GOBIERNO DE COLOMBIA

DANE INFORMACIÓN ESTRATÉGICA

DIRECCIÓN DE REGULACIÓN, PLANEACIÓN, ESTANDARIZACIÓN Y NORMALIZACIÓN (DIRPEN)

REGULACIÓN / NORMAS Y ESTÁNDARES

Guía para la anonimización de bases de datos en el Sistema Estadístico Nacional

Agosto 2018

DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA - DANE

MAURICIO PERFETTI DEL CORRAL
Director

CARLOS FELIPE PRADA LOMBO
Subdirector

MARY LUZ CARDENAS FONSECA
Secretaria General

Directores:

MIGUEL ÁNGEL CÁRDENAS CONTRERAS
Dirección de Geoestadística (DIG)

ANDREA CAROLINA RUBIANO FONTECHA
Dirección de Metodología y Producción Estadística

EDUARDO EFRAÍN FREIRE DELGADO
Dirección de Censos y Demografía

ANA PAOLA GÓMEZ ACOSTA
Dirección de Regulación, Planeación, Estandarización y Normalización

GIOVANNI BUITRAGO HOYOS
Dirección de Síntesis y Cuentas Nacionales

RAMÓN RICARDO VALENZUELA GUTIÉRREZ
Dirección de Difusión, Mercadeo y Cultura Estadística

Equipo de trabajo

Dirección de Regulación, Planificación, Estandarización y Normalización (DIRPEN):

ANA PAOLA GÓMEZ ACOSTA
Directora Técnica

SHEILA ISABEL CENTENO MARTINEZ
Coordinadora Investigación

JUAN CAMILO CALDERÓN
JENNIFFER ESCUDERO
Coordinación Investigación

JOSÉ ANDERSON CASTAÑEDA
Coordinación de Regulación

LAURA LÓPEZ FONSECA
Asesora

ALBA LIZETH PABÓN PUSEY
Diseño editorial

Agosto 2018

TABLA DE CONTENIDO

LISTA DE TABLAS

INTRODUCCIÓN

Los Sistemas Estadísticos Nacionales (SEN) tienen como propósito proveer información estadística oficial, relevante, comprensiva, confiable y objetiva a los diferentes usuarios, siendo la información un elemento fundamental para la toma de decisiones. Este propósito, al mismo tiempo, implica importantes retos para los productores de información estadística; especialmente al observar una mayor demanda de información desagregada, un aumento en el uso microdatos y, por tanto, una mayor necesidad de explotar los datos disponibles en los sistemas estadísticos.

El SEN colombiano establece dentro de sus objetivos promover, entre sus miembros, el acceso y uso de microdatos para la producción y difusión de estadísticas oficiales (Decreto 1743 de 2016: Art. 2.2.3.1.2). De igual forma, el Código Nacional de Buenas Prácticas Estadísticas del SEN, en su principio 10 sobre accesibilidad de la información, incentiva a los miembros del SEN para que implementen prácticas que permitan 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: 10). De igual forma, el principio 11 del Código que trata sobre la confidencialidad, pretende incentivar en los productores de información estadística, la utilización de técnicas para la anonimización de los microdatos, garantizando la protección de la identificación o localización geográfica de las fuentes empleadas en el proceso estadístico (DANE; 2017: 11).

Bajo estas premisas, el Departamento Administrativo Nacional de Estadística (DANE), en su rol de coordinador del SEN, presenta 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 se construye a partir de la experiencia del DANE al implementar procesos propios de anonimización en distintas bases de datos. Se espera que los integrantes del Sistema puedan identificar en este documento, buenas prácticas, herramientas e instrumentos cuando se implementen procesos de anonimización para la producción de sus propias estadísticas, así como para otros usos de la información anonimizada. La guía también presenta a lo largo de las etapas, la ejemplificación del proceso mediante el uso de una base de datos simulada, siendo este un insumo para que las entidades del SEN interesadas puedan seguir paso a paso el proceso y comparar los resultados que se presentan aquí.

La Guía se divide en siete partes, siendo la introducción la primera de ellas. En la segunda parte, se presentan algunos antecedentes internacionales y nacionales sobre los usos de la anonimización y algunos beneficios. Posteriormente, se relaciona el alcance

del documento frente a los integrantes del SEN y los conceptos sobre la anonimización, buscando identificar la importancia y potencialidad de su uso en la producción y difusión de estadísticas.

Posteriormente, se presenta el proceso de anonimización adelantado desde el DANE, que podrá ser de utilidad para los productores de estadísticas del SEN, identificando en este apartado seis etapas que son: i) revisiones previas de la información a anonimizar; ii) análisis de riesgos del proceso; iii) selección de técnicas para avanzar en la anonimización; iv) análisis de viabilidad; v) aplicación de la técnica de anonimización en la información y vi) la evaluación de resultados del proceso.

Finalmente, en la sexta parte se presentan algunas recomendaciones finales y la séptima parte las referencias bibliográficas.

1. ANTECEDENTES INTERNACIONALES Y NACIONALES EN EL USO DE LA ANONIMIZACIÓN

Son varias las experiencias internacionales que permiten observar la implementación de la anonimización; su principal propósito es incrementar la desagregación de la información, manteniendo los niveles de confidencialidad, así como generar un mayor aprovechamiento estadístico de la misma. Esta sección presentará algunas experiencias de países como Reino Unido, Estados Unidos y Holanda y los avances que se han tenido en el contexto nacional.

1.1. Contexto internacional

La mayor parte de las experiencias internacionales sobre la anonimización de información estadística parten de las indicaciones o recomendaciones propias de los países establecidas en sus sistemas nacionales de estadística. Este proceso se ha observado también como un posible mecanismo para dar seguimiento a la legislación internacional en materia de protección, privacidad y confidencialidad de la información.

En Reino Unido, por ejemplo, la Oficina del Comisionado de Información desarrolló el Código de Buenas Prácticas para la Anonimización (Oficina del Comisionado de información , 2012), en el cual se presentan los antecedentes jurídicos de la protección de datos de ese país; explica los beneficios de la anonimización y el por qué se debe hacer, todo enmarcado en los principios que deben guiar dicho proceso. Se señalan también los riesgos que se pueden presentar al cruzar información en bases que ya se encuentran anonimizadas, generando posibles identificaciones de usuarios y por qué se debe realizar el control de los datos en el sistema estadístico en este país.

Otro esfuerzo internacional destacable en la documentación de procesos de anonimización para orientar a las entidades que producen estadísticas en los Sistemas Estadísticos Nacionales de los distintos países, es el realizado por la Comisión Económica para Europa de las Naciones Unidas (UNECE, por sus siglas en inglés), este organismo publicó el *Manual sobre el control de divulgación estadística*, en el cual se proveen lineamientos técnicos para el control de la revelación de la información, describiendo los métodos aplicables para la protección de la privacidad de la información y explicando detalladamente el programa de anonimización ARGUS, una iniciativa que viene liderando UNECE para implementar mecanismos de anonimización en los productores de estadística.

Igualmente, la Red Internacional de Encuestas de Hogares (IHSN por sus siglas en inglés), publicó en 2014 una introducción sobre los controles a tener en cuenta en la divulgación estadística de información para uso de los integrantes del SEN, explicando los métodos para difundir la información de datos confidenciales, teniendo en cuenta los distintos riesgos a los cuales se enfrentan los productores de información estadística, así como posibles mecanismos para valorar este tipo de riesgos, y describe los métodos de anonimización (IHSN, 2014).

Por otra parte, la anonimización permite atender de una mejor manera las demandas y necesidades de información de distintos usuarios. Algunos de los sectores que se han caracterizado por generar este tipo de requerimientos, respecto a mayores desagregaciones de la información; especialmente en países como Estados Unidos, han sido investigadores, centros de investigación, universidades, y el sector público.

Precisamente, en este último país, se ha observado una fuerte relación entre la información y la investigación, siendo esta, uno de principales focos para el desarrollo de procesos de anonimización. La Oficina de Censos de Estados Unidos ha liderado el estudio de técnicas de anonimización e integración de información, así como la opción de acceso a microdatos no anonimizados con propósitos académicos y de investigación. Algunas de las técnicas implementadas se basan, por ejemplo en eliminar las variables de identificación directa, utilizar técnicas como umbrales geográficos o categóricos de las variables, redondeo, infusión de ruido, recodificación, intercambio de registros basado en rangos o en proximidad (Morales, 2017: 9).

Dentro de las bases anonimizadas publicadas por esta institución estadounidense, se destacan las de censos demográficos y de encuestas económicas y, para el caso del Censo Agropecuario, se disponen al público algunas tablas agregadas por Estado y Condado. Todo esto bajo el marco de la normatividad nacional e internacional de confidencialidad de la información de los usuarios.

Otra experiencia en este sentido es la de la Oficina de Estadísticas de Holanda (CBS, por sus siglas en holandés), entidad que ha trabajado conjuntamente con la UNECE (por sus siglas en inglés) para el avance de la investigación sobre procesos de anonimización de la información; destacándose de esta cooperación, la implementación de la iniciativa, *Control de Divulgación Estadística*, un estudio exhaustivo sobre la importancia de la confidencialidad, que busca mediante la implementación de proyectos, generar mecanismos que permitan garantizarla.

Uno de los proyectos adelantados en este marco, es el desarrollo del software µ-ARGUS, cuyo objetivo se centra en realizar un control de divulgación de microdatos, acompañado del paquete τ-ARGUS que realiza el manejo de datos tabulares. 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 (Morales, 2017:10).

1.2. Antecedentes en Colombia

En Colombia, el DANE ha sido pionero en el desarrollo de los procesos de anonimización de microdatos. En 2014 se publicaron los *Lineamientos para la anonimización de microdatos* que identificaban tres grandes etapas en el proceso: preanonimización, anonimización de microdatos de uso interno y anonimización (DANE, 2014: 11).

El Ministerio de Salud, basado en este documento del DANE, generó sus propios lineamientos para anonimizar microdatos a través del documento *Lineamientos para la Anonimizacion de Datos del Sistema Nacional de Estudios y Encuestas Poblacionales para la Salud* (Ministerio de Salud, s.f.). En este ejercicio, el Ministerio incluyó las etapas previstas por el DANE y agregó con mayor detalle, las técnicas de anonimización que se usarían en sus bases de información; especificamente presenta el uso de los métodos de perturbación o los métodos de reducción en la información (Ministerio de Salud, s.f.:12-13).

Adicional a los lineamientos y orientaciones en materia de anonimización, Colombia ha gestionado la implementación de un marco legal en esta materia, partiendo del derecho a la intimidad personal y familiar y el derecho a conocer, actualizar y rectificar las informaciones que se hayan recogido en las bases de datos y en archivos de entidades públicas y privadas, que aparece consignado en el artículo 15 de la Constitución Política de 1991.

En términos de la confidencialidad, enmarcados en las directrices para el SEN, es importante tener en cuenta la Ley Estatutaria 1266 de 2008 que establece el *Habeas Data* y regula el manejo de la información contenida en bases de datos personales, financieras,

crediticias, comerciales, de servicios y provenientes de terceros países, además de establecer disposiciones sobre la recolección, tratamiento y circulación de datos personales en el país (Congreso de Colombia, 2008).

Junto con estas leyes se cuenta con la Ley 1581 de 2012, que trata sobre la protección de datos personales, reglamentada parcialmente por el decreto 1377 de 2013, en el que se señala que el acceso a los datos se debe restringir y la información debe estar sujeta a tratamiento por parte del responsable, como lo indica en su artículo 4, manteniendo los principios de acceso y circulación restringida, de seguridad y de confidencialidad. Para el DANE, particularmente, la Ley 79 de 1993 establece la reserva estadística, en la cual se establece mantener la confidencialidad de las fuentes cuando se realizan procesos de recolección de información a través de censos o encuestas.

Respecto a la transparencia y el acceso de la información pública, la Ley 1712 de 2014, regula el derecho de acceso a la información pública, haciendo énfasis en el establecimiento de una política de datos abiertos por parte de las entidades públicas. Ademásel Decreto 2573 de 2014 del Ministerio de Tecnología de la información y las Comunicaciones (MinTIC), establece los ***Lineamientos generales de la Estrategia de Gobierno en línea***, indicando los principios y fundamentos a tener en cuenta en las entidades públicas destacando entre ellos, la excelencia en el servicio ciudadano, la apertura y reutilización de datos públicos, la estandarización, la innovación, entre otros. Igualmente, se establecen cuatro componentes que facilitarán la masificación de la oferta y la demanda en gobierno en línea, resaltando entre estos, la seguridad y la privacidad de la información, siendo un componente transversal.

Estas normas han tenido su complemento con la reglamentación del SEN, a través de la Ley 1753 de 2015 y del Decreto 1743 de 2016. En este último, se establecen entre los varios objetivos para el SEN, el fortalecimiento y aprovechamiento amplio e intensivo de los registros administrativos como fuente para la producción de estadísticas oficiales y la promoción al acceso y uso de microdatos para la producción y difusión de estadísticas oficiales y el fortalecimiento de la calidad y coherencia de las mismas (DANE, 2016).

Finalmente, en 2017 con la actualización del Código Nacional de Buenas Prácticas del SEN, se plantea la implementación de prácticas en materia de acceso y confidencialidad de la información por parte del SEN, incentivando un mayor acceso y uso de la información de los productores de estadísticas en Colombia, así como la difusión de información anonimizada, garantizando la confidencialidad de la misma.

2. OBJETIVO Y ALCANCE DE LA GUÍA

La *Guía para la anonimización de bases de datos en el Sistema Estadístico Nacional* tiene como propósito orientar a las entidades integrantes del SEN sobre el proceso de anonimización de bases de datos que provienen de registros administrativos y de operaciones estadísticas.

De esta manera esta guía puede ser consultada por:

- ✓ Entidades que producen y difunden información estadística mediante operaciones estadísticas por muestreo, censos, a partir de registros administrativos y operaciones estadísticas derivadas.
- ✓ Entidades que poseen o son responsables de registros administrativos.

3. MARCO CONCEPTUAL PARA LA 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 (Morales, 2017: 5).

El concepto de **microdato** es fundamental en la concepción del proceso de anonimización; este se entiende como:

> "… 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" (Decreto 1743 de 2016: Art. 2.2.3.1.1).

A su vez, el Sistema de Consulta de Conceptos del DANE define que una **base de datos** hace referencia a:

> "… un conjunto o colección de datos interrelacionados entre sí, que se utilizan para la obtención de información de acuerdo con el contexto de los mismos y que son almacenados sistemáticamente para su posterior uso"
> (Sistema de Consultas de Conceptos del DANE, 2018)

Según lo anterior, el proceso de anonimización se aplica a aquellos datos que, por su naturaleza, son sensibles al público debido a la posibilidad de que sea violada la confidencialidad de la información.

El proceso de anonimización puede aplicarse tanto a los microdatos obtenidos de **operaciones estadísticas como a los de los registros administrativos** que posee una entidad.

Respecto al primer tipo de datos, el Decreto 1743 de 2016 ha definido operación estadística como la "aplicación de un proceso estadístico sobre un objeto de estudio que conduce a la producción de información estadística" (Decreto 1743 de 2016: Art. 2.2.3.1.1).

El **proceso estadístico** se entiende como un:

> "… conjunto sistemático de actividades encaminadas a la producción de estadísticas que comprende, entre otras, la detección de necesidades, el diseño, la recolección, el procesamiento, el análisis y la difusión" (Decreto 1743 de 2016: Art. 2.2.3.1.1).

Por tanto, el proceso de anonimización que se aplique a las operaciones estadísticas, debe contar con una metodología y documentación a lo largo de las fases de producción, buscando garantizar la calidad de la información estadística a generar.

La información que se obtiene como producto del proceso estadístico, a nivel de los datos de las unidades de observación (hogares, personas, empresas, establecimientos, entre otros) sirve como insumo para que los usuarios puedan aprovechar estadísticamente la información de dichas operaciones, así como generar estudios o investigaciones propias que permitan mejorar la toma de decisiones, además de otros usos que consideren pertinentes los responsables de la información a anonimizar.

De manera que la anonimización permite fomentar en las entidades del SEN, la transparencia de la información, priorizando en este caso, la preservación de la confidencialidad.

