumen de unidades de

PÚBLICA

observación riesgosas obtenida en el tercer subproceso de esta etapa, fecha de emisión del informe.

Al finalizar con esta etapa se contará con los siguientes productos:

Producto Etapa II:
- Clasificación de variables por su tipo de sensibilidad
- Planteamiento de Riesgos de Identificación
- Identificación de las unidades de observación riesgosas
- Informe de Riesgos de identificación

Fuente: DANE, 2018: 35.

6.1.3. Etapa III. Identificación y selección de técnicas de anonimización.

Esta etapa se compone de dos subprocesos, así:
1) Identificación de técnicas de anonimización más comunes para variables cuantitativas y categóricas;
2) Selección de una o más técnicas para aplicar a cada uno de los riesgos planteados en la etapa anterior. Para esto es necesario hacer una valoración teniendo en cuenta que se debe minimizar el riesgo de: 1. Identificación o singularización de una unidad de observación; 2. Vinculabilidad entre los datos de la misma base o con datos provenientes de otra(s) bases de datos; 3. Inferencia o deducir a partir de los valores o a partir del análisis de la información contenida en la base de datos.

6.1.3.1. Métodos de reducción o generalización.

Son técnicas basadas en la no perturbación de datos, mediante las supresiones parciales, reducción o recodificación de la información para minimizar el riesgo de identificación de las unidades de observación, ya que la reducción o generalización que producen supresiones o reducción del nivel de detalle del conjunto original de manera parcial o total evita que los datos atípicos sean de fácil identificación.

Técnica	Tipo de Variable	Descripción	Referencia Bibliográfica
Eliminación de variables	Categóricas	Esta técnica suprime toda la información de una variable. Se usa cuando la variable contiene información de identificación directa de la unidad de observación o pueden contener información sensible y no siempre se puede aplicar otros métodos de anonimización; también es muy posible que las bases de datos contengan información irrelevante para los responsables temáticos. Algunos ejemplos de variables que deben ser suprimidas son: nombres, documentos de identidad, números telefónicos, correos, fotografías, etc.	Hundepool et al. (2010)
Recodificación Global	Cuantitativas o categóricas	Combina diversas categorías de las variables categóricas en una más general que tenga mayor frecuencia y menor información, es	Hundepool et al., (2012)

Cualquier copia impresa de este documento se considera como COPIA NO CONTROLADA.

Técnica	Tipo de Variable	Descripción	Referencia Bibliográfica
		decir, colapsa las categorías con el fin de hacerlas menos específicas. Para variables continuas agrupa por medio de intervalos, manteniendo la utilidad de los datos. Esta técnica es recomendable cuando se desean proteger unidades de observación con riesgo de identificación a partir de las variables pseudo-identificadoras.	Templ et al., IHSN Working Paper No. 007 (2014).
Codificación superior e inferior	Cuantitativas o categóricas	Consiste en proteger la identificación de las unidades de observación que presentan los valores más altos o bajos de cada variable. Se utiliza cuando se presentan valores máximos y mínimos en el nivel de desagregación geográfico o temático que son de fácil identificación.	Hundepool et al. (2012)
Supresión local o Supresión de celdas	Categóricas	Reemplazar los valores de una o más variables de las unidades de observación identificadas como riesgosas por valores faltantes. Esta técnica se usa cuando la combinación entre las variables pseudo-identificadoras permita la identificación de las unidades de observación. Sí al hacer cruce de información a partir de tablas de contingencia, estas muestran celdas que pueden revelar información (frecuencias bajas) que conduzca a la identificación de alguna(s) persona(s), se puede suprimir una o varias celdas de la tabla. Es útil también cuando se trata de información presentada en gráficos.	Hundepool y De Wolf (2012) Templ et al., IHSN Working Paper No. 007 (2014) MINSALUD, 2016, p 15
Eliminación de registros	Categóricas	Cuando al aplicar alguna(s) técnica(s) de anonimización de datos, no surta el efecto deseado porque siguen siendo identificables algunas de las personas, se puede recurrir a la eliminación de registros; cabe señalar que este debe ser un procedimiento de último recurso.	MINSALUD, 2016, p 15

Fuente: Construcción a partir de DANE, 2018: 41-42, MINSALUD, 2016: 15.

