Alcaldía de Medellín
Distrito de
Ciencia, Tecnología e Innovación

GUÍA PARA LA ANONIMIZACIÓN DE BASES DE DATOS QUE SE PRODUCE Y DIFUNDE EL DISTRITO ESPECIAL DE CIENCIA, TECNOLOGÍA E INNOVACIÓN DE MEDELLÍN

DEPARTAMENTO ADMINISTRATIVO DE PLANEACIÓN
SUBDIRECCIÓN DE PROSPECTIVA, INFORMACIÓN Y EVALUACIÓN ESTRATÉGICA
UNIDAD DE PRODUCCIÓN DE INFORMACIÓN ESTRATÉGICA
EQUIPO DE GESTIÓN DE INFORMACIÓN ESTADÍSTICA

Versión 1

DISTRITO ESPECIAL DE CIENCIA, TECNOLOGIA E INNOVACIÓN DE MEDELLÍN
2022

INTRODUCCIÓN

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).

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 la entidad.

En la Constitución Política de Colombia de 1991, específicamente en el artículo 15, se establece elderecho que tienen todas las personas a conservar su intimidad, mantener su buen nombre y a laprotección y garantía del Habeas Data. En relación con la protección de datos personales, el régimen normativo colombiano dispone de la Ley 1266 de 2008 y la Ley 1581 de 2012. En términos de la confidencialidad, 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, además de establecer disposiciones sobre la recolección, tratamiento y circulación dedatos personales en el país (Congreso de Colombia, 2008).

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.

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. Como se establece en el principio 11: Confidencialidad, el cual indica: “Las entidades del SEN deben establecer protocolos de seguridad y confidencialidad que protejan la privacidad de las fuentes en el proceso estadístico o en el intercambio de microdatos.”

ALCANCE:

La Guía para la anonimización de bases de datos va dirigido a los encargados de la producción y difusión de información estadística del Distrito de Medellín, con el propósito orientar sobre el proceso de anonimización de bases de datos que provienen de registros administrativos y de operaciones estadísticas (por muestreo, censos, a partir de registros administrativos y operaciones estadísticas derivadas, de acuerdo a los lineamientos del Departamento

Administrativo Nacional de Estadística – DANE, como ente coordinador del Sistema Estadístico Nacional SEN.

1. ANONIMIZACIÓN DE LOS MICRODATOS

El proceso de anonimización consiste en la creación de mecanismos y estrategias que eviten la identificación de las fuentes de información, ya que con los datos capturados por las diferentes variables que integran la operación estadística, al cruzarse es posible identificar fácilmente a las personas o empresas que proveen la información y dichos datos pueden ser utilizados con fines inapropiados.

Este ejercicio a las bases de datos se convierte en una necesidad dada la normatividad vigente respecto a la reserva estadística, la protección de la confidencialidad y los datos personales. La anonimización debe realizarse teniendo presente que hay que conservar las variables claves paracumplir con el objetivo de la operación estadística, esto implica modificar lo menos posible la información y preservar su potencial de aprovechamiento, tratando de introducir el menor ruido posible en los resultados y protegiendo la privacidad de las fuentes de información.

Para realizar un correcto proceso de anonimización, desde el momento en el que se están seleccionando las variables que va a capturar la operación estadística, estas se deben clasificar encategorías

I. Variables geográficas, por ejemplo: dirección, barrio, comuna, departamento, distrito
II. Variables de identificación de personas o empresas: NIT, actividad económica, edad, sexo,sector, nivel educativo o etnia
III. Variables con datos numéricos: ingresos, edad, estatura, número de hijos, número dehermanos, área, diámetro
IV. Variables de carácter sensible o confidencial: número de identificación, NIT, ingresos, gastos, costos, impuestos declarados, teléfono, e-mail
V. Variables sin restricción de acceso al público: sexo, edad, años de escolaridad

La segunda recomendación es tratar de reducir la cantidad de información personal o confidencial que se solicita a la fuente, hecho que se relaciona de forma directa con los métodos y mecanismos de captura y registros de la información para consolidar la base de datos. Posteriormente las variables confidenciales que se haya determinado deben ser recolectadas, se les debe asignar un acceso limitado de consulta y modificación, una forma sencilla de cumplir con este paso es sacar una copia de la base de datos bruta y a esta suprimirle o eliminarle las variables confidenciales, de manera que los integrantes del equipo de trabajo puedan acceder a la base de datos depurada y solo algunos usuarios permitidos puedan acceder a la base de datos bruta. A partir de este

momento los demás trabajos de análisis, procesamiento y validación, deben hacerse con la base de datos depurada y la base de datos bruta quedará como un respaldo de acceso limitado.