Frente a los registros administrativos, estos se entienden como un:

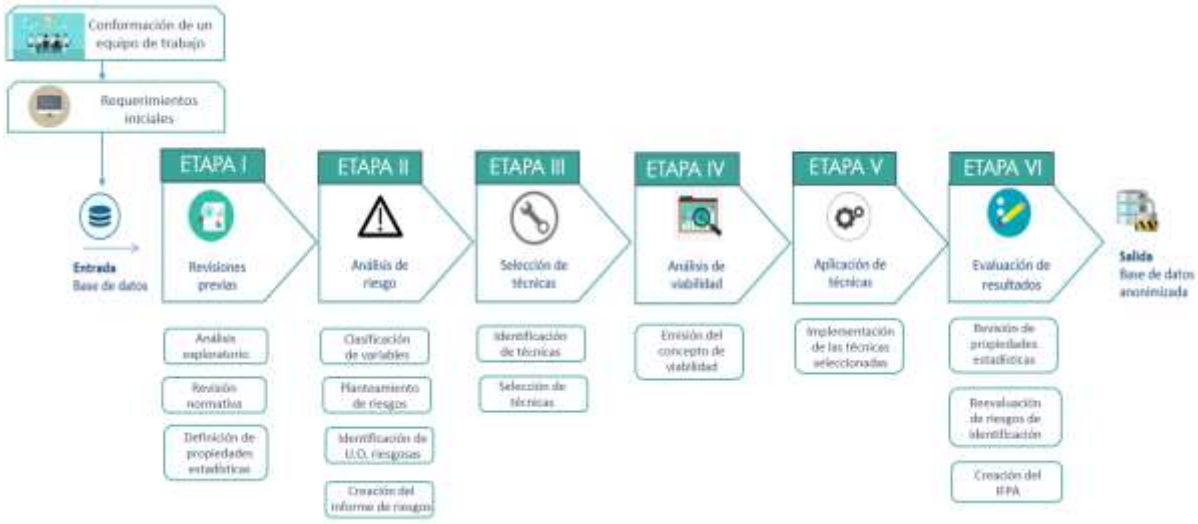
> "…Conjunto de datos que contiene la información recogida y conservada por entidades u organizaciones en el cumplimiento de sus funciones o competencias misionales" (Decreto 1743 de 2016: Art. 2.2.3.1.1).

Para esta fuente en particular, las entidades del SEN pueden contar con procesos de diagnósticos que permiten implementar prácticas para mejorar la captura de la información, mediante la implementación de indicadores de calidad sobre el registro de la entidad[1]. Los resultados de los diagnósticos de los registros administrativos, permitirán al mismo tiempo, desarrollar mejores procesos de anonimización y obtener mejoras en la publicación de su información estadística.

4. PROCESO DE ANONIMIZACIÓN DE BASES DE DATOS

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. Estas fases se presentan en la Gráfica 1:

Gráfica 1. Esquema general del proceso de anonimización



Fuente: DANE-DIRPEN

Estas etapas presentan distintos subprocesos y actividades a desarrollar para realizar la anonimización de la base de datos deseada.

En la medida en que es un proceso de anonimización, los contenidos de las fases se encuentran orientados a presentar entradas y salidas, así como recomendaciones sobre pasos previos que permitirán preparar a la entidad al iniciar el proceso de anonimización.

[1] Para mayor información sobre los procesos de diagnóstico de los registros administrativos, se puede consultar la *Metodología de diagnóstico de los Registros administrativos para su Aprovechamiento estadístico del DANE.* Disponible en: http://www.dane.gov.co/files/sen/registros-administrativos/Metodologia-de-Diagnostico.pdf

Igualmente, a lo largo de las etapas, se realizan indicaciones y recomendaciones de documentación que permitan conocer la trazabilidad del proceso de anonimización dentro de la entidad. Al finalizar la etapa 6 del proceso, podrá encontrarse un modelo de ***Informe Final del Proceso de Anonimización – IFPA***, el cual recoge todas las recomendaciones de documentación cuando se desarrolla la anonimización de la base de datos.

4.1. Pasos previos al proceso de anonimización

El proceso de anonimización debe ser ejecutado por un **equipo de trabajo** que tenga acceso y conocimiento de la base de datos a anonimizar. Es importante que este equipo de trabajo cuente con la capacidad de:

- **Conocer temáticamente el contenido de la base de datos a anonimizar**: por ejemplo, si la base es insumo de una operación estadística de un área temática particular, es importante que el equipo de trabajo se encuentre conformado por una persona (o varias) que conozca(n) el fenómeno registrado en la base de datos. La vinculación de este personal especializado tiene como fin apoyar la toma de decisiones de la anonimización a desarrollar, principalmente en las etapas de riesgos de identificación (Etapa II) y de análisis de viabilidad del proceso (Etapa IV).

- **Manejar herramientas que permitan el análisis exploratorio de datos:** en este caso se necesitará que el equipo cuente con miembros con habilidades para el manejo de paquetes estadísticos como *R, SAS, SPSS, Stata*, entre otros. Igualmente, se recomienda que conozcan técnicas estadísticas que permitan apoyar la etapa de análisis exploratorio de la base de datos (Subproceso de la Etapa I) y la etapa de aplicación de las técnicas de anonimización (Etapa V).

Cuando la entidad conforme estratégicamente su equipo de trabajo, éste deberá tener en cuenta que existen **requerimientos iniciales** que permiten la ejecución satisfactoria del proceso de anonimización. Estos son:

- ✓ **Disponer de una base de datos:** la base definida por la entidad y que será difundida para el SEN. Esta puede contener una o varias tablas relacionadas entre sí[2]. La entidad debe saber si la base de datos es resultado de una operación estadística o es la resultante de un registro administrativo (Recuadro 1).

[2] Modelo Entidad-Relación de la base de datos.

- ✓ **Contar con el diccionario de datos de la base a anonimizar[3]:** este documento debe especificar claramente las propiedades básicas de las variables contenidas en la base de datos. Algunas de estas propiedades son: nombre, longitud, obligatoriedad de respuesta, descripción, reglas de validación, entre otros, así como la relación entre ellas. Un ejemplo sobre el uso de diccionario de datos se encuentra disponible en el Anexo A.

- ✓ **Disponer de una infraestructura tecnológica:** la infraestructura definida por la entidad debe permitir el manejo estadístico de datos; en este caso, debe tener en cuenta 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. Cuando las bases de datos son de grandes dimensiones, el equipo de trabajo debe tener en cuenta que algunos paquetes de *software* no son compatibles; por lo cual requerirá revisiones sobre programas estadísticos y el alcance de estos sobre grandes cantidades de información.

- ✓ **Definir mecanismos de seguridad sobre la base de datos a anonimizar**: la entidad debe prever condiciones mínimas para salvaguardar la información, así como definir protocolos de acceso a la información por parte del equipo de trabajo que participará en el proceso de anonimización. Por ejemplo, acuerdos de confidencialidad del personal involucrado en el proceso (equipo de trabajo) debidamente firmados, usos de permisos y contraseñas para el uso de la información, entre otros.

[3] Diccionario ejemplo dispuesto por el DANE disponible en el Anexo A.

Recuadro 1. Verificación especial de la base de datos de acuerdo a su origen

La entidad del SEN debe verificar si la base de datos a anonimizar es la resultante de alguna operación estadística o de un registro administrativo.

Caso1: Resultado de operaciones estadísticas;

- ✓ La entidad tendrá que verificar que la base de datos para la anonimización es la resultante de la finalización de la Fase de Ejecución del Proceso Estadístico. Esto significa que la consistencia de los datos recolectados debe haber sido validada, según: i) las técnicas definidas por la entidad responsable; ii) las variables creadas (o calculadas) a partir de la información recolectada; y, iii) que se encuentren en la base de datos. En caso de valores faltantes, tendrá que verificar que los métodos de imputación hayan sido aplicados.

Caso 2: Base de datos de registros administrativos

- ✓ La entidad debe verificar que la base de datos sea la versión más actual. En este caso, se recomienda que el periodo de tiempo del registro administrativo que se anonimizará, se defina con base en la periodicidad de consolidación de la información. Además, se recomienda revisar la consistencia y calidad de la base de datos teniendo en cuenta la sección Revisión de la consistencia de la base de datos de la Metodología de Diagnóstico de Registros Administrativos.

Ejemplo:

El Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales (SISBEN) tiene una periodicidad de recolección diaria y de consolidación de la información en el aplicativo SISBENNET mensual. En este caso, se recomendaría que SISBEN realice un corte para anonimizar la base de datos de un periodo, teniendo como fecha de cierre la última consolidación de la información.

Fuente: Elaboración propia

Después de que el equipo de trabajo confirme la disponibilidad de los *requerimientos previos*, empezará con la primera etapa del proceso de anonimización.

Insumos:
- Equipo de trabajo definido
- Base de datos
- Diccionario de datos completo
- Infraestructura definida

4.2. Etapa I: Revisiones previas al proceso de anonimización

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:

- ✓ Análisis exploratorio de la base de datos
- ✓ Revisión normativa sobre protección de datos e identificación de usuarios de la información
- ✓ Definición de las propiedades estadísticas a conservar en la base de datos

4.2.1. *Análisis exploratorio de la base de datos*

En este subproceso se busca caracterizar la base de datos a anonimizar, teniendo en cuenta para ello, aspectos procedimentales y temáticos. Está compuesto de cuatro pasos:

Caracterización de la base de datos

Para iniciar con la etapa, el equipo de trabajo caracterizará cada una de las variables contenidas en la base de datos teniendo en cuenta si son cuantitativas (continuas o discretas) o categóricas (Recuadro 2).

Recuadro 2. Clasificación de las variables por su tipo

Las variables, en general, se pueden clasificar por el tipo de valores que toman:

- **Variable continua:** Es una característica medida en las unidades de observación que puede tomar valores dentro de un intervalo específico (infinito) (Sistema de Consulta de Conceptos del DANE, 2018).

  Por ejemplo: la altura medida en centímetros, el peso medido en kilogramos, la temperatura medida en grados centígrados, entre otros.

- **Variable discreta:** Es una característica medida en las unidades de observación donde su conjunto de posibles valores, no corresponde a un intervalo ya que presenta interrupciones (Sistema de Consulta de Conceptos del DANE, 2018).

  Por ejemplo: número de hijos, número de intentos en un experimento, número de personas de un hogar, entre otros.

- **Variable categórica:** Es una característica medida en las unidades de observación que asigna una de varias categorías cualitativas (Sistema de Consulta de Conceptos del DANE, 2018).

  Por ejemplo: una variable dicotómica, que asigna el valor de 1 si el individuo cumple cierta característica y 0 si no la cumple, o estado civil, que asigna a cada individuo alguna de las categorías: soltero, casado, separado, divorciado o unión libre.

Para desarrollar este subproceso, el equipo de trabajo, con ayuda del diccionario de la base de datos (o de toda la documentación disponible), describirá las dimensiones de la base en términos del número de variables y número de registros y la distribución de todas las variables con respecto a su tipo (cuantitativa o categórica).

Otra clasificación que el equipo de trabajo debe tener en cuenta para las variables de la base de datos es el tipo de información que contiene de la unidad de observación. Estas se clasifican en variables de identificación, de ubicación y temáticas. Por ejemplo, el número de identificación de una persona, es una variable de identificación; el municipio; la localidad y el departamento de una entidad, son variables de ubicación; y, el ingreso promedio mensual, el nivel de escolaridad y el número de hijos, son variables temáticas.

La caracterización de la base de datos se puede hacer teniendo en cuenta como ejemplo la Tabla 1.

Tabla 1. Clasificación de variables de base de datos

VARIABLES/TIPO DE VARIABLE	CUANTITATIVAS	CATEGÓRICAS	TOTAL
IDENTIFICACIÓN	*Escriba en este espacio el número de variables cuantitativas de identificación*		
UBICACIÓN		*Escriba en este espacio el número de variables categóricas de ubicación*	
TEMÁTICAS			*Escriba en este espacio el número de variables temáticas*
TOTAL	*Escriba en este espacio el número de variables cuantitativas*		*Escriba en este espacio el total de variables de la base de datos*

Fuente: DANE- DIRPEN

Cálculo de medidas descriptivas de las variables cuantitativas

Después de que el equipo realice la clasificación de las variables de la base de datos, procederá a calcular las medidas descriptivas estadísticas para cada una de las variables cuantitativas como se ilustra en la siguiente tabla.

Tabla 2. Principales medidas descriptivas para variables cuantitativas

VARIABLE CUANTITATIVA	MEDIA	DESVIACIÓN ESTÁNDAR	MÍNIMO	CUARTIL 1	CUARTIL 2	CUARTIL 3	MÁXIMO
Escriba en este espacio el nombre de la variable cuantitativa				*Escriba en este espacio el primer cuartil de los datos de la variable*			

Fuente: DANE-DIRPEN

Estas medidas servirán como insumo para el análisis de riesgos (Etapa II), el análisis de viabilidad del proceso de anonimización (Etapa IV) y la evaluación de resultados (Etapa VI). Este tipo de medidas se conocen como propiedades globales de las variables y estarán sujetas a verificación por parte del equipo de trabajo para examinar el resultado del proceso de anonimización.

Cálculo de frecuencias de las variables categóricas

Las distribuciones de frecuencias para las variables categóricas hacen parte de las medidas para verificar el proceso de anonimización de la base de datos. Estas distribuciones pueden realizarse teniendo en cuenta la siguiente tabla:

Tabla 3. Distribución de frecuencias para una variable con dos categorías

VARIABLE CATEGÓRICA	NÚMERO DE UNIDADES DE OBSERVACIÓN QUE CUMPLEN LA CATEGORÍA	PORCENTAJE DE REGISTROS QUE CUMPLEN LA CATEGORÍA
Categoría 1		*Escriba en este espacio el porcentaje de unidades de observación que cumplen con la categoría 1*
Categoría 2	*Escriba en este espacio el número de unidades de observación que cumplen con la categoría 2*	
TOTAL		

Fuente: DANE-DIRPEN

Revisión temática del contenido de la base de datos

El equipo de trabajo, después de analizar cada una de las variables, debe realizar una revisión temática de la base de datos a anonimizar, teniendo en cuenta la documentación de la operación estadística o el registro administrativo. Para ello se sugiere responder las siguientes preguntas:

- ¿Cuál es el objetivo de la operación estadística (o del registro administrativo)?
- ¿Qué cambios metodológicos ha presentado la operación estadística (o registro administrativo) a través del tiempo?
- ¿Qué tipo de estándares, clasificaciones o nomenclaturas siguen las variables de la base de datos obtenida por la operación estadística o por el registro administrativo? ¿Se están usando adecuadamente?
- ¿Qué documentos existen acerca de la operación estadística (o registro administrativo)? ¿Son de fácil acceso para el equipo de trabajo que realizará la anonimización?
- ¿Existen otras operaciones estadísticas (o registros administrativos) relacionadas con la temática de la base de datos?

Esta revisión temática servirá como insumo en el planteamiento de los riesgos de identificación de las unidades de observación (Etapa II) y en el análisis de viabilidad del proceso (Etapa IV). Con esto, finaliza el subproceso de *Análisis exploratorio de la base de datos*.

4.2.2. *Revisión normativa sobre protección de datos e identificación de necesidades de información*

Este subproceso busca realizar una revisión normativa que pueda afectar la publicación de la información sujeta a anonimizar; así mismo, se sugiere identificar las necesidades de información que presentan los usuarios sobre la base de datos. El subproceso se compone de dos pasos:

1. Revisión de restricciones de publicación de la información
2. Revisión de las necesidades de los usuarios de la información

Revisión de restricciones de publicación de la información

El equipo de trabajo debe realizar una revisión de la normatividad que puede afectar la publicación de la información sujeta a ser anonimizada; en este caso es importante que la entidad verifique las normas y cláusulas de confidencialidad que tengan alcance en la información estadística que se espera dejar disponible para el SEN.

Para iniciar esta actividad, el equipo de trabajo tendrá que realizar la revisión de leyes, decretos, resoluciones, convenios institucionales, acuerdos de confidencialidad de la información, estatutos y toda la normatividad que fundamenta el origen del registro administrativo o de la operación estadística de la entidad. El resultado de este análisis determinará si existe alguna norma que impida la publicación de la información de la base de datos a anonimizar.