6.1.3.2 Métodos de aleatorización o perturbación.

Estas técnicas. basadas en la perturbación de datos, son procedimientos que implican la modificación sistemática de datos (a veces en pequeñas cantidades aleatorias), que puede ser vista como una modificación de la veracidad de estos, de manera que las cifras no sean lo suficientemente precisas como para revelar información sobre casos individuales. Pueden incluirse nuevos datos, suprimir y/o modificar los existentes beneficiando la confidencialidad estadística.

Técnica	Tipo de Variable	Descripción	Referencia Bibliográfica
Microagregación	Cuantitativa	Consiste en reemplazar los valores de algunas unidades de observación, por el valor promedio calculado sobre ellas. El mínimo aceptado de unidades de observación es 3, sin embargo, en la elección de las unidades de observación se debe sopesar la perdida de información y de variabilidad. Comúnmente, se usa cuando la unidad de observación por nivel de desagregación geográfica es de fácil identificación	Hundepool et al., (2012) Templ et al., IHSN Working Paper No. 007 (2014)

Cualquier copia impresa de este documento se considera como COPIA NO CONTROLADA.

Técnica	Tipo de Variable	Descripción	Referencia Bibliográfica
Adición de ruido	Cuantitativa	Consiste en agregar una cantidad aleatoria definida por el equipo de trabajo sobre los valores de las unidades de observación (ruido aleatorio). Es útil cuando se desea proteger las unidades de observación y se ha identificado que por medio de cruces de información con bases de datos externas se expone la información confidencial, en este caso los atributos pueden causar un importante efecto adverso en las personas y por eso es necesario la modificación de atributos mediante la generación de valores aleatorios lo que proporciona un conjunto de datos menos exactos. La adición de ruido no es una medida suficiente para anonimizar un conjunto de datos; esta técnica se debe combinar con otra (s) con el fin de minimizar los riesgos. Por otro lado es importante tener en cuenta que el ruido introducido en un conjunto de datos debe respetar la lógica de los atributos; por ejemplo no se pueden superar los límites naturales de una variable.	Hundepool A. 2012, p.54 Templ et al., IHSN Working Paper No. 007 (2014), p.9
Permutación o intercambio de datos	Cuantitativas o categóricas	Intercambia la información de las unidades de observación identificadas con riesgo, con la información de las unidades de observación que no tienen riesgo de identificación. Este intercambio de datos se realiza de manera aleatoria entre pares de observaciones (con riesgo de identificación y sin riesgo). Esta permutación garantiza que el rango y la distribución de valores sean idénticos además de que mantiene el nivel de detalle; no obstante las correlaciones entre los valores y los individuos pueden quedar destruidas.	Hundepool et al., 2010, p. 58
Redondeo	Cuantitativas	Consiste en la sustitución del valor de las unidades de observación originales por valores redondeados (cero decimales). Comúnmente, se usa después de aplicar Microagregación y cuando la información de las variables técnicamente se debe expresar en unidades enteras y no decimales.	Hundepool et al. (2012)
Reajuste de los pesos o factores de expansión	Cuantitativas	En los casos donde se conoce el tipo de muestreo utilizado es posible revertir el proceso, lo cual aumenta la posibilidad de identificar de manera puntual a una persona; por tanto, es conveniente hacer una modificación de los pesos con el fin de disminuir este riesgo.	MINSALUD, 2016: 14

Fuente: Construcción a partir de DANE, 2018: 43-44, MINSALUD, 2016: 14.

Una vez identificada las técnicas de anonimización más comunes por tipo de variable, se seleccionará una o más técnicas que permitan minimizar la ocurrencia de cada uno de los riesgos identificados en la Etapa II.

Producto Etapa III:
- Características de las técnicas de anonimización
- Técnicas de anonimización para cada uno de los riesgos identificados

Fuente: DANE, 2018: 47.