Algunas de las variables que son susceptibles de eliminación son las siguientes:

- Nombres.
- Fecha de nacimiento.
- Fecha de constitución en Cámara de Comercio.
- Números de teléfono y fax.
- Números de identificación: cédula de ciudadanía, pasaporte, tarjeta de identidad, númerosasociados a la seguridad social, licencias de conducción, NIT, RUT, RUP, RUE, etc.
- Direcciones de correo electrónico.
- Números de cuentas bancarias.
- Identificadores del vehículo, placa, etc.
- Identificadores de dispositivos móviles y números de serie.
- Direcciones de IP.
- Identificadores biométricos.
- Fotografías e imágenes similares.
- Cualquier otro número único de identificación.
- Dirección de domicilio.

Los microdatos para difusión no incluirán por ningún motivo las variables confidenciales que se determine permitan la identificación del registro, o identificadores directos o indirectos de carácter personal.

Los archivos anonimizados deber ser verificados y aprobados por los expertos temáticos del equipo y personal administrativo, así como con el equipo del Plan Estadístico Territorial perteneciente al Departamento Administrativo de Planeación – DAPM para asegurar la reserva estadística antes de su publicación.

1.1. 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.

Se recomienda que el equipo de trabajo este compuesto por:

- Profesionales especializados que conozcan la temática de la base de datos de la operaciónestadística o del registro administrativo que se va a anonimizar, ya que sobre

ellos recae laidentificación de las variables sensibles y con riesgo y, la determinación de la viabilidad de la anonimización.

- El equipo de trabajo debe manejar paquetes estadísticos, como R, SAS, SPSS, Stata, para que realicen los análisis exploratorios de las bases de datos y la aplicación de las técnicas de anonimización.

Para la ejecución satisfactoria del proceso de anonimización, se debe tener en cuenta los siguientes requerimientos:

- Los responsables de la operación estadística o del registro administrativo, deben proporcionar la base de datos a anonimizar en su última versión y con la verificación de la consistencia y coherencia.
- Así mismo, deben proporcionar el modelo entidad- relación[1] y el respectivo Diccionario dedatos[2] "el diccionario de datos especifica 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".
- Se debe disponer de una infraestructura tecnológica para salvaguardar la información, los paquetes estadísticos con licencias, las claves de acceso y la administración de la base anonimizada junto con las notas metodológicas de uso que se requieran. En este sentido, es aconsejable para cuando se carece de estos requerimientos, establecer acuerdos o convenios con entidades o dependencias que cuenten con los recursos y la infraestructurarequerida.

2. PROCESO DE ANONIMIZACIÓN:

Los encargados de anonimizar la base de datos, deben contemplar 6 etapas, estas etapas presentan distintos subprocesos y actividades a desarrollar para realizar la anonimización de la base de datos deseada. Las etapas son las siguientes:

- Revisiones previas de la información a anonimizar

[1] El modelo entidad - relación es un modelo conceptual para bases de datos relacionales que permite representar cualquier abstracción, percepción y conocimiento en un sistema de información formado por un conjunto de objetos denominados entidades y relaciones, a partir de una representación gráfica llamada diagrama entidad-relación, adoptando el enfoque más natural del mundo real. Página 5. En: https://www.dane.gov.co/files/sen/lineamientos/Recomendaciones-paraelaborar-modelos-entidad-relacion.pdf

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

- Análisis de riesgos del proceso
- Selección de técnicas para avanzar en la anonimización
- Selección de técnicas y viabilidad de anonimización
- Evaluación de resultados del proceso.

2.1. Fase I: Revisiones previas de la información a anonimizar

En esta etapa, el equipo encargado debe realizar una revisión de los insumos disponibles para la ejecución del proceso. La etapa se compone de tres subprocesos:

a- Análisis exploratorio de la base de datos. El equipo de trabajo debe:

- Caracterizar cada una de las variables contenidas en la base de datos teniendo en cuenta sison cuantitativas (continuas o discretas) o categóricas, de igual manera, 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.

La caracterización de la base de datos se puede hacer teniendo en cuenta como ejemplo la Tabla1.

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 temática*
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

- Calcular las medidas descriptivas estadísticas para cada una de las variables cuantitativas, estas medidas servirán como insumo para el análisis de riesgos, el análisis de la viabilidaddel proceso de anonimización y la evaluación de resultados. Ver ejemplo Tabla 2.