Un insumo a tener en cuenta en esta actividad para el equipo de trabajo es el alcance de las normas asociadas a la confidencialidad de la información que tiene impacto sobre las entidades del SEN; por ejemplo el principio de confidencialidad descrito en la Ley Estatutaria 1266 de 2008 o el artículo 2.2.3.3.5 del Decreto 1743 de 2016, en donde se establecen indicaciones sobre la no exposición de la identificación y ubicación de las unidades de observación al momento de publicar información (Recuadro 3.).

Recuadro 3. Normas Referentes a la confidencialidad de la información

Ley Estatutaria 1266 de 2008: El principio de confidencialidad descrito en la ley indica que "Todas las personas naturales o jurídicas que intervengan en la administración de datos personales que no tengan la naturaleza de públicos están obligadas en todo tiempo a garantizar la reserva de la información".

El artículo 2.2.3.3.5 del decreto 1743 de 2016 establece que "las entidades que conforman el SEN …, deberán guardar la confidencialidad de los datos que permitan la identificación y/o localización espacial de las fuentes, cuando estos fueren recolectados exclusivamente para la producción de las estadísticas oficiales y para fines estadísticos".

Después de la revisión normativa y de tener en cuenta los principios de confidencialidad, el equipo de trabajo describirá los hallazgos de la correspondiente revisión. Esta descripción es necesario incluirla en el informe final de anonimización; en ella puede describir aspectos como:

- ✓ **Normatividad identificada** como decretos, leyes, artículos, entre otros, que impidan la publicación de la información.
- ✓ **Normas que respalden la publicación de la información** a los diferentes usuarios y demás entidades del SEN.
- ✓ **Observaciones y sugerencias** a tener en cuenta al momento de realizar la publicación de la información, para evitar sanciones legales y judiciales a la entidad del SEN.
- ✓ **Personal y fecha en la cual se hizo la revisión de la normatividad** esto con el fin de presentar una trazabilidad del trabajo sobre la base de datos a anonimizar.

Revisión de las necesidades de los usuarios de la información

Este paso tiene como propósito realizar una caracterización de las necesidades de información de los usuarios sobre la base de datos a anonimizar. En este caso, el equipo de trabajo de la entidad del SEN debe tener en cuenta las demandas o requerimientos realizados por parte de los usuarios.

Para el caso en que la entidad del SEN cuente con demandas de información por distintos usuarios, se recomienda elaborar un listado que contenga los requerimientos solicitados, las fechas de solicitud, las variables requeridas, la frecuencia con la que se hacen las solicitudes y las respuestas dadas a dichos requerimientos.

Se recomienda al equipo de trabajo realizar una revisión del histórico de las solicitudes, menor a dos años, teniendo en cuenta el volumen de las solicitudes recibidas, la recurrencia y la frecuencia que tiene cada solicitud. Algunas de las solicitudes que podrá revisar el equipo de trabajo se presentan a continuación:

- ✓ Solicitudes de información recibidas por distintos usuarios: ciudadanos, academia, empresas, entidades públicas del orden nacional o territorial, organismos internacionales, entre otros.
- ✓ Derechos de petición radicados en la entidad durante el último año.
- ✓ Necesidades de información propias del área temática de la entidad.
- ✓ Otro tipo de requerimientos

Con la información recolectada, el equipo de trabajo podrá clasificar y cuantificar las demandas de información, teniendo en cuenta:

- Tipo de usuarios
- Tipo de solicitudes
- Variables solicitadas
- Nivel de desagregación requerido de la información
- Periodos de la información (tiempo en años) o bases con determinados cortes
- Frecuencia de las solicitudes
- Objetivo, finalidad, uso de la información requerida

El Recuadro 4 presenta un ejemplo de clasificación de este tipo de solicitudes de información.

Recuadro 4. Ejemplo de clasificación de solicitudes recibidas

Categoría	Descripción	No de solicitudes
Tipo de usuario	Entidades Públicas	4
	Entidades Privadas	3
	Investigadores	20
	Instituciones Académicas	15
	Otros	8
Tipo de Solicitudes	Derechos de petición	1
	Acceso a información	30
	Actualización de datos	19
Clasificación de variables	Variables de identificación	15
	Variables de ubicación	10
	Variables temáticas	25
Niveles de desagregación	Nacional	3
	Departamental	10
	Municipal	30
	Temática	5
	Otros	2
Períodos solicitados de la información	Anual	16
	Mensual	30
	Trimestral	20
	Diaria	23
Frecuencia de la solicitud	Mensual	120
	Diaria	26
Objetivo, finalidad, uso	Investigaciones o estudios	36
	Académica	12
	Política pública	59

Fuente: DANE- DIRPEN

Basados en la clasificación y cuantificación de las solicitudes, el equipo de trabajo podrá identificar las variables, los niveles de desagregación, los periodos (tiempo en años) de la información de la base de datos que tienen mayor demanda lo cual le dará información para definir la base de datos y desagregaciones que satisficieran de la mejor forma las necesidades de los usuarios, y que, además, no expongan la identificación de las unidades de observación.

Una ventaja de publicar las bases anonimizadas, para la entidad del SEN, es la disminución de las cargas operativas para responder solicitudes de información que pueden ser repetitivas o recurrentes.

En el caso en que la entidad del SEN no tenga identificadas las demandas de información o desconozca los usuarios que potencialmente podrían usar su información estadística a nivel de microdato, el equipo de trabajo podrá tener en cuenta distintos recursos; algunos ejemplos son:

- **Vacíos de información del SEN:** Estos hacen referencia a los requerimientos de información nacional e internacional que el país requiere, y que actualmente no es producida por alguna entidad del SEN. Los vacíos de información priorizados en materia estadística se pueden consultar en el Plan Estadístico Nacional –PEN- vigente. El Plan cuenta con la oferta de operaciones estadísticas del país, así como los requerimientos de información que aún no se producen[4].

  Algunos ejemplos de compromisos internacionales previstos en el PEN hacen referencia a los requerimientos de información de la Organización para la Cooperación y el Desarrollo Económicos (OCDE); los Objetivos de Desarrollo Sostenible (ODS); la Organización de las Naciones Unidas (ONU), entre otros. Frente a los requerimientos nacionales, el PEN tienen en cuenta las necesidades de información de acuerdo a lo definido en la política pública, políticas sectoriales, cuentas macroeconómicas y la normatividad del país.

- **Prioridades de política:** Otro posible insumo para evaluar la potencialidad de información se encuentra asociada a las prioridades de política definida por el país; estas se encuentran fijadas en documentos como el Plan Nacional de Desarrollo; los documentos CONPES; y otros instrumentos de planeación del país para el mediano y largo plazo (por ejemplo, el Plan Nacional Decenal de Educación). En este tipo de documentos también se pueden identificar posibles necesidades de información que se tienen y los posibles usos de la información que produce la entidad del SEN.

  Uno de los documentos que ha generado un importante aporte para la producción de información a partir de registros administrativos y que da respuesta a una prioridad de mediano y largo plazo, es el Documento CONPES 3918 de 2018[5], el cual establece la hoja de ruta para dar respuesta a los Objetivos de Desarrollo Sostenible (ODS). En este documento se han fijado los indicadores nacionales para el seguimiento de los ODS[6], identificando la información que se requeriría en el monitoreo de las metas fijadas por el país.

[4] El Plan Estadístico Nacional 2017-2022; en su anexo B, presenta en total 229 vacíos de información que pueden ser consultados por la entidad del SEN. Disponible en el siguiente enlace: https://www.dane.gov.co/files/sen/PEN-2017-2022.pdf
[5] https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3918.pdf
[6] https://colaboracion.dnp.gov.co/CDT/Sinergia/Documentos/Indicadores_Globales_ODS_oficial.pdf

- **Sistemas de información del SEN**: Los sistemas de información también son una fuente de demanda de información estadística; en este caso se tienen ejemplos como el Sistema Unificado de Inversión y Finanzas Públicas del DNP, el cual hace seguimiento a la inversión pública del país, mediante el uso de indicadores; el Sistema Integrado de Información sobre Violencias de Género (SIVIGE) donde se definen indicadores que permitan monitorear distintos tipo de violencia en la población; el Sistema de Estadísticas en Justicia, el Sistema de Información Ambiental de Colombia (SIAC), entre otros, donde se reflejan algunos vacíos y posibles oportunidades de producción estadística.

Teniendo en cuenta la revisión de los diferentes recursos para encontrar un uso potencial, el equipo de trabajo de la entidad del SEN, tomará la decisión del tipo de información que puede suministrar en la base de datos anonimizada para responder a las necesidades de los usuarios.

Después de la revisión de las necesidades de los usuarios de la información, el equipo de trabajo, mediante una bitácora, indicará los resultados encontrados:

- ✓ Periodo de revisión de las solicitudes, peticiones, requerimientos de información revisado por el equipo de trabajo.
- ✓ Listado de variables más solicitadas, los niveles de desagregación más demandados.
- ✓ Listado de los usuarios que con mayor frecuencia hacen solicitudes de información.
- ✓ Los periodos o bases de datos con determinados cortes que más son solicitados.
- ✓ Listado de los diferentes usos identificados para los cuales son radicadas la mayoría de las solicitudes y requerimientos.

4.2.3. *Definición de las propiedades estadísticas a conservar en la base de datos*

En este subproceso el equipo de trabajo deberá establecer las propiedades estadísticas que se deben mantener en la base de datos anonimizada, en relación con la base de datos sin anonimizar. Algunas de estas propiedades estadísticas son:

- ✓ **Mantener tendencias en las variables a través del tiempo**: Esta propiedad hace referencia a que las variables conserven un comportamiento en determinados periodos de tiempo. Por ejemplo, si la base de datos de una operación estadística de temática económica contiene la variable *ingreso de los hogares colombianos* y esta variable ha presentado un comportamiento creciente en el primer trimestre de 2016, al publicar la base de datos anonimizada el equipo de trabajo desea garantizar que esta tendencia se conserve.

- ✓ **Mantener propiedades globales de las variables:** El equipo de trabajo debe definir cuáles de las medidas estadísticas descritas en el análisis exploratorio de datos (sección 4.2.1), para las variables categóricas y cuantitativas, se deben mantener sin variación y para qué niveles de desagregación geográfica o temática. Así mismo, debe decidir cuáles de las propiedades globales pueden presentar alguna variación significativa y hasta qué porcentaje de variación es permitido en la base de datos anonimizada.

  Por ejemplo, un equipo decidió que la propiedad global que desea mantener es el promedio de la variable "Ingreso por hogar". Además, aceptará el proceso de anonimización, solamente si el promedio de la variable en la base de datos anonimizada difiere del promedio en la base de datos sin anonimizar en menos del 1%.

- ✓ **Mantener cifras por niveles de desagregación geográfica o temática:** El equipo de trabajo debe definir cuáles medidas estadísticas se deben conservar sin variación en los niveles de desagregación geográfica o temática, para garantizar a los usuarios análisis de estadísticas más sectorizados.

  Por ejemplo, un equipo de trabajo decidió mantener para la variable grupos étnicos los totales de cada categoría a nivel departamental; esta propiedad permite caracterizar la población étnica en cada departamento y con esta información los usuarios pueden realizar análisis estadístico por regiones.

- ✓ **Mantener correlaciones entre variables**: Esta propiedad busca conservar las posibles relaciones lineales o no lineales que se puedan presentar entre las variables. El equipo de trabajo definirá si mantiene los coeficientes de correlación entre dos o más variables (cuantitativas o categóricas) en la base de datos anonimizada, con el fin de no distorsionar los resultados finales. Se recomienda que el equipo de trabajo garantice que las correlaciones entre variables se mantienen para que los datos anonimizado no conlleven a interpretaciones erróneas por parte de los usuarios de la información.

Finalmente, en este subproceso el equipo de trabajo definirá qué propiedades estadísticas deberá tener la base de datos anonimizada, dado que son insumo para evaluar el proceso de anonimización. Además, describirá las propiedades estadísticas a conservar en la base de datos, teniendo en cuenta:

- ✓ Descripción de las propiedades estadísticas elegidas por el grupo de trabajo para conservar en la base de datos anonimizada.
- ✓ Listado de las variables con la respectiva propiedad estadística a conservar en la base de datos anonimizada.

- ✓ Nivel de desagregación geográfica o temática en el que se desea conservar las propiedades estadísticas.
- ✓ Porcentajes de variación permitidos por variables y niveles de desagregación geográfica o temática para las propiedades globales en la base de datos anonimizada.

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

Para visualizar estos elementos de los subprocesos de la Etapa I, se presenta a continuación un ejemplo.

Ejemplo de la Etapa I. Análisis exploratorio de la base de datos

La base de datos anonimizada de la Encuesta Anual de Comercio (EAC) del año 2016[7], cuenta con 64 variables y 10.242 unidades de observación, en este caso empresas. Las variables se pueden caracterizar de la siguiente forma:

Tabla 4. Clasificación por tipo de variable de la EAC en el 2016

VARIABLES/TIPO DE VARIABLE	CUANTITATIVAS	CATEGÓRICAS	TOTAL
IDENTIFICACIÓN	0	2	2
UBICACIÓN	0	0	0
TEMÁTICAS	59	3	62
TOTAL	**59**	**5**	**64**

Fuente: DANE- EAC, Cálculos DANE - DIRPEN

Además, las medidas descriptivas para cinco variables cuantitativas temáticas de la operación estadística, son:

[7] Disponible en el Archivo Nacional de Datos

Tabla 5. Medidas descriptivas de algunas variables cuantitativas de la EAC en el 2016

VARIABLE	MEDIA	VARIANZA	CUARTIL 1	CUARTIL 2	CUARTIL 3
TOTAL SUELDOS*	973.902	2.76E+13	137.877	274.157	629.193
TOTAL PRESTACIONES*	491.284	9.30E+12	58.905	118.604	280.252
VALOR AGREGADO*	3.220.106	2.40E+14	360.691	834.109	2.058.826
VENTAS CAUSADAS*	24.022.327	1.89E+16	2.789.874	6.335.314	14.654.517
PERSONAL REMUNERADO**	52.5	114575.7999	12	19	38

*Cifras en millones de pesos
**Cifra en número de personas

Fuente: DANE- EAC, Cálculos DANE - DIRPEN

Además, se presenta la distribución de frecuencias de la variable categórica Organización Jurídica:

Tabla 6. Distribución de Frecuencias de la variable Organización Jurídica de la EAC

ORGANIZACIÓN JURÍDICA	NUMERO DE REGISTROS POR CATEGORÍA	%DE REGISTROS POR CATEGORÍA
Sociedad en comandita simple	111	1.08%
Sociedad en comandita por acciones	37	0.36%
Sociedad limitada	1,676	16.36%
Sociedad anónima	1,567	15.30%
Sucursal de sociedad extranjera	30	0.29%
ORGANIZACIÓN JURÍDICA	**NUMERO DE REGISTROS POR CATEGORÍA**	**% DE REGISTROS POR CATEGORÍA**
Empresa unipersonal	63	0.62%
Persona natural	1,946	19.00%
Organizaciones de economía solidaria	82	0.80%
Entidades sin ánimo de lucro	50	0.49%
Sociedad por acciones simplificada	4,652	45.42%
Otro	28	0.27%
Total	**10,242**	**100.00%**

Fuente: DANE- EAC, Cálculos DANE - DIRPEN

Finalmente, la revisión temática del contenido de la base de datos permite concluir que:

- El objetivo de la Encuesta Anual de Comercio es conocer la estructura y el comportamiento económico del sector comercio a nivel nacional y por grupo de actividad comercial, de manera que permita el análisis de la evolución del sector y de la conformación de agregados económicos.

- En la página web del DANE[8] se encuentran disponibles todos los metadatos y microdatos de la EAC. Los documentos contienen información sobre:

  - ✓ Recolección de los datos
  - ✓ Procesamiento de la información
  - ✓ Políticas de acceso a los microdatos
  - ✓ Diccionario de datos
  - ✓ Descripción de cada una de las variables (incluidos los estándares utilizados)
  - ✓ Referentes internacionales
  - ✓ Cuestionario
  - ✓ Metodología y ficha metodológica.