6.1.4 Etapa IV. Análisis de viabilidad del proceso.

Esta etapa busca establecer el beneficio que puede proveer el equipo de trabajo al generar mayores desagregaciones de la información, frente a los riesgos de identificación de las unidades de observación que se encuentran en la base de datos. Se deberá analizar las necesidades de los usuarios, las limitaciones normativas, las políticas de la entidad y los aspectos temáticos de la base de datos, según los siguientes criterios:

1) Revisión temática y normativa, se analiza si existe alguna norma, ley o una directriz temática de la entidad productora de la información, que impida la publicación de las variables incluidas en la base de datos, o de las variables más útiles para los usuarios. A partir de esta revisión, el equipo podría considerar que el proceso de anonimización de la base de datos no es viable.
2) Técnicas de anonimización, al analizar la selección de las técnicas vistas anteriormente, se podría identificar que ninguna de éstas permite que la base de datos anonimizada conserve las propiedades estadísticas iniciales, por lo que podría considerar que no es viable realizar el proceso de anonimización.
3) Nivel de utilidad de la información, se analiza el nivel de utilidad de la información que podrá ser publicada a los usuarios. Si esta utilidad es baja y no permite la réplica de las cifras publicadas por la entidad, se considerará que el proceso de anonimización no es viable.

El producto de esta etapa es:

Producto Etapa IV:
- Informe y concepto de viabilidad del proceso de anonimización.

Fuente: DANE, 2018: 47.

6.1.5 Etapa V. Aplicación de técnicas de anonimización.

En esta etapa se implementará las técnicas de anonimización asociadas a los riesgos de identificación seleccionadas en la Etapa III, obteniendo así una primera versión de la base de datos anonimizada que será examinada cuidadosamente en la siguiente etapa. Para la aplicación de estas técnicas se debe tener en cuenta los siguientes pasos:

1) Clasificación y planteamiento de técnicas de anonimización para cada uno de los riesgos identificados.
2) Elección del software a implementar las técnicas de anonimización.
3) Rutinas (algoritmos) que ayuden a implementar las técnicas, para que el riesgo de identificación de las unidades de observación disminuya. Se recomienda que en las rutinas el equipo de trabajo siga la siguiente estructura:
    a. Cargue de base de datos a anonimizar.
    b. Tipos de riesgos identificados y la explicación de éstos.

Cualquier copia impresa de este documento se considera como COPIA NO CONTROLADA.

c. Consolidación de riesgos identificados.
d. Técnica de anonimización a aplicar
e. Verificación de riesgos.
f. Exportación de base de datos anonimizada.

Al finalizar esta etapa se contará con el siguiente producto:

Producto Etapa V:
- Base de datos anonimizada
- Rutina del proceso de anonimización

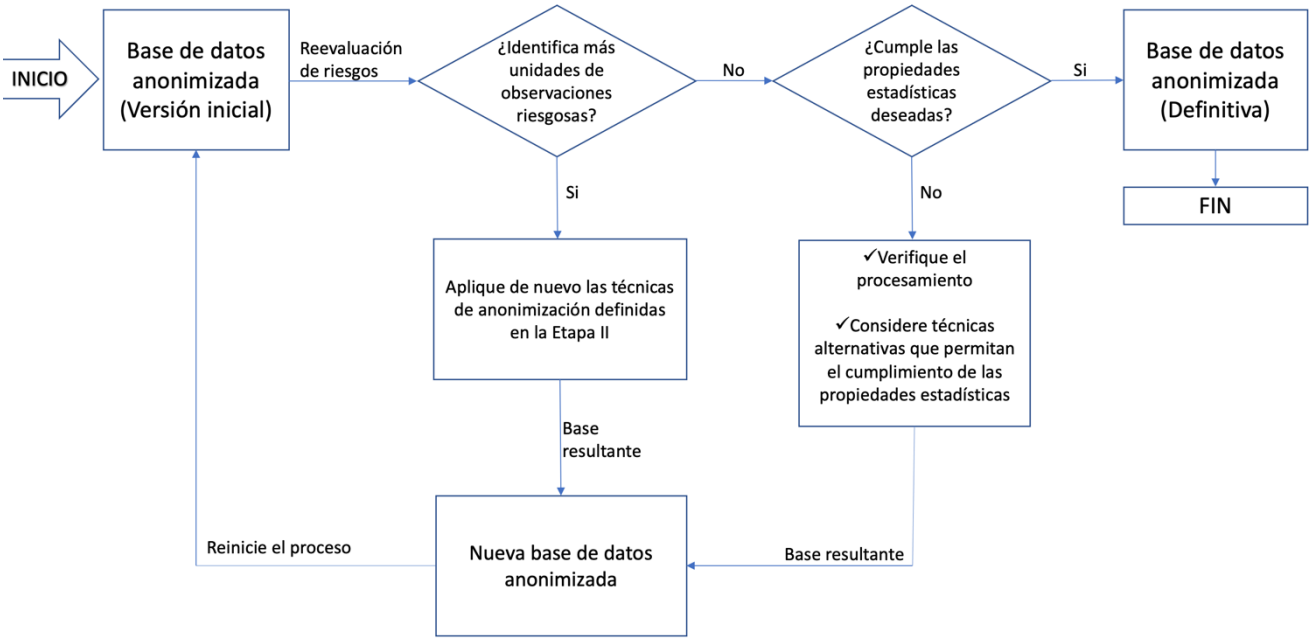
Fuente: DANE, 2018: 52.