Revisión de restricciones de publicación de la información

La Encuesta Anual de Comercio, cuenta con los siguientes fundamentos normativos:

✓ Ley 2ª de 1962, que permite el levantamiento de encuestas nacionales y en especial, las de industria, comercio y servicios.

✓ El Decreto 1633 de 1960 en su artículo 74 establece que todas las personas naturales o jurídicas, domiciliadas en el territorio nacional y los empleados públicos, en todos sus niveles, están obligados a suministrar información al DANE, dentro de los plazos establecidos para el efecto que se señalen, para el cumplimiento de sus finalidades, además, establece que los datos suministrados a la entidad tienen un carácter estrictamente reservado y, por lo tanto, no podrán darse a conocer al público ni a las entidades oficiales, sino únicamente en resúmenes numéricos.

✓ La Ley 0079 de octubre 20 de 1993 regula la realización de los censos y encuestas decreta que las personas naturales o jurídicas, de cualquier orden o naturaleza, domiciliadas o residentes en el territorio nacional, están obligadas a suministrar información al (DANE), así mismo especifica que la entidad podrá imponer multas a quienes incumplan esta disposición.

[8] https://www.dane.gov.co/index.php/estadisticas-por-tema/comercio-interno/encuesta-anual-de-comercio-eac

- ✓ El principio de confidencialidad descrito en la ley estatutaria 1266 de 2008 indica "Todas las personas naturales o jurídicas que intervengan en la administración de datos personales que no tengan la naturaleza de públicos están obligadas en todo tiempo a garantizar la reserva de la información"

- ✓ El artículo 2.2.3.3.5 del decreto 1743 de 2016 indica "guardar la confidencialidad de los datos que permitan la identificación y/o localización espacial de las fuentes, cuando estos fueren recolectados exclusivamente para la producción de las estadísticas oficiales y para fines estadísticos".

- ✓ Ley 79 de 1993 Artículo 5: Los datos suministrados al Departamento Administrativo Nacional de Estadística DANE, en desarrollo de censos y encuestas, no podrán darse a conocer al público ni a entidades u organismos oficiales, ni a las autoridades públicas, sino únicamente en resúmenes numéricos, que no hagan posible deducir de ellos información alguna de carácter individual que pudiera utilizarse para fines comerciales, de tributación fiscal, de investigación judicial o cualquier otro diferente del propiamente estadístico.

Acorde con la revisión de esta normatividad, no se evidencia alguna norma que nos impida publicar la información siempre y cuando se proteja la identificación de las unidades de observación.

Revisión de las necesidades de los usuarios de la información

El grupo de trabajo, después de revisar las solicitudes de información en la entidad durante los últimos dos años, evidencia que:

- ✓ La mayoría de las solicitudes de información que llegan al DANE provienen de Gremios, Asociaciones, Investigadores, Académicos, Centros de Investigación, Universidades.

- ✓ Las variables Ingreso por ventas, Gastos de personal (sueldos y prestaciones), Costos de la mercancía vendida, Gastos de operación, Personal ocupado, Inventarios, Movimientos de activos fijos, son las que mayor demanda de información presentan.

- ✓ La información correspondiente al año 2016 se encuentra de forma recurrente entre los periodos de tiempo solicitado, a través de las diferentes solicitudes de información que ha recibido la entidad.

- ✓ La información que los diferentes usuarios manifiestan en sus respectivas solicitudes corresponden a la información desagregada a nivel nacional.

Definición de las propiedades estadísticas a conservar en la base de datos

En la Encuesta Anual de Comercio – EAC, el equipo de trabajo decide que las propiedades estadísticas que la base de datos anonimizada debe conservar son:

- ✓ Para las variables: ventas y el personal ocupado **mantener la participación % directo de las divisiones** CIIU Rev. 4. A.C. del sector comercio.

- ✓ Para las variables: Número de empresas, Ventas, Costo de la Mercancía, Producción Bruta, Consumo Intermedio, Valor agregado, Remuneración, **mantener los totales a nivel nacional.**

- ✓ La información correspondiente al subsector vehículos automotores, motocicletas, sus partes, piezas y accesorios debe **mantener los totales a nivel nacional.**

- ✓ Con la base de datos anonimizada los usuarios puedan replicar los diferentes cuadros de salida que son publicados en el boletín técnico[9] de la encuesta anual de comercio para el año 2016.

4.3. Etapa II: Análisis de riesgos de identificación de las unidades de observación:

En esta etapa el equipo de trabajo se planteará todos los posibles escenarios de riesgo de identificación de las unidades de observación de la base de datos.

Un escenario de riesgo de identificación de información confidencial es aquel en el cual existe una posibilidad de que, mediante la combinación de variables de la base de datos, se puedan identificar características de las unidades de observación que deben ser protegidas por el tipo de información que contienen.

La etapa de análisis de riesgos se compone de cuatro subprocesos así:

- ✓ Clasificación de variables por su nivel de sensibilidad
- ✓ Planteamiento de riesgos de la base de datos
- ✓ Identificación de unidades de observación riesgosas
- ✓ Creación del informe de riesgos

[9] https://www.dane.gov.co/files/investigaciones/boletines/eac/bol_eac_2016.pdf

4.3.1. *Clasificación de variables por su nivel de sensibilidad*

La etapa de análisis de riesgos inicia con la clasificación de las variables de la base de datos por su nivel de sensibilidad. El equipo debe realizar la revisión de los riesgos teniendo en cuenta el contexto de la base de datos y la(s) persona(s) encargada(s) del procesamiento de la base de datos.

Tenga en cuenta que una **variable se considera sensible** si:

> "… contiene información privada de la unidad de observación de la base de datos. Inicialmente, las variables sensibles son todas aquellas que permiten la identificación y ubicación de las fuentes de información. Sin embargo, otro tipo de variables sensibles, son las variables con contenido temático (social, económico, entre otros) que combinadas entre sí permiten la identificación de las fuentes de información" (Concepto propio DANE, 2018)

En este punto los expertos temáticos que componen el equipo de trabajo juegan un rol preponderante.

Las variables pueden clasificarse en:

- ✓ **Identificadores directos:** Las variables denominadas como identificadores directos son todas aquellas variables que contienen información sensible de identificación o ubicación de las unidades de observación. Estas variables pueden coincidir con las variables denominadas en el análisis exploratorio de la base de datos (Etapa I) como variables de identificación o de ubicación. Un ejemplo de este tipo de variables, es el número de cédula de una persona, NIT de una empresa, dirección de una entidad, entre otros.

- ✓ **Pseudoidentificadores:** Las variables denominadas como pseudoidentificadores son todas aquellas que combinadas con otras variables, conllevan a la identificación de las unidades de observación. Estas variables pueden coincidir comúnmente, con las variables denominadas temáticas o de ubicación en el análisis exploratorio de la base de datos (Etapa I). Un ejemplo de este tipo de variables, es la combinación del nivel de escolaridad con el ingreso promedio mensual en cierto municipio. En este caso, son 3 variables consideradas como pseudoidentificadores, que permiten la identificación de *algunas* unidades de observación.

- ✓ **No confidenciales:** Las variables denominadas como no confidenciales son todas aquellas que no permiten la identificación de las unidades de observación de la base de datos, ni siquiera cuando son combinadas con pseudoidentificadores. Algunos

ejemplos de este tipo de variables de acuerdo con el contexto de la base de datos podrían ser: habilidades de conducción de un vehículo cuando el contexto de la base está relacionada con el estudio de una enfermedad particular; gusto por el deporte o gustos culturales, en una base de datos relacionada con el mercado laboral.

El equipo de trabajo puede resumir la clasificación de acuerdo con la siguiente tabla:

Tabla 7. Tabla resumen de la clasificación de las variables por su tipo de sensibilidad

TIPO DE VARIABLE/TIPO DE SENSIBILIDAD	IDENTIFICADORES DIRECTOS	PSEUDO IDENTIFICADORES	NO CONFIDENCIALES	TOTAL
CUANTITATIVAS	*Escriba en este espacio el número de variables cuantitativas clasificadas como identificadores directos*		*Escriba en este espacio el número total de variables cuantitativas clasificadas como no confidenciales*	
CATEGÓRICAS		*Escriba en este espacio el número total de variables categóricas clasificadas como pseudoidentificadores*		
TOTAL				*Escriba en este espacio el número total de variables*

Fuente: DANE- DIRPEN

4.3.2. Planteamiento de riesgos de la base de datos

Después de la clasificación de todas las variables de la base de datos de acuerdo con su nivel de sensibilidad, el equipo de trabajo deberá establecer los riesgos de identificación de la base de datos que será anonimizada. Los riesgos que defina el equipo de trabajo servirán como insumo en la selección de técnicas de anonimización (Sección 4.4.2).

En general, los riesgos se pueden entender como todas las posibles combinaciones de las variables (entre identificadores directos y pseudoidentificadores) 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.

Tenga en cuenta que al realizar combinaciones entre variables, se pueden identificar unidades de observación que deben ser protegidas dentro de sus niveles de desagregación geográfica o temática.

Por ejemplo:

- ✓ Si en la base de datos de un hospital de la ciudad de Cartagena, se tiene la información de los pacientes atendidos en el año 2016, al combinar la información de la edad, tipo de enfermedad, medicamentos recibidos y dirección, es posible identificar a las mujeres de más de 70 años que han tenido cáncer de seno para el barrio la Chinita. Esta combinación de las variables permite reconocer a la unidad de observación.
- ✓ Al tener la información de los ingresos por hogar de los colombianos, al combinar las variables, Ingresos anuales, edad, sexo, corregimiento o vereda, nivel escolar, podemos identificar que hay un hombre de 75 años con nivel educativo profesional en el corregimiento de Puerto Salazar, con ingresos anuales de $ 35.601.804. Con esta combinación podría identificarse que se trata del único Médico del corregimiento.

- ✓ Al tener la información de ventas por catálogo de la empresa Yanbal, al combinar las variables departamento, sexo, ventas reportadas, artículos, se puede identificar que en Yopal hay un hombre y dos mujeres que poseen las ventas más altas de la región, por lo que se puede inferir que se traten de los líderes Premium del departamento.

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. Algunas recomendaciones sobre cómo definir los niveles de riesgos de identificación son:

- ✓ Verificar los identificadores que definitivamente deben ser suprimidos de la base de datos por su nivel de sensibilidad.

- ✓ Verificar los identificadores directos que podrían ser agrupados con el propósito de minimizar el riesgo de identificación

- ✓ Listar todas las posibles combinaciones de las variables que representen un riesgo de identificación de las unidades de observación.

- ✓ Analizar los niveles mínimos de desagregación de los identificadores directos o pseudoidentificadores que puedan ser más riesgosos para determinar la identificación de las unidades de observación. Usualmente estos niveles de desagregación hacen referencia a desagregaciones geográficas o temáticas.

- ✓ Revisar las medidas descriptivas estadísticas calculadas para las variables cuantitativas en el ***Análisis exploratorio de la base de datos (Etapa I).***

- ✓ Revisar la distribución de frecuencias de las variables categóricas calculadas en el ***Análisis exploratorio de la base de datos*** (Etapa I) y utilizarlas en la identificación de categorías con frecuencias considerablemente bajas. Usualmente este tipo de categorías se asocian como riesgosas porque permiten fácilmente identificar a las unidades de observación.

- ✓ Considerar la recodificación de las categorías de las variables que presentan frecuencias considerablemente bajas. Esto dependerá del contexto de la base de datos que se anonimizará.

- ✓ Verificar las bases de datos de entidades externas encontradas en la revisión temática realizada en el ***Análisis exploratorio de la base de datos (subproceso de la Etapa I*)**. El equipo de trabajo deberá identificar si todas las variables contenidas en esas bases podrían convertirse en pseudoidentificadores respecto a la base de datos original sujeta a anonimizar. Esta revisión permitirá establecer qué variables de las bases de datos externas, combinadas con las variables de la base de datos a anonimizar, aumentan el riesgo de identificación de las unidades de observación.

Cuando el equipo de trabajo obtenga el listado priorizado de los riesgos de identificación de las unidades de observación de acuerdo a la frecuencia en la que ocurran estos riesgos; procederá a evaluar temáticamente la base de datos con el equipo de trabajo.

En este caso, el equipo revisará si es necesario considerar todos los riesgos identificados, si se pueden considerar sólo algunos, o si se pueden construir nuevos riesgos a partir de la primera ronda de riesgos planteados.

Después de que el equipo de trabajo haya evaluado todos los posibles riesgos de identificación y haya decidido finalmente cuáles son los definitivos (o más probables), procederá a identificar qué unidades de observación se considerarán riesgosas bajo los escenarios de riesgos definidos en este subproceso.

4.3.3. *Identificación de unidades de observación riesgosas*

Con el listado definitivo de riesgos de identificación priorizados, el equipo de trabajo procederá a identificar con mayor nivel de detalle las unidades de observación que son riesgosas bajo todos los riesgos planteados en el subproceso anterior.

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.**

La identificación de las unidades riesgosas, por su parte, puede ser una actividad a realizar por parte del equipo del procesamiento de la base de datos dada su experticia y capacidades técnicas en el manejo de este tipo de archivos de información.

A continuación, la
Tabla ***8*** presenta un resumen sobre la identificación de las unidades de observación que resultan riesgosas.

Tabla 8. Resumen de unidades de observación riesgosas en la anonimización teniendo en cuenta **3 riesgos**

RIESGO	VARIABLES INVOLUCRADAS	¿CUÁNDO SE CONSIDERA UNA UNIDAD DE OBSERVACIÓN RIESGOSA?	NÚMERO DE UNIDADES DE OBSERVACIÓN RIESGOSAS	PORCENTAJE DE UNIDADES DE OBSERVACIÓN RIESGOSAS
RIESGO 1	*Escriba en este espacio qué variables están involucradas en este riesgo*			*Escriba en este espacio el porcentaje de unidades de observación riesgosas por el riesgo 1con respecto al total de unidades de observación*
RIESGO 2		*En este espacio explique qué condiciones debe cumplir una unidad de observación para considerarse riesgosa*		
RIESGO 3			*Escriba en este espacio cuántas unidades de observación se consideran riesgosas bajo el riesgo 3.*	
TOTAL	*Escriba en este espacio el número de variables involucradas en el análisis de riesgos*			*Escriba en este espacio el porcentaje de unidades de observación riesgosas con respecto al total de unidades de observación*

Fuente: DANE- DIRPEN

4.3.4. Creación del informe de riesgos

Finalmente, el equipo de trabajo creará un ***Informe de Riesgos*** que será utilizado en la identificación y aplicación de técnicas de anonimización (Etapa III), el cual describirá cómo se clasifican las variables según su 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.

Se recomienda que el informe contenga al menos la siguiente información:

- Criterios y aspectos considerados en la definición de los riesgos
- Listado de riesgos definitivos priorizados.
- La tabla resumen de unidades de observación riesgosas obtenida en el tercer subproceso de esta etapa (
- Tabla ***8***).
- Fecha de emisión del informe

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

Para visualizar estos elementos de los subprocesos de la Etapa I, se presenta a continuación un ejemplo.

Ejemplo de la Etapa II: Análisis de riesgos.

Para ejemplificar la etapa de análisis de riesgo, se utilizará una base de datos simulada a la cual se llamará **COL20**[10], la cual contiene información de 32 variables de identificación, ubicación y socioeconómicas, medidas en 496 personas que se encuentran presentes en 342 municipios a nivel nacional. La descripción de cada una de las variables de **COL20** se encuentra disponible en el Anexo A.

[10] Base de datos disponible en el Anexo B.

Para el análisis de riesgos, con base en la experiencia del DANE, se iniciará con la clasificación de las variables por su tipo de sensibilidad. Las 32 variables de **COL20**, pueden caracterizarse así:

Tabla 9. Clasificación de las variables por tipo de sensibilidad de COL20

TIPO DE VARIABLE/TIPO DE SENSIBILIDAD	IDENTIFICADORES DIRECTOS	PSEUDOIDENTIFICADORES	INFORMATIVAS	TOTAL
CUANTITATIVAS	0	8	0	8
CATEGÓRICAS	6	14	3	23
TOTAL	**6**	**22**	**3**	**31**

Fuente: DANE- DIRPEN

De la anterior tabla, se puede concluir que 28 variables son sensibles, entre identificadores directos y pseudoidentificadores, donde 20 de las variables son categóricas. Algunas variables contenidas en **COL20** que son identificadores directos son nombres, apellidos, dirección, número de identificación, y respecto a los pseudoidentificadores, se tienen variables como RH, grupo étnico, ingresos anuales, número de bienes raíces, entre otras.

Al revisar la distribución de las variables, se observa que la mayor parte de estas son categóricas, recordando que para este tipo de variables, las unidades de observación cuentan una valoración cuando cumplen una característica particular. Es el caso de la variable Grupo étnico que toma valores *afrocolombiano, gitano, indígena o ninguno*. Teniendo en cuenta estas características de las variables categóricas, es posible utilizar como una medida aproximada para la definición de riesgos de identificación, las distribuciones de frecuencias.

Posteriormente, siguiendo las recomendaciones propuestas en esta guía, se elaboró el siguiente listado de riesgos de la base de datos **COL20**:

- **Los identificadores directos como** número de identificación, tipo de identificación, nombre, apellidos, fecha de nacimiento, barrio y dirección **deben ser suprimidos** definitivamente de la base de datos porque representa una identificación inmediata de las unidades de observación.

- Las siguientes combinaciones entre pseudoidentificadores se consideran riesgosas porque el cruce entre ellas podría, eventualmente, permitir la identificación de una unidad de observación de la base de datos. Sin embargo, es posible que se identifiquen otra serie de combinaciones que pueden ser identificadas como riesgos:

  - ✓ Ingresos anuales (o mensuales) y nivel de escolaridad a nivel municipal y departamental.
  - ✓ RH y edad a nivel municipal y departamental.
  - ✓ Ocupación, nivel de escolaridad, ingresos anuales (o mensuales) a nivel municipal.
  - ✓ Grupo étnico a nivel municipal y departamental.
  - ✓ Número de habitaciones de la vivienda, materiales de los pisos del hogar e ingresos anuales (o mensuales) a nivel municipal y departamental.
  - ✓ Número de bienes raíces con la variable “tiene vehículo” (marca y modelo) a nivel municipal y departamental.
  - ✓ Número de viajes fuera del país, ocupación e ingresos anuales (o mensuales).
  - ✓ Valores frecuentes de la variable asistencia a eventos culturales (o deportivos), edad e ingresos a nivel municipal y departamental.

- La desagregación a nivel municipal es altamente riesgosa para la identificación de las unidades de observación con respecto a algunas variables temáticas como es el caso de la variable *Ingresos anuales*, porque se podría tener con certeza que en el municipio El Castillo en el departamento del Meta se encuentran 2 personas con ingresos anuales superiores a 10 millones de pesos.

- A partir de las medidas descriptivas estadísticas calculadas en la Etapa I, se identificaron las variables cuantitativas más sensibles. A continuación, se presentan los resultados de este ejercicio para **COL20**:

Tabla 10. Medidas descriptivas de las variables cuantitativas de **COL20**

VARIABLE	MEDIA	VARIANZA	CUARTIL 1	CUARTIL 2	CUARTIL 3
Ingresos Anuales*	$ 52,174,008	2.15679E+15	$ 13,188,150	$ 40,048,606	$ 89,000,899
Ingresos Mensuales*	$ 4,347,834	1.49777E+13	$ 1,099,012	$ 3,337,384	$ 7,416,742
Número de hijos nacidos vivos**	1.97	2.84	0	2	3
Número de bienes raíces**	2.38	3.25	1	2	4
Número de viajes fuera del país**	3.47	6.74	1	3	6

*Unidades: millones de pesos
**Números enteros

Fuente: DANE- DIRPEN

- Con el propósito de identificar las variables categóricas más sensibles, se buscan aquellas categorías con menor participación a nivel nacional. En este ejemplo, se utilizaron las distribuciones de frecuencia calculadas en la sección 4.2.1 (Tabla 3), de las variables categóricas que el equipo consideró eran las más sensibles, como por ejemplo: grupo étnico, nivel de escolaridad y RH.

  Es importante recordar que las categorías de las variables cualitativas con frecuencias considerablemente bajas con respecto al total presentan altos niveles de riesgo. Por ejemplo, se identificaron las variables de RH; nivel de escolaridad y grupo étnico como variables que permiten fácilmente la identificación de las unidades de observación.

- Estas categorías deben revisarse cuidadosamente, ya que las unidades de observación que pertenezcan a esta categoría pueden ser riesgosas a nivel municipal.

A continuación, se presentan la frecuencias para las variables que tienen un mayor nivel de sensibilidad en la base de datos **COL20**:

Tabla 11. Distribución de frecuencias del RH en **COL20**

RH	NÚMERO DE REGISTROS	PORCENTAJE DE REGISTROS
O+	244	49.2%
A+	212	42.7%
B+	5	1.0%
AB+	3	0.6%
O-	23	4.6%
A-	6	1.2%
B-	1	0.2%
AB-	2	0.4%
Total	496	100%

Fuente: DANE- DIRPEN

En este caso, las categorías de RH **B-; B+; AB-; AB+ y A-,** se consideran como las más sensibles, dada su baja frecuencia de aparición en las unidades de observación

Tabla 12. Distribución de frecuencias del nivel de escolaridad en **COL20**

ESCOLARIDAD	NÚMERO DE REGISTROS	PORCENTAJE DE REGISTROS
Primaria	54	10.9%
Secundaria	26	5.2%
Básica Media	70	14.1%
Técnico	30	6.0%
Tecnólogo	88	17.7%
Profesional	130	26.2%
Posgrado	98	19.8%
Total	496	100%

Fuente: DANE- DIRPEN

Para este segundo caso, se tendrían las categorías de **secundaria y técnico** como los niveles de escolaridad más sensibles dada su baja frecuencia en las unidades de observación

Tabla 13. Distribución de frecuencias del grupo étnico en **COL20**

GRUPO ÉTNICO	NÚMERO DE REGISTROS	PORCENTAJE DE REGISTROS
Afrocolombiano	20	4.0%
Indígena	12	2.4%
Rrom	4	0.8%
Ninguno	460	92.7%
Total	496	100%

Fuente: DANE- DIRPEN

En este último caso, las categorías **Rrom, indígena y afrocolombiano** son los grupos étnicos más sensibles dada su baja frecuencia en las unidades de observación.

Además, se presenta la tabla de identificación de unidades de observación riesgosas, la cual sirve como insumo para la identificación de las técnicas de anonimización a utilizar y definir la viabilidad del proceso de anonimización:

Tabla 14**. Unidades de observación riesgosas para los cinco riesgos más frecuentes para la anonimización de **COL20

RIESGO	VARIABLES INVOLUCRADAS	¿CUÁNDO SE CONSIDERA UNA UNIDAD DE OBSERVACIÓN RIESGOSA?	NÚMERO DE UNIDADES DE OBSERVACIÓN RIESGOSAS	PORCENTAJE DE UNIDADES DE OBSERVACIÓN RIESGOSAS
1	Ingresos mensuales y departamento	Las 3 personas con el ingreso anual más alto en cada departamento	99	20%
2	Grupo étnico, departamento	Las personas pertenecientes a un grupo étnico particular a nivel departamental	36	7.3%
3	Número de habitaciones de la vivienda, departamento	Todas las personas que vivan en una vivienda con un número de habitaciones por encima del promedio departamental	235	47.4%
4	Número de viajes fuera del país y departamento	Todas las personas que hayan viajado fuera del país más veces que el promedio departamental	225	45.4%
5	Nivel de escolaridad y departamento	Todas las personas con posgrado en aquellos departamentos con menos de 4 personas a ese nivel de escolaridad	31	6.3%

Fuente: DANE- DIRPEN

La anterior tabla evidencia que de los cinco riesgos más probables seleccionados, más del 40% de las unidades de observación de la base de datos tienen riesgo de ser identificadas por las variables *número de habitaciones de la vivienda (47.4%)* y *número de viajes fuera del país (45.4%)* (Riesgos 3 y 4). De la misma forma, 99 unidades de observación tienen riesgo de ser identificadas por sus ingresos mensuales (Riesgo 1).

Teniendo en cuenta que estas tres variables son cuantitativas, es posible identificar las técnicas de anonimización más idóneas para minimizar el riesgo de identificación de aquellas unidades de observación.

4.4. Etapa III: Identificación y selección de técnicas de anonimización

En esta etapa el equipo de trabajo conocerá e identificará as técnicas de anonimización más comunes para variables cuantitativas y categóricas. Además, seleccionará una o más técnicas para aplicar a cada uno de los riesgos planteados en la etapa anterior.

La etapa de identificación y selección de técnicas de anonimización se compone de dos subprocesos así:

- ✓ Identificación de técnicas de anonimización más comunes
- ✓ Selección de técnicas de anonimización

4.4.1. *Identificación de técnicas de anonimización más comunes*

En este subproceso se presentan de **manera general** las técnicas de anonimización más comunes por tipo de variable, que permiten minimizar el riesgo de identificación de las unidades de observación.

Tenga en cuenta que las técnicas de anonimización se dividen en:

- Técnicas basadas en la no perturbación de datos
- Técnicas basadas en la perturbación de datos

Técnicas basadas en la no perturbación de datos: Estas técnicas utilizan supresiones parciales, reducción o recodificación de la información para minimizar el riesgo de identificación de las unidades de observación. Este tipo de técnicas son comúnmente utilizadas para evitar que los datos atípicos sean de fácil identificación.

Técnicas basadas en la perturbación de datos: Estas técnicas se refieren a procedimientos que implican la modificación sistemática de datos (a veces en pequeñas cantidades aleatorias), de manera tal 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.[11]

[11] González M. 2017, p. 29

A continuación, se describen brevemente las técnicas y su implementación teniendo en cuenta el tipo de variables que pueden estar contenidas en las bases de datos. En este caso para aquellas técnicas basadas en no perturbación.

Tabla 15. Técnicas basadas en la no perturbación de datos según el tipo de variable.

TÉCNICAS	DESCRIPCIÓN	TIPO DE VARIABLE	EJEMPLO VARIABLES (BASE COL20)	REFERENCIA BIBLIOGRÁFICA
ELIMINACION DE VARIABLES	Esta técnica suprime toda la información de una variable. Se usa cuando la variable contiene información de identificación directa de la unidad de observación.	En variables categóricas	CEDULA;TIPO DE IDENTIFICACION; NOMBRE; APELLIDOS;DIRECCIÓN; BARRIO; MUNICIPIO;FECHA DE NACIMIENTO;RH Estas variables son eliminadas porque debido a la información contenida, es posible identificar directamente a las unidades de observación. Algunas de ellas contienen información sensible, por lo tanto, permitirían reconocer las unidades de observación.	Hundepool et al. (2010)
RECODIFICACIÓN GLOBAL	Esta técnica consiste en combinar diversas categorías de las variables categóricas en una más general que tenga mayor frecuencia y menor información. En el caso de las variables continuas, consiste en agrupar 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 pseudoidentificadoras.	En variables cuantitativas o categóricas	GRUPO ÉTNICO;NUMERO DE HABITACIONES DE LA CASA;NUMERO DE VIAJES REALIZADOS FUERA DEL PAÍS; NIVEL DE ESCOLARIDAD Estas variables son recodificadas para combinar las categorías de las variables y poder tener en cada nueva categoría una mayor frecuencia de unidades de observación.	Hundepool et al., (2012) Templ et al., IHSN Working Paper No. 007 (2014).

TÉCNICAS	DESCRIPCIÓN	TIPO DE VARIABLE	EJEMPLO VARIABLES (BASE COL20)	REFERENCIA BIBLIOGRÁFICA
CODIFICACIÓN SUPERIOR E INFERIOR	Esta técnica consiste en proteger la identificación de las unidades de observación que presentan los valores más altos o más 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.	En variables continuas o categóricas	TÉCNICA NO UTILIZADA EN LA ANONIMIZACIÓN DE LA BASE DE DATOS EJEMPLO COL20	Hundepool et al. (2012)
SUPRESION LOCAL	Esta técnica consiste en 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 pseudoidentificadoras permita la identificación de las unidades de observación.	En variables categóricas	TÉCNICA NO UTILIZADA EN LA ANONIMIZACIÓN DE LA BASE DE DATOS EJEMPLO COL20	Hundepool y De Wolf (2012) Templ et al., IHSN Working Paper No. 007 (2014)

Fuente: DANE- DIRPEN

Respecto a las técnicas basadas en perturbación, se tienen las siguientes recomendaciones:

Tabla 16. Técnicas basadas en la perturbación de datos según el tipo de variable

TÉCNICAS	DESCRIPCIÓN	TIPO DE VARIABLE	EJEMPLO VARIABLES (BASE COL20)	REFERENCIA BIBLIOGRÁFICA
MICRO AGREGACIÓN	Esta técnica consiste en reemplazar los valores de algunas unidades de observación, por el valor promedio calculado sobre ellas. Comúnmente, se usa cuando la unidad de observación por nivel de desagregación geográfica es de fácil identificación	En variables cuantitativas	EDAD;INGRESOS ANUALES;INGRESOS MENSUALES;NÚMERO DE HIJOS; NACIDOS VIVOS;NÚMERO DE PERSONAS QUE COMPONEN EL HOGAR;NUMERO DE HABITACIONES DE LA CASA;NUMERO DE BIENES RAICES;NUMERO DE VIAJES FUERA DEL PAÍS Estas variables son microagregadas porque se busca proteger la identificación de las unidades de observaciones con los ingresos anuales más altos por departamento.	Hundepool et al., (2012) Templ et al., IHSN Working Paper No. 007 (2014)
REDONDEO	Esta técnica consiste en sustituir los valores de las unidades de observación en aquellas variables que tienen decimales 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. Ej. Número de hijos.	En variables cuantitativas	EDAD;NÚMERO DE HIJOS NACIDOS VIVOS;NUMERO DE HABITACIONES DE LA CASA;NÚMERO DE BIENES RAÍCES;NUMERO DE VIAJES FUERA DEL PAÍS. Estas variables son redondeadas para mantener la información de las variables en unidades enteras después de su microagregación.	Hundepool et al. (2012)
INTERCAMBIO DE DATOS	Esta técnica consiste en intercambiar 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).	En variables cuantitativas o categóricas	TÉCNICA NO UTILIZADA EN LA ANONIMIZACIÓN DE LA BASE DE DATOS EJEMPLO COL20	*Hundepool et al., 2010, p. 58*

TÉCNICAS	DESCRIPCIÓN	TIPO DE VARIABLE	EJEMPLO VARIABLES (BASE COL20)	REFERENCIA BIBLIOGRÁFICA
AGREGAR RUIDO	Esta técnica consiste en añadir una cantidad aleatoria definida por el equipo de trabajo sobre los valores de las unidades de observación (ruido aleatorio). Comúnmente, es utilizada 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 variables cuantitativas	TÉCNICA NO UTILIZADA EN LA ANONIMIZACIÓN DE LA BASE DE DATOS EJEMPLO COL20	*Hundepool A. 2012, p.54* *Templ et al., IHSN Working Paper No. 007 (2014), p.9*

Fuente: DANE- DIRPEN

Otra de las técnicas que ha ganado un espacio en la literatura y en los procesos de anonimización de bases de datos, especialmente en Estados Unidos, es el uso de **datos sintéticos**.

Está técnica consiste en el uso de datos simulados, siendo una alternativa a los métodos previamente explicados. En este caso, se produce una nueva base de datos mediante el uso de algoritmos de simulación, que conserve las propiedades estadísticas de la base de datos no anonimizada. Para la generación de los datos simulados se puede hacer uso de métodos como regresión cuantílica, imputación adicional y datos combinados (Hundepool, et al., 2010: 58)

Comúnmente, cuando las propiedades estadísticas sobre la base de datos sin anonimizar incluyen el no perturbar la información, o hacerlo lo menos posible, los métodos de datos sintéticos no son los más adecuados, ya que proporcionan las mismas tendencias, propiedades globales o correlaciones; sin embargo, modifican todos los campos a nivel de microdato. Cuando se modifican todos los campos, el nivel de utilidad de la información puede disminuir considerablemente.