6.1.6 Etapa VI. Evaluación de resultados del proceso.

En esta última etapa se procederá a evaluar los resultados del procedimiento de anonimización, se confirma que los riesgos de identificación de las unidades de observación se hayan subsanado y que las variables de la base de datos conserven las propiedades estadísticas iniciales. Esta etapa se divide en tres subprocesos así:

1) Revisión de propiedades estadísticas de la base de datos original contra la base de datos anonimizada.

2) Reevaluación de riesgos de identificación, para esto se sigue el siguiente flujograma.



Fuente: DANE, 2018: 61.

Cualquier copia impresa de este documento se considera como COPIA NO CONTROLADA.

3) Por último, se procede con la creación del Informe Final del Proceso de Anonimización – IFPA, el DANE recomienda la siguiente estructura de presentación:
   a. Características del Equipo de Trabajo e Insumos
   b. Revisiones Previas al Proceso de Anonimización
   c. Análisis de Riesgos de Identificación de las Unidades de Observación
   d. Selección de Técnicas a Implementar
   e. Análisis de Viabilidad
   f. Aplicación de las Técnicas de Anonimización
   g. Evaluación de Resultados

Esta última sección documenta los hallazgos encontrados en la evaluación de resultados y responde a las siguientes preguntas: "*¿Las propiedades estadísticas esperadas se cumplen en la base de datos anonimizada con respecto a la base de datos original? ¿Debido al incumplimiento de las propiedades estadísticas esperadas tuvo que verificar el procesamiento de las técnicas de anonimización? ¿encontró algún error? ¿Replanteó las técnicas de anonimización propuestas debido al incumplimiento de las propiedades estadísticas esperadas? En la reevaluación de los riesgos de identificación, ¿encontró nuevas unidades de observación riesgosas?, ¿cuántas?*" Fuente: DANE, 2018: 61

Finalmente, con la creación del IFPA el proceso de anonimización se dará por concluido y se tendrá un procedimiento debidamente documentado lo que permite que el proceso de anonimización se aplique a futuras bases de datos de la misma operación estadística o del mismo registro administrativo. Por otra parte, en caso de que el proceso no sea viable el informe incluirá las razones normativas, temáticas y procedimentales por las cuales el procedimiento de anonimización no es viable. En esta etapa si el proceso fue viable se obtendrá el siguiente producto:

Producto Etapa VI:
- Base de datos anonimizada definitiva
- IFPA

Fuente: DANE, 2018: 63.

7 DOCUMENTOS DE REFERENCIA

- DANE (2018). Guía para la Anonimización de Bases de Datos en el Sistema Estadístico Nacional.
- DANE (2017). Código Nacional de Buenas Prácticas para las Estadísticas Oficiales.
- P16.GTI Procedimiento de Liberación de Información
- P17.GTI Procedimiento para el procesamiento y generación de información estadística,
- P18.GTI Procedimiento para generar análisis de información oficial

Cualquier copia impresa de este documento se considera como COPIA NO CONTROLADA.

- MINSALUD (2016). Lineamientos para la Anonimización de Datos del Sistema Nacional de Estudios y Encuestas Poblacionales para la Salud. Ministerio de Salud y Protección Social. Dirección de Epidemiología y Demografía.
- Unión Europea (1997). Directiva 95/46/CE del Parlamento Europeo. Agencia de Protección de Datos.
- Hundepool Anco et al. (2012). Statistical Disclosure Control, John Wiley & Sons.
- Templ, Matthias, Bernhard Meindl, Alexander Kowarik, and Shuang Chen. (2014). Introduction to Statistical Disclosure Control (SDC). IHSN Working Paper No. 007.

8 CONTROL DE CAMBIOS

Fecha	Versión	Descripción del Cambio
NA	NA	N/A

PÚBLICA