Algunos ejemplos y desarrollos teóricos de este tipo de técnicas se pueden consultar en la siguiente bibliografía:

- Domingo-Ferrer J., Drechsler J. and Polettini S (2009) Report on synthetic data files. Technical report, Deliverable of Project ESSNET-SDC

- Drechsler J., Bender S. and R¨assler S. (2008a) “Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel” *Transactions on Data Privacy* **1**(3): 105–130.

- Fienberg S.E. (1994) “A radical proposal for the provision of micro-data samples and the preservation of confidentiality” *Technical Report 611*, Carnegie Mellon University Department of Statistics.

- Fienberg S.E. and Makov U.E. (1998) “Confidentiality, uniqueness and disclosure limitation for categorical data“ *Journal of Official Statistics* 14(4): 385–397.

- Liew C.K., Choi U.J. and Liew C.J. (1985) “A data distortion by probability distribution”. *ACM Transactions on Database Systems* 10: 395–411.

- Reiter J.P. (2005a) “Releasing multiply-imputed, synthetic public use microdata: an illustration and empirical study“ *Journal of the Royal Statistical Society, Series A* **168**:185–205.

- Woodcock S.D. and Benedetto G. 2007 *Distribution-preserving statistical disclosure limitation.* Disponible en SSRN: http://ssrn.com/abstract=931535.

4.4.2. *Selección de técnicas de anonimización con base en los riesgos identificados*

Cuando el equipo de trabajo identifique las técnicas de anonimización más comunes por tipo de variable, seleccionará una o más técnicas que permitan minimizar la ocurrencia de cada uno de los riesgos identificados en la Etapa II.

Con base en la tabla Clasificación de las variables por tipo de sensibilidad de la base de datos simulada construida en la Etapa II (Tabla 7), el equipo agregará la columna: Técnica(s) de anonimización a utilizar.

Tabla 17. Planteamiento de técnicas de anonimización para cada uno de los riesgos identificados

RIESGO	VARIABLES INVOLUCRADAS	¿CUÁNDO SE CONSIDERA UNA UNIDAD DE OBSERVACIÓN RIESGOSA?	NÚMERO DE UNIDADES DE OBSERVACIÓN RIESGOSAS	PORCENTAJE DE UNIDADES DE OBSERVACIÓN RIESGOSAS	TÉCNICA(S) DE ANONIMIZACIÓN A UTILIZAR
RIESGO 1	*Escriba en este espacio qué variables están involucradas en este riesgo*			*Escriba en este espacio el porcentaje de unidades de observación riesgosas por el riesgo 1 con respecto al total de unidades de observación*	*Escriba en este espacio la(s) técnica(s) de anonimización que minimizarán la ocurrencia del riesgo 1.*
RIESGO 2		*En este espacio explique qué condiciones debe cumplir una unidad de observación para considerarse riesgosa*			
RIESGO 3			*Escriba en este espacio cuántas unidades de observación se consideran riesgosas bajo el riesgo 3.*		
TOTAL	*Escriba en este espacio el número de variables involucradas en el análisis de riesgos*			*Escriba en este espacio el porcentaje de unidades de observación riesgosas con respecto al total de unidades de observación*	*Escriba en este espacio todas las técnicas de anonimización que utilizará para minimizar la ocurrencia de todos los riesgos.*

Fuente: DANE- DIRPEN

Es importante destacar que el equipo de trabajo deberá realizar la selección de las técnicas de anonimización, teniendo en cuenta el tipo de variables involucradas en cada uno de los riesgos y las propiedades estadísticas que se desean conservar en la base de datos.

Las técnicas de anonimización más comunes fueron presentadas en el subproceso anterior, con una referencia bibliográfica que permitirán al lector profundizar en su aplicación.

La Tabla 17, *Planteamiento de técnicas de anonimización para cada uno de los riesgos identificados*, sirve como insumo para la toma de decisiones en el análisis de viabilidad que se realizará en la siguiente etapa.

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

4.5. Etapa IV: Análisis de viabilidad

En esta etapa, se describen algunos criterios a tener en cuenta por el equipo de trabajo, en el momento de analizar la utilidad del proceso de anonimización de la base de datos.

En términos generales, 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. Frente a esto, el equipo de trabajo deberá analizar las necesidades de los usuarios, las limitaciones normativas, las políticas de la entidad y los aspectos temáticos de la base de datos.

Para adelantar este proceso, se emitirá el concepto de viabilidad del proceso de anonimización, basándose en los siguientes criterios:

1. **Revisión temática y normativa:** Después de la revisión normativa, planteada en la *sección 4.2.2.* y la revisión de la documentación temática prevista en la *sección 4.4.2.*, el equipo analizará si se encontró alguna norma, ley o una directriz temática de la entidad productora de la información, que impida la publicación de la **mayoría** 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**: Después de analizar la selección de las técnicas previstas en la *sección 4.4.2.*, el equipo podría identificar que ninguna de éstas permite que la base de datos anonimizada conserve las propiedades estadísticas definidas en la *sección 4.2.3.*, por lo que podría considerar que **no es viable** realizar el proceso de anonimización.

3. **Nivel de utilidad de la información**: cuando el equipo revise todas las limitaciones a nivel normativo y las técnicas disponibles para las variables contenidas en la base de datos, deberá analizar 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**.

Cuando el equipo identifique que el proceso de anonimización se ve afectado por el primer criterio, definitivamente, no es viable. Sin embargo, cuando se presentan los criterios 2 y 3, existe una forma alternativa de anonimizar la base datos:

- ✓ El equipo definirá propiedades estadísticas más flexibles para que la base de datos anonimizada pueda conservar y garantizar la utilidad de la información para los usuarios. La flexibilidad de las propiedades estadísticas puede darse al cambiar el nivel de desagregación geográfica o temática, modificándolas a categorías más generales, al aumentar la posible variación en las propiedades globales o al eliminar variables sensibles que son de alto riesgo de identificación.

  Por ejemplo: Un equipo deseaba publicar la variable “Ingreso promedio por hogar” a nivel departamental. Sin embargo, con las técnicas existentes, evidenció que aún se podían identificar algunas unidades de observación. Por lo tanto, decidió modificar la desagregación de nivel departamental a nivel regional, que, aunque no conserva la utilidad que se esperaba, sigue siendo información relevante para el usuario y además permite la réplica de las cifras publicadas por la entidad.

Finalmente, si el equipo logra proponer nuevas propiedades estadísticas que permiten que la base de datos anonimizada siga siendo útil para los usuarios, el proceso se considera **viable**.

A modo de resumen frente a los tres criterios, el equipo de trabajo podría diseñar un marco para establecer el riesgo máximo tolerable para anonimizar la base de datos, teniendo en cuenta el siguiente esquema.

Tabla 18. Criterios para analizar la viabilidad de la anonimización de la base de datos

Criterio	Nivel de afectación	Decisión
1. **Revisión temática y normativa**	**Alta** Todas las variables presentan riesgos	No es viable la anonimización de la base de datos
2. **Técnicas de anonimización**	**Medio** Se podrían identificar algunas unidades de observación	Flexibilizar las propiedades estadísticas de la base de datos (Etapa III. Selección de las técnicas de anonimización)
3. **Nivel de utilidad de la información**		

Fuente: DANE- DIRPEN

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

En términos conceptuales, el análisis de viabilidad que realiza el equipo de trabajo en esta etapa, puede verse de forma detallada en el Recuadro 5

Recuadro 5. Análisis de viabilidad en términos conceptuales

En en la literatura, al buscar identificar el nivel máximo de riesgo tolerable para anonimizar la información, esto es el balance entre los beneficios y los riesgos de anonimizar la información, se puede utilizar el siguiente esquema de revisión.

En esta gráfica, se muestra la relación entre la **utilidad de los datos,** o alguna medida cuantitativa de la calidad estadística de la base de datos, y el **riesgo de divulgación de la información,** o la probabilidad de re-identificación de la información.

Al hacer el balance entre la base de datos original, y la base de datos anonimizada, se pueden tener varios escenarios; por ejemplo, en el punto (1), se observa que la base de datos original presenta un alto nivel de utilidad para el usuario de la información, y a su vez, tiene una alta probabilidad de que las unidades de observación sean identificadas. De hecho, la base de datos original, se encuentra por encima del *máximo riesgo tolerable*, punto (2), riesgo que el equipo encargado del proceso de anonimización consideró no podría sobrepasarse para así conservar la confidencialidad de **todas** las unidades de observación.

El segundo escenario, puede darse cuando el equipo decide no publicar la base de datos, punto (4), dado que la utilidad de la información para el usuario es nula, así como el riesgo de identificar las unidades de observación. Por lo que no tendría beneficios explícitos la decisión de anonimizar o mantener la base de datos en su estado original.

Finalmente, en el último escenario, después de la aplicación de las técnicas de anonimización, el dato publicado, punto (3), es el que se considera que conserva el equilibrio adecuado entre la utilidad de la información para el usuario y el riesgo de identificación de las unidades de observación, además, se encuentra por debajo del *máximo riesgo tolerable.*

Fuente: Adaptado de Hundepool et al., (2012): 5

4.6. Etapa V: Aplicación de técnicas de anonimización

En esta etapa el grupo de trabajo 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 las técnicas el equipo debe tener en cuenta los siguientes pasos:

- ✓ Clasificación de las técnicas asociadas a cada riesgo identificado, tal y como se evidencia en la Tabla 17 *Planteamiento de técnicas de anonimización para cada uno de los riesgos identificados*

- ✓ Elegir el *software* que le permita implementar las técnicas de anonimización elegidas con el fin de planear de manera eficiente el proceso de anonimización. Dentro de los paquetes de *software* a tener en cuenta están μ-Argus, desarrollado por el instituto de estadística de Holanda, τ-Argus cuando se tienen datos agregados o tablas, SAS, también se ha publicado un paquete de anonimización para R, entre otros.

- ✓ Es importante que el equipo de trabajo cree rutinas (algoritmos) que le 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:

    - ✓ Cargue de base de datos a anonimizar
    - ✓ Tipos de riesgos identificados y la explicación de éstos
    - ✓ Consolidación de riesgos identificados
    - ✓ Técnica de anonimización a aplicar
    - ✓ Verificación de riesgos
    - ✓ Exportación de base de datos anonimizada.

    A continuación, se presenta una parte de la rutina utilizada por el DANE en el ejercicio simulado para la anonimización de la Encuesta Anual de Comercio (EAC), utilizando el paquete estadístico SAS:

```
libname EAC "\\BASES_ANONIMIZADAS\EAC"; run;
data base11;
set work.'2011_if_0000'n;
run;
```

En esta parte de la rutina, se presenta la forma en que fue **cargada la base de datos** a anonimizar.

```
PROC SQL;/*RIESGO 1 MAXIMOS A NIVEL NACIONAL*/
CREATE TABLE MAXIMOS
AS SELECT *,  MAX(VENTA) AS MAX_VENTA
FROM BASE11;
QUIT;
DATA MAXIMOS_VENTAS;
SET MAXIMOS;
IF VENTA = MAX_VENTA THEN ID_RIESGO1 = 1;
RUN;
```

Posteriormente, se presenta la forma en que identificó los **valores máximos** de la variable **Ventas,** que corresponde al riesgo 1 previsto en la base de la EAC.

```
PROC FREQ DATA=EAC.EAC_RIESGO2016 ORDER=INTERNAL;
    TABLES Riesgo1 /  SCORES=TABLE;
    TABLES Riesgo2 /  SCORES=TABLE;
    TABLES Riesgo3 /  SCORES=TABLE;
    TABLES Riesgo4 /  SCORES=TABLE;
    TABLES RiesgoTotal /  SCORES=TABLE;
RUN;
```

Seguidamente, se relaciona la forma en que se consolida el **número de unidades de observación riesgosas;** de acuerdo con los diferentes riesgos planteados en la EAC.

```
/*GENERAR UN IDENTIFICADOR*/
data ranqueo (where =(ranqueo in (1,2,3)));
set orden;
retain ranqueo 0;
if first.llave then do ranqueo=0;
end;
ranqueo=ranqueo+1;
by llave;
run;

/*APLICA TÉCNICA: PROMEDIAR LOS DATOS MAS ALTOS
DE TODAS LAS VARIABLES POR LLAVE*/
proc sql;
create table Metodo_1 as select
CIIU4,
intio, mean (intio) as promedio_intio,
/*VARIABLES, MEAN(VARIABLE) AS PROMEDIO_VARIABLE,*/
from ranqueo
group by llave; quit;
```

Finalmente, se presenta la programación para aplicar la **técnica de microagregación** cuyo objetivo es eliminar, uno de los riesgos de identificación planteado en la base de datos de la EAC.

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

Ejemplo de la Etapa V: Aplicación de técnicas de anonimización

Retomando la base **COL20**, esta cuenta con 6 variables que son identificadores directos, 22 variables que son pseudoidentificadores y 3 variables no confidenciales.

Antes de identificar y aplicar las técnicas de anonimización, se define que, en este ejercicio simulado, el objetivo del proceso de anonimización es ***mantener los promedios de las variables cuantitativas (edad, ingresos anuales, ingresos mensuales, número de hijos nacidos vivos, número de personas que componen el hogar, número de bienes raíces)*** con una variación inferior al 5% a nivel departamental.

El equipo de trabajo lista los riesgos y acorde con el tipo de variable, indica la técnica de anonimización a utilizar, como se evidencia en la Tabla 19.

Tabla 19. Riesgos identificados y técnica de datos a utilizar

RIESGO	DESCRIPCIÓN	VARIABLES INVOLUCRADAS	TIPO DE VARIABLE	NÚMERO DE UNIDADES DE OBSERVACIÓN RIESGOSAS	TECNICA DE ANONIMIZACIÓN
1	Las 3 personas con los ingresos anuales más altos por departamento	- DEPARTAMENTO - INGRESOS ANUALES	- CATEGÓRICA - CUANTITATIVA	99	Microagregación
2	Las personas pertenecientes a un grupo étnico a nivel departamental	- DEPARTAMENTO - GRUPO ÉTNICO	- CATEGÓRICA - CATEGÓRICA	36	Recodificación
3	Todas las personas que vivan en una vivienda con un número de habitaciones por encima del promedio departamental	- DEPARTAMENTO - NUMERO DE HABITACIONES DE LA CASA	- CATEGÓRICA - CUANTITATIVA	235	Redondeo; Recodificación
4	Todas las personas que hayan viajado fuera del país más veces que el promedio departamental	- DEPARTAMENTO - NUMERO DE VIAJES REALIZADOS FUERA DEL PAÍS	- CATEGÓRICA - CUANTITATIVA	225	Redondeo; Recodificación
5	Todas las personas con posgrado en aquellos departamentos con menos de 4 personas a ese nivel de escolaridad	- DEPARTAMENTO - NIVEL DE ESCOLARIDAD	- CATEGORICA - CATEGORICA	31	Recodificación

Fuente: Elaboración propia

Para la base de datos anonimizada a publicar se han eliminado las siguientes variables: **CÉDULA, TIPO DE IDENTIFICACIÓN, NOMBRE, APELLIDOS, DIRECCIÓN, BARRIO, MUNICIPIO, FECHA DE NACIMIENTO, RH.** Esta acción se realiza dada la naturaleza de la información registrada en estas nueve variables, consideradas como identificadores directos, y por tanto sensibles, puesto que permitirían reconocer a las unidades de observación en la base de datos.

Posteriormente, se identifica el porcentaje de unidades riesgosas teniendo en cuenta el listado de riesgos previstos en la Tabla 19 y priorizándolos de acuerdo con el número de riesgo que pueden afectar a las unidades de observación, tal y como se detalla en la Tabla 20.

Tabla 20. Porcentajes de Unidades de observación por números de riesgos

Riesgo Total	N° de casos	% de casos
Ningún riesgo	109	22%
Un solo riesgo	204	41%
Dos riesgos	137	27%
Tres riesgos o más	46	10%
TOTAL	**496**	**100%**

Fuente: Elaboración propia

A partir de la información de la Tabla 20, se encuentra que el 77% de las unidades de observación presentan algún tipo de riesgo; por lo tanto, se define para este ejercicio que aquellas unidades que presenten uno o más riesgos, son las que deben ser anonimizadas.

Adicional al establecer el porcentaje de unidades que presentan algún riesgo, es importante detallar el porcentaje de unidades de observación que pertenecen a cada uno de los riesgos priorizados como se muestra en la Tabla 21.

Tabla 21. Porcentajes de Unidades de observación para cada riesgo priorizado

Tipo de riesgo		N° de casos	% de casos
Riesgo 1	Sin riesgo	397	80%
	En riesgo	99	20%
Riesgo 2	Sin riesgo	460	92%
	En riesgo	36	7%
Riesgo 3	Sin riesgo	261	52%
	En riesgo	235	47%
Riesgo 4	Sin riesgo	271	54%
	En riesgo	225	45%
Riesgo 5	Sin riesgo	465	93%
	En riesgo	31	6%

Fuente: Elaboración propia del equipo de trabajo

El procedimiento de anonimización para **COL20,** se describe a continuación:

1. Para lograr minimizar el riesgo de identificación que se genera por la variable *ingresos anuales* (Riesgo 1), se analizan cada una de las técnicas expuestas para variables **continuas** (Tabla 16).

   El riesgo 1 se presenta cuando los ingresos anuales son desagregados geográficamente ya que se tendrían tres unidades de observación riesgosas por departamento. Para este caso, la técnica de redondeo **no** minimizaría el riesgo de identificación dado que no perturbaría la información lo suficiente.

   La técnica de intercambio de datos sólo cambiaría los ingresos anuales de la persona y mantendría los mismos valores altos en los ingresos, lo cual permitiría la identificación de las unidades de observación.

   La adición de ruido perturbaría la información en un nivel muy bajo y los ingresos altos aún seguirían siendo riesgosos.

   Por estas razones, se decidió que la **técnica de microagregación** es la que mejor perturba la información, dado que reemplaza los 3 valores más altos por un mismo valor, y la probabilidad de identificar alguna de las unidades de observación, disminuye considerablemente.

   Para iniciar la aplicación de la técnica, se deben:

   ✓ Identificar las unidades de observación asociadas al riesgo 1, es decir las 3 personas con ingresos altos para cada departamento.

- ✓ Al identificarlas, se verifica que cumplan con características similares; por ejemplo, se revisó el grado de escolaridad y la frecuencia con la estas personas viajan por fuera del país. En este caso, se observa que se tienen tres unidades de observación que cuentan con posgrado, y al mismo tiempo, han viajado una o más veces por fuera del país.

2. Para aplicar la microagregación, se recodificó la variable nivel de escolaridad así:

   - ✓ **Bachilleres** para las unidades de observación que habían reportado en su nivel de escolaridad primaria (1), secundaria (2), educación media (3).

   - ✓ **Técnicos** para las unidades de observación que habían reportado en su nivel de escolaridad técnico (4), tecnólogos (5).

   - ✓ **Profesionales** para las unidades de observación que habían reportado en su nivel de escolaridad profesional (6), posgrado (7), maestría (8), doctorado (9).

   - ✓ **No reporta** para las unidades de observación que no reportaron información referente a su nivel de escolaridad.

3. Con esta recodificación, se logra obtener en la base de datos anonimizada, el mismo grado de escolaridad entre las tres observaciones riesgosas por la variable ingresos mensuales.

4. Respecto a los departamentos, se verifica que las tres unidades de observación con ingresos más altos en la base de datos anonimizada, coincidan con la frecuencia asociada a los viajes que realizan fuera del país.

5. Posteriormente, se calcula el promedio de las variables edad, ingresos anuales, ingresos mensuales, número de hijos nacidos vivos, número de personas que componen el hogar, número de habitaciones de la casa, número de bienes raíces, número de viajes fuera del país.

6. Los promedios encontrados en el punto anterior, se reemplazan en los valores originales de las tres unidades de observación para lograr así minimizar el riesgo de identificación. Este proceso es repetido para cada departamento con las tres unidades de observación con ingresos anuales más altos.

Para ejemplificar esta actividad de reemplazamiento de los valores (punto 6) se presenta a continuación para el departamento del Amazonas, la aplicación de la técnica de microagregación de los ingresos anuales para las unidades de observación 71, 127 (Unidad de observación Riesgosa) y 417 que se encuentran en la base de datos **COL20**.

Las siguientes tablas presentan la información que tienen dichas unidades de observación en la base de datos, antes de la aplicación de la técnica (Tabla 22), y después la información que tendrán dichas unidades de observación en la base de datos anonimizada (Tabla 23).

Tabla 22. Unidades de observación riesgosas en Amazonas. Datos originales

IDPERSONA	DEPARTAMENTO	EDAD	INGRESOS ANUALES
71	AMAZONAS	52	$ 87,595,509
127	AMAZONAS	49	$ 127,500,652
417	AMAZONAS	86	$ 114,654,116

Fuente: DANE- DIRPEN

Tabla 23. Unidades de observación riesgosas en Amazonas. Datos anonimizados

IDPERSONA	DEPARTAMENTO	EDAD	INGRESOS ANUALES
71	AMAZONAS	62	$ 109,916,759
127	AMAZONAS	62	$ 109,916,759
417	AMAZONAS	62	$ 109,916,759

Fuente: DANE- DIRPEN

7. Al momento de asignarle el valor promedio a las tres unidades de observación con los ingresos más altos por departamento, las variables ***edad, número de hijos nacidos vivos, número de habitaciones de la casa, número de bienes raíces, número de viajes fuera del país,*** se observa que estos valores registran **decimales.**

   A partir de ese resultado, se procede a aplicar la técnica de Redondeo, esto es, dejar los registros sin unidades decimales. Este procedimiento se realiza para cada una de las variables, de esta forma se mantienen los valores exactos, estableciendo como condición que los valores sean números enteros.

   Por ejemplo, una mujer que tiene en la variable número de hijos nacidos vivos y que tenga reportado 3,2, en la base anonimizada este valor será redondeado a 3.

Con la aplicación de estas técnicas se logra minimizar el riesgo de identificación de las unidades de observación para el Riesgo 1 (*Las 3 personas con los ingresos anuales más altos por departamento*), y el Riego 5 (*Todas las personas con posgrado en aquellos departamentos con menos de 4 personas en ese nivel de escolaridad*) (Tabla 19).

8. Para el Riesgo 2, es decir para *Las personas pertenecientes a un grupo étnico a nivel departamental*, se aplica la técnica de Recodificación a la variable ***grupo étnico.***

   En este caso, se define que las observaciones que reportaron ser afrocolombiano (1), indígena (2) y Rrom (3), se les asignaría la categoría "*Pertenece a una etnia"* y las que no habían reportado pertenecer a una etnia, se les asignaría la categoría *"No Pertenece a una etnia".* Al aplicar esta técnica, se enmascara la pertenencia al grupo étnico por parte de las unidades de observación, siendo la única técnica que nos permite este tipo de proceso.

   En la base de datos anonimizada no se podrá diferenciar a qué etnia pertenecen las unidades de observación; sin embargo, los usuarios sí podrán calcular el porcentaje de las unidades de observación que pertenecen o no a un grupo étnico por departamento.

9. Para el Riesgo 3, es decir para *Todas las personas que vivan en una vivienda con un número de habitaciones por encima del promedio departamental,* se aplica la técnica de Recodificación a la variable, ***número de habitaciones de la casa,*** teniendo en cuenta los valores obtenidos en el paso número 5.

10. Se definen los siguientes rangos a la variable número de habitaciones de la casa: de "1 – 3"; "4 – 10" y, posteriormente, se asignó a cada unidad de observación, su correspondiente categoría.

    Con esta técnica, se minimiza la identificación de las unidades de observación que viven en una vivienda con un número de habitaciones por encima del promedio departamental. Con la recodificación realizada en la base anonimizada, se puede tener la frecuencia de las viviendas en cada uno de los diferentes rangos establecidos.

11. Para el riesgo 4, esto es, *Todas las personas que hayan viajado fuera del país más veces que el promedio departamental,* se aplica la técnica de Recodificación para la variable ***número de viajes realizados fuera del país.***

    A partir de esto, se definen las siguientes categorías para la variable: entre 0 y 2 viajes "0 - 2" y para más de 2 viajes "más de 2"

Con esta técnica, en la base de datos anonimizada se puede tener la frecuencia del número de viajeros en cada uno de los diferentes rangos establecidos.

Al aplicar estas técnicas de anonimización se obtiene una primera versión de la base de datos anonimizada. En este caso, se recomienda tener copias de las bases de datos, indicando claramente los criterios aplicados para disminuir los riesgos identificados en la base de datos original.

En resumen, para comprender los pasos y recapitular las acciones desarrolladas en la base **COL20**, es importante destacar que durante el proceso de anonimización se aplicó la técnica de *Recodificación* a tres variables (***número de viajes fuera del país, grupo étnico, número de habitaciones de la casa***) y se eliminaron el 28,1% de las variables iniciales, esto es, 9 de las 32 variables iniciales.

Con la primera versión de la base de datos anonimizada se procede a evaluar si las unidades de observación identificadas como riesgosas ya no presentan algún riesgo, en caso de presentarse nuevas observaciones con riesgo se deben buscar técnicas alternativas de anonimización. La primera versión de la base de datos anonimizada será insumo para la etapa de evaluación de los resultados obtenidos.

4.7. Etapa VI: Evaluación de resultados del proceso

En esta etapa el equipo de trabajo, procederá a evaluar los resultados del proceso de anonimización, al validar que los riesgos de identificación de las unidades de observación se hayan minimizado y que las variables de la base de datos conserven las propiedades estadísticas deseadas.

Esta etapa se divide en tres subprocesos así:

- ✓ Revisión de propiedades estadísticas de la base de datos original contra la base de datos anonimizada
- ✓ Reevaluación de riesgos de identificación
- ✓ Creación del Informe Final del Proceso de Anonimización – IFPA

4.7.1. Revisión de propiedades estadísticas de la base de datos original contra la base de datos anonimizada

Después de la aplicación de técnicas de anonimización, se obtiene una primera versión de lo que sería la base de datos anonimizada.

En esta etapa, se comparan las propiedades estadísticas de la base de datos anonimizada con respecto a la base de datos original, y se proponen medidas correctivas (verificación del proceso, aplicación de nuevas técnicas de anonimización, entre otros), en caso de que no se cumplan las propiedades como se esperaban.

El equipo de trabajo calculará las principales medidas estadísticas sobre todas las variables de la base de datos anonimizada. Con estos resultados, se podrá comparar y concluir, si el proceso de anonimización conservó o no las propiedades estadísticas esperadas con respecto a la base original. Estas propiedades pueden variar según el objetivo que el equipo de trabajo se haya trazado; en algunos casos, se requiere que las medidas globales de las variables (media, varianza, coeficiente de variación, entre otros) se conserven en la base de datos anonimizada, por lo que el equipo de trabajo verificará que las diferencias entre estas medidas no sean significativas.

Por otro lado, existen casos en que el objetivo del proceso es que las propiedades estadísticas de la base de datos anonimizada conserven la tendencia de ciertas variables, la relación lineal entre una variable de estudio y algunas variables explicativas (regresión lineal), o medidas descriptivas sobre segmentaciones (o subpoblaciones) de interés en la población analizada. En estos casos el equipo de trabajo definirá y calculará las medidas que le permitan comparar si las propiedades estadísticas se conservan en la base de datos anonimizada.

Cuando el equipo de trabajo revise la base de datos anonimizada y considere que cumple las propiedades estadísticas, procederá al subproceso de reevaluación de riesgos de identificación. Sin embargo, cuando considere que la base de datos anonimizada no las cumple, el equipo:

1. Revisará detalladamente la aplicación de las técnicas de anonimización propuestas en la aplicación de técnicas de anonimización. Verificará que no haya **errores de procesamiento** o que no se esté usando inadecuadamente la técnica.

2. Cuando tenga certeza de que las técnicas planteadas han sido aplicadas adecuadamente, volverá a la Etapa III, donde identificará y aplicará técnicas de anonimización ***alternativas***[12] que no perturben demasiado la base de datos original y permitan el objetivo estadístico planteado.

[12] En la etapa selección de técnicas de anonimización se muestran las técnicas más adecuadas por tipo de variable y nivel de perturbación de la información.

4.7.2. Reevaluación de riesgos de identificación

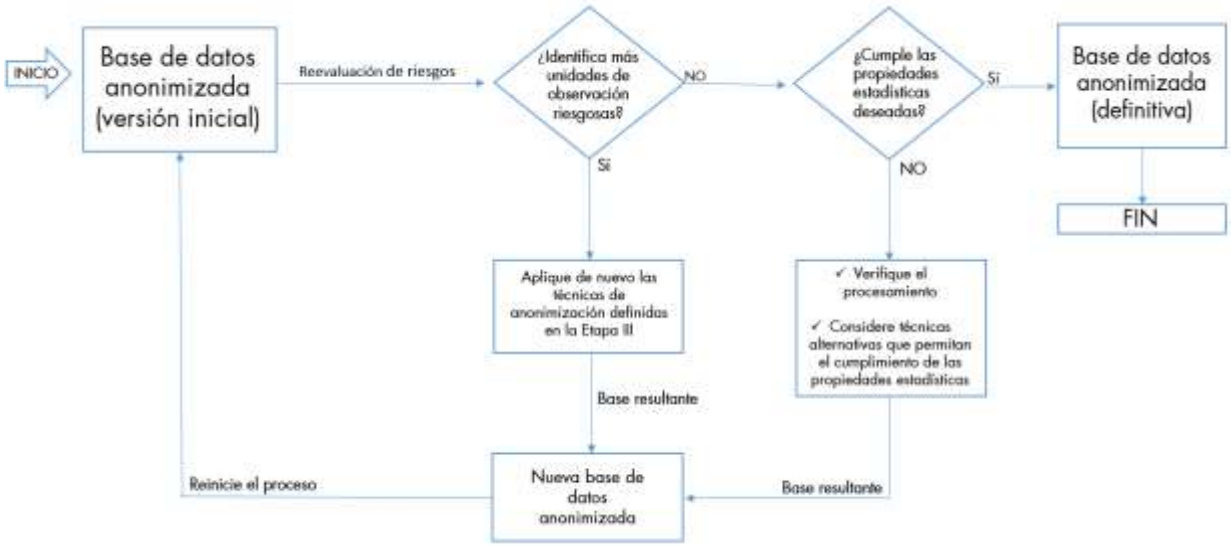
Un criterio que el equipo de trabajo debe tener en cuenta para evaluar el proceso de anonimización es la reevaluación de riesgos de identificación.

Para ello, el equipo de trabajo, con base en el listado de riesgos priorizados planteado en el Análisis de Riesgos (Etapa II), hará un nuevo análisis de riesgos de identificación, tal y como se hizo con la base original, con el objetivo de identificar nuevas unidades de observación riesgosas.

Es probable, que las técnicas de anonimización aplicadas enmascaren la información sensible identificada, pero a su vez, conviertan nuevas unidades de observación en unidades riesgosas.

En la Figura 1, se presenta un flujograma de cómo el equipo de trabajo puede realizar los dos primeros subprocesos de esta etapa.

Figura 1. Flujograma reevaluación de riesgos de identificación



Fuente: DANE- DIRPEN

4.7.3. *Creación del Informe Final del Proceso de Anonimización – IFPA*

La etapa de evaluación de resultados finaliza cuando el equipo de trabajo obtiene la base de datos anonimizada en su versión definitiva. Esta base de datos final debe cumplir las propiedades estadísticas esperadas y no permitir la identificación de información sensible de las unidades de observación que se había previsto en la base de datos original.

El equipo de trabajo debe construir el **Informe Final del Proceso de Anonimización - IFPA**, el cual debe seguir la siguiente estructura:

1. CARACTERÍSTICAS DEL EQUIPO DE TRABAJO E INSUMOS

En esta sección es importante que el equipo describa con qué insumos inicia el proceso de anonimización. Debe tener en cuenta:

- ✓ El equipo de trabajo encargado del proceso de anonimización *(Sección 4.3.4)*. Incluir el rol que tiene cada persona en el equipo, esto es, persona con conocimiento temático de la base de datos o la persona encargada del procesamiento de la base de datos.
- ✓ Paquete de *software* utilizado.
- ✓ Descripción de la base de datos original. Esta sección puede incluir: dimensiones de la base de datos, formato, periodo, la operación estadística o el registro administrativo a que corresponde la información.
- ✓ Diccionario de datos (en caso de que cuente con uno).

2. REVISIONES PREVIAS AL PROCESO DE ANONIMIZACIÓN:

En esta sección el equipo documentará los hallazgos encontrados en la Etapa I del proceso. Debe tener en cuenta:

- ✓ Análisis exploratorio de la base de datos.
- ✓ Fundamentos legales que respalden o impidan la publicación de la información.
- ✓ Historial de solicitudes de información por parte de los usuarios.
- ✓ Propiedades estadísticas que se espera que la base de datos anonimizada conserve.

3. ANÁLISIS DE RIESGOS DE IDENTIFICACIÓN DE LAS UNIDADES DE OBSERVACIÓN

En esta sección el equipo documentará todos los riesgos planteados en la Etapa II. Debe incluir el Informe de Riesgos como se indica en la sección 4.3.4.

4. SELECCIÓN DE TÉCNICAS A IMPLEMENTAR

En esta sección, el equipo documentará las técnicas que escogió para cada uno de los riesgos planteados en la Etapa II. Es suficiente con que el equipo incluya la Tabla 17: *Planteamiento de técnicas de anonimización para cada uno de los riesgos identificados.*

5. ANÁLISIS DE VIABILIDAD

En esta sección se documentará el concepto de viabilidad sobre el proceso de anonimización al que el equipo llegó. Debe justificar claramente, cuáles son las razones para considerar viable o no viable el proceso.

6. APLICACIÓN DE LAS TÉCNICAS DE ANONIMIZACIÓN

En esta sección el equipo documentará las rutinas utilizadas en la programación del proceso de anonimización.

7. EVALUACIÓN DE RESULTADOS

En esta sección el equipo documentará los hallazgos encontrados en la evaluación de resultados. Puede tener en cuenta:

- ✓ ¿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?

Finalmente, con la creación del IFPA el proceso de anonimización quedará debidamente documentado. En caso de que el proceso no sea viable el informe incluye las razones normativas, temáticas y procedimentales. Por otro lado, si el proceso es viable, el informe permite la continuación del proceso de anonimización en próximas bases de datos de la misma operación estadística o del mismo registro administrativo.

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

Ejemplo de Etapa VI: Evaluación de resultados

Continuando con el ejemplo de **COL20**, después de aplicar las técnicas de anonimización definidas en la sección anterior, en esta etapa se verifica el cumplimiento de las propiedades estadísticas esperadas y se re evalúan los riesgos de identificación.

Como se definió en la Etapa I, la propiedad estadística que se desea mantener en la base de datos anonimizada, es que los promedios de las variables numéricas por departamento no presenten una variación superior al 5%. Para esto, se calculan las variaciones entre los promedios departamentales de las variables que en la base de datos anonimizada continúan siendo cuantitativas. Algunas variables, como *número de habitaciones en la vivienda* y *número de viajes fuera del país* se convirtieron en variables categóricas para disminuir el riesgo de identificación. Las variaciones en las variables se presentan a continuación:

Tabla 24. Variaciones de los promedios de las variables numéricas a nivel departamental

DEPARTAMENTO	INGRESOS ANUALES	INGRESOS MENSUALES	NÚMERO DE PERSONAS COMPONEN EL HOGAR	NÚMERO DE BIENES RAICES
AMAZONAS	0.00%	0.00%	-2,44%	-3,57%
ANTIOQUIA	0.00%	0.00%	1,18%	0,00%
ARAUCA	0.00%	0.00%	-2,22%	4,00%
SAN ANDRÉS	0.00%	0.00%	-3,45%	0,00%
ATLÁNTICO	0.00%	0.00%	-1,23%	0,00%
BOGOTÁ, D.C.	0.00%	0.00%	1,30%	-2,13%
BOLÍVAR	0.00%	0.00%	0,00%	3,85%
BOYACÁ	0.00%	0.00%	0,00%	-2,70%
CALDAS	0.00%	0.00%	0,00%	0,00%
CAQUETÁ	0.00%	0.00%	2,63%	6,67%
CASANARE	0.00%	0.00%	0,00%	-3,57%
CAUCA	0.00%	0.00%	-3,12%	0,00%
CESAR	0.00%	0.00%	0,00%	0,00%
CHOCÓ	0.00%	0.00%	0,00%	0,00%
CÓRDOBA	0.00%	0.00%	-1,79%	-2,94%
CUNDINAMARCA	0.00%	0.00%	1,08%	1,82%
GUAINÍA	0.00%	0.00%	-4,17%	-3,70%
GUAVIARE	0.00%	0.00%	-3,23%	0,00%
HUILA	0.00%	0.00%	1,45%	2,38%
LA GUAJIRA	0.00%	0.00%	3,23%	-3,13%

DEPARTAMENTO	INGRESOS ANUALES	INGRESOS MENSUALES	NÚMERO DE PERSONAS COMPONEN EL HOGAR	NÚMERO DE BIENES RAICES
MAGDALENA	0.00%	0.00%	0,00%	4,55%
META	0.00%	0.00%	0,00%	0,00%
NARIÑO	0.00%	0.00%	-2,94%	0,00%
NORTE DE SANTANDER	0.00%	0.00%	0,00%	0,00%
PUTUMAYO	0.00%	0.00%	0,00%	5,88%
QUINDIO	0.00%	0.00%	0,00%	-2,17%
RISARALDA	0.00%	0.00%	1,85%	3,13%
SANTANDER	0.00%	0.00%	0,00%	0,00%
SUCRE	0.00%	0.00%	0,00%	0,00%
TOLIMA	0.00%	0.00%	-1,41%	1,47%
VALLE DEL CAUCA	0.00%	0.00%	-1,05%	-1,85%
VAUPÉS	0.00%	0.00%	-4,17%	5,26%
VICHADA	0.00%	0.00%	0,00%	-2,94%

Fuente: DANE- DIRPEN

Se observa que las variaciones en las variables *ingresos anuales* e *ingresos mensuales* es de 0% para todos los departamentos, esto se debe a que las variables son continuas, en cambio, las variables *número de bienes raíces* y *número de personas que componen el hogar*, que fueron microagregadas y además redondeadas (pues deben ser un número entero) presentan variaciones mayores al 0%.

Por otro lado, es evidente que la variable "*número de bienes raíces*" es la única que **no** conserva en todos los departamentos una variación inferior al 5%. En el ejercicio se verificó que la variación no corresponde a un error de procesamiento.

El equipo de trabajo considera que las variaciones superiores al 5% para los departamentos de Caquetá y Putumayo son permitidas desde el punto de vista temático, por lo tanto, se considera que con las técnicas utilizadas se mantienen las propiedades estadísticas establecidas al inicio del proceso de anonimización

Un paso que debe tenerse en cuenta para la evaluación de resultados, es la re-identificación de unidades de observación riesgosas. En este caso, sobre la base de datos anonimizada, se buscan unidades de observación que cumplan con los riesgos planteados en la Etapa II. Se obtiene:

Tabla 25. Re-identificación de unidades de observación riesgosas

Riesgo	Unidades de observación riesgosas	Porcentaje
1	0	0%
2	0	0%
3	0	0%
4	0	0%
5	8	1.61%

Fuente: DANE- DIRPEN

La re-identificación de nuevas unidades de observación riesgosas, muestra que la ocurrencia de los riesgos 1,2, 3 y 4 fue minimizada correctamente. Sin embargo, dado que aún aparecen 8 unidades de observación riesgosas por el riesgo 5, el equipo propone nuevas categorías para la variable *nivel de escolaridad, unificando las categorías técnicos y profesionales en "técnico o profesional" y dejar la categoría Bachiller* y así minimiza el riesgo de identificación.

Finalmente, después de que se tuvieron en cuenta los cambios que sugiere la etapa de evaluación de resultados, se obtiene la base de datos anonimizada **en su versión final**. Esta base cumple con las propiedades estadísticas esperadas y no tiene riesgo de que las unidades de observación sean identificadas.

5. RECOMENDACIONES FINALES

A partir de los puntos revisados en la Guía, como un paso final, el DANE presenta algunas recomendaciones que el equipo puede tener en cuenta para un proceso de anonimización efectivo:

1. La base de datos que será anonimizada debe cumplir con los requerimientos iniciales previstos en la primera etapa de la guía. Esto permitirá que el proceso de anonimización no se vea afectado por errores de captura de la información o falta de reglas de validación.

2. El planteamiento de los riesgos de identificación de las unidades de observación debe estar debidamente documentado. De manera que, en caso de cambios en el equipo de trabajo responsable de la anonimización, se pueda asegurar una trazabilidad del proceso.

3. Las rutinas del *software* estadístico utilizadas en el proceso de anonimización, deben estar debidamente explicadas para permitir que el proceso pueda repetirse cuando se cuenten con actualizaciones de las bases de datos.

4. La documentación correspondiente a la aplicación de las técnicas de anonimización debe ser información restringida para el equipo de trabajo, debido a que esta información permitiría a terceros revertir el proceso de anonimización y exponer la información de las unidades de observación que son anonimizadas.

5. Si la base de datos que se dispone para uso de los diferentes usuarios relaciona cuadros de salida publicados por la entidad del SEN a través de sus diferentes medios, es importante que genere una documentación que permita entender la información contenida en la base de datos anonimizada, para que los usuarios puedan replicar las cifras que son publicadas.

6. Utilizar diferentes medios de difusión y estrategias de publicidad para visualizar la información anonimizada que la entidad posee ante las demás entidades pertenecientes al SEN, permitiendo el acceso y uso de microdatos para la producción y difusión de estadísticas oficiales.

6. BIBLIOGRAFÍA

Libros, papers y otros documentos

CBS. (2008). *Manual del usuario de mu-Argus.* Disponible en CBS: http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf. Recuperado el 20 de Mayo de 2018

Departamento Administrativo Nacional de Estadística (DANE) (2017). Plan Estadístico Nacional del Sistema Estadístico Nacional (SEN) 2017-022. Disponible en : https://www.dane.gov.co/files/sen/PEN-2017-2022.pdf Consultado el 28 de junio de 2018. (2017). Estrategias del Plan Estadístico Nacional.

Departamento Administrativo Nacional de Estadística (DANE) (2017). *Código Nacional de Buenas Prácticas del Sistema Estadístico Nacional.* Disponible en: https://www.dane.gov.co/files/sen/bp/Codigo_nal_buenas_practicas.pdf

Departamento Administrativo Nacional de Estadística (DANE) (2017). *Norma Técnica de la Calidad del Proceso Estadístico. NTC PE 1000.* Disponible en: http://www.dane.gov.co/files/sen/normatividad/NTC_Proceso_Estadistico.pdf. Consultado el 28 de junio de 2018..

Departamento Administrativo Nacional de Estadística (DANE). (2016). *Metodología Encuesta Anual de Comercio – EAC.* Disponible en: http://www.dane.gov.co/files/sen/normatividad/NTC_Proceso_Estadistico.pdf. Consultado el 28 de junio de 2018.

Departamento Nacional de Planeación (DNP) (2018). *Documento Conpes 3918: Estrategia para la implementación de los objetivos de desarrollo sostenible ODS en Colombia.* Disponible en: https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3918.pdf Consultado el 20 de junio de 2018. (marzo 2018)..

Domingo-Ferrer J., Drechsler J. and Polettini S. (2009) Report on synthetic data files. Technical report, Deliverable of Project ESSNET-SDC

Drechsler J., Bender S. and R¨assler S. (2008a) “Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel”. Transactions on Data Privacy 1(3),): 105–130.

Fienberg S.E. (1994) “A radical proposal for the provision of micro-data samples and the preservation of confidentiality”. Technical Report 611, Carnegie Mellon University Department of Statistics.

Fienberg S.E. and Makov U.E. (1998) “Confidentiality, uniqueness and disclosure limitation for categorical data. “ Journal of Official Statistics 14(4),): 385–397.

Hundepool Anco et al. (2012). *Statistical Disclosure Control*, John Wiley & Sons.

Liew C.K., Choi U.J. and Liew C.J. (1985) “A data distortion by probability distribution”. ACM Transactions on Database Systems 10, : 395–411.

Ministerio de Salud. (s.f.). *Lineamientos para la Anonimizacion de Datos del Sistema Nacional de Estudios y Encuestas Poblacionales para la Salud.* Disponible en: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/GCFI/lineamientos-anonimizacion-sistema-encuestas.pdf. Recuperado el 20 de mayo de 2018

Morales, B. (2017). *Introducción a la anonimización de datos.* DANE. Bogotá.

Oficina de Información del Comisionado, 2012 *Anonymisation:Managing Data Protection Risk Code Of Practice. Disponible en: https://ico.org.uk/media/1061/anonymisation-code.pdf*

Ramaswamy, R., Franconi, L. , & Polettini, S. (2008). *User's Manualrisk Models) Peter-paul De Wolf (pram) Josep Domingo and Vicenc Torra (numerical Micro Aggregation and Rank Swapping) about the Name Argus*. Disponible en: https://www.semanticscholar.org/paper/User's-Manualrisk-Models)-Peter-paul-De-Wolf-(pram)-Ramaswamy-Franconi/60133f3338740eb13cb14034d45a3151300af092. Recuperado el 20 de mayo de 2018

Reiter J.P. (2005a) “Releasing multiply-imputed, synthetic public use microdata: an illustration and empirical study. “ Journal of the Royal Statistical Society, Series A 168,:185–205.

Templ, Matthias, Bernhard Meindl, Alexander Kowarik, and Shuang Chen. (2014). “Introduction to Statistical Disclosure Control (SDC).” IHSN Working Paper No. 007.

Woodcock S.D. and Benedetto G. 2007 Distribution-preserving statistical disclosure limitation. Disponible en Available at SSRN: http://ssrn.com/abstract=931535.

Leyes y normatividad

Asamblea Nacional Constituyente (1991) *Constitución Política de Colombia.* Disponible en: http://www.secretariasenado.gov.co/senado/basedoc/constitucion_politica_1991.html

Congreso de la República de Colombia, Ley 1753 de 2015 “Por la cual se expide el Plan Nacional de Desarrollo 2014-2018 “Todos por un nuevo país”. Disponible en: http://www.secretariasenado.gov.co/senado/basedoc/ley_1753_2015.html

Congreso de la República de Colombia, 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 y se dictan otras disposiciones”. Disponible en: http://www.secretariasenado.gov.co/senado/basedoc/ley_1266_2008.html

Congreso de la República de Colombia, Ley Estatutaria 1581 de 2012 “Por la cual se dictan disposiciones generales para la protección de datos personales”. Disponible en: http://www.secretariasenado.gov.co/senado/basedoc/ley_1581_2012.html

Congreso de la República de Colombia, Ley Estatutaria 1712 de 2014 “Por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional y se dictan otras disposiciones.”. Disponible en: http://www.secretariasenado.gov.co/senado/basedoc/ley_1712_2014.html

Departamento Nacional de Estadística, Ley 79 de 1993, “Por la cual se regula la realización de los censos de población y vivienda en todo el territorio nacional.”. Disponible en: https://www.dane.gov.co/files/acerca/Normatividad/Ley79_1993.pdf

Departamento Nacional de Estadística, Decreto 1743 de 2016, “Por el cual se reglamenta el artículo 160 de la ley 1753 de 2015, se adiciona el título 3 a la parte 2 del libro 2 del Decreto 1170 de 2015 Único del Sector Administrativo de Información Estadística”. Disponible en: https://www.dane.gov.co/files/sen/normatividad/decreto_1743_noviembre_1_2016.pdf

Presidencia de la República de Colombia. Decreto 2573 de 2014 “Por el cual se establecen los lineamientos generales de la Estrategia de Gobierno en línea, se reglamenta parcialmente la Ley 1341 de 2009 y se dictan otras disposiciones”, Disponible en: http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=60596

Presidencia de la República de Colombia. Decreto 1377 de 2013 “Por el cual se reglamenta parcialmente la Ley 1581 de 2012”, Disponible en: http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=53646

Sistemas de información

Departamento Administrativo Nacional de Estadística (DANE). http//:www.dane.gov.co. Consultado el 4 de julio de 2018. Elementos conceptuales básicos sobre el Sistema Estadístico Nacional.

ANEXO A

Ejemplo diccionario de datos. Disponible en formato xlsx.

ANEXO B

Base de datos COL20. Disponible en formato xlsx.