Tabla 2. Principales medidas descriptivas para variables cuantitativas

VARIABLE CUANTITATIV	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

- Calcular las frecuencias de las variables categóricas, estas distribuciones puedenrealizarse teniendo en cuenta la siguiente tabla:

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

VARIABLE CATEGÓRICA Ejemplo: ORGANIZACIÓN JURÍDICA	NÚMERO DE UNIDADES DE OBSERVACIÓN QUE CUMPLEN LA CATEGORÍA	% DE REGISTROS QUE CUMPLEN LA CATEGORÍ
Categoría 1 *Ejemplo:* Sociedad en comandita simple		*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

- Realizar una revisión temática de la base de datos a anonimizar, teniendo en cuenta la documentación de la operación estadística o el registro administrativo. Esta revisión temática servirá como insumo en el planteamiento de los riesgos de identificación de lasunidades de observación y en el análisis de viabilidad del proceso.

b- Revisión normativa sobre protección de datos e identificación de usuarios 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 (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).

Para la identificación de los usuarios se recomienda realizar una revisión histórica de las

solicitudes 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

Tabla 4. Ejemplo de clasificación de solicitudes recibidas

Categoría	Descripción	No. de solicitudes
Tipo de usuario	Entidades Públicas	
	Entidades Privadas	
	Investigadores	
	Instituciones Académicas	
	Otros	
Tipo de solicitudes	Derechos de petición	
	Acceso a información	
	Actualización de datos	
Clasificación de variables	Variables de identificación	
	Variables de ubicación	
	Variables temáticas	
Niveles de desagregación	Nacional	
	Departamental	
	Municipal	
	Temática	
	Otros	
Períodos solicitados de la información	Anual	
	Mensual	
	Trimestral	
	Diaria	
Frecuencia de la solicitud	Mensual	
	Diaria	
Objetivo, finalidad, uso	Investigaciones o estudios	
	Académica	
	Política Pública	

Basados en la clasificación y cuantificación de las solicitudes, el equipo de trabajo podrá identificarlas 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 tipo de información que puede suministrar en la base de datos anonimizada para responder a lasnecesidades de los usuarios, y que, además, no expongan la identificación de las unidades de observación.

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

El equipo de trabajo deberá establecer las propiedades estadísticas que se deben mantener en labase 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
- Mantener propiedades globales de las variables
- Mantener cifras por niveles de desagregación geográfica o temática
- Mantener correlaciones entre variables

2.2. Fase II: Análisis de riesgos del proceso

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 y deben ser protegidas por el tipo o combinación de información que contienen (datos personales, datos sensibles, valores únicos, extremos, atípicos, valores combinados con poca participación), para ello se debe:

a- Clasificar las variables por su nivel de sensibilidad[3]:

Las variables pueden clasificarse en:

- Identificadores directos:
  Son todas aquellas variables que contienen información sensible de identificación o ubicación de las unidades de observación. (nombre, apellidos, fecha de nacimiento, barrioy dirección, número de cédula de una persona, NIT de una empresa, dirección de una entidad, entre otros.)

- Pseudoidentificadores:
  Son todas aquellas que, combinadas con otras variables, conllevan a la identificación de las unidades de observación. Ejemplo: combinación del nivel de escolaridad con el ingreso promedio mensual en cierto municipio. En este caso, son 3 variables

[3] 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)

consideradas como pseudoidentificadores, que permiten la identificación de algunas unidades de observación.

- No confidenciales:
  Son todas aquellas que no permiten la identificación de las unidades de observación de labase de datos, ni siquiera cuando son combinadas con pseudoidentificadores.

  Calcular y analizar la distribución de frecuencias para las variables categóricas consideradas "sensibles", para identificar las clases que presentan poca participación paraun nivel de desagregación temático o geográfico y que pueden ser consideradas riesgosas.

b- Planteamiento de riesgos de la base de datos

- Identificar, caracterizar y evaluar, a partir de la base de datos y del diccionario de datos, lasvariables o registros que presentan riesgos, debido a que contienen información de la unidad de observación:
  - Con datos sensibles que afectan la intimidad de las personas y pueden causar discriminación
  - Con datos personales que permite su identificación o ubicación
  - Con variables temáticas (pseudoidentificadores) que al combinarlas pueden ser riesgosas, por presentar baja frecuencia.
- Establecer y realizar posibles combinaciones entre las variables directas y las variables temáticas para diferentes niveles de desagregación, a fin de verificar las unidades de observación susceptibles de identificar por los usuarios y que requieren ser protegidas aplicando las técnicas de anonimización.
- Definir los niveles de desagregación temática y geográfica a los cuales se puede presentar la información sin que presente riesgos de identificación y sobre los cuales se deben conservar las propiedades globales.
- Definir las variables "importantes" que deben conservar las propiedades estadísticas (porcentajes, promedios, tasas, porcentajes de variación, distribuciones) para desagregaciones en las bases de datos anonimizadas.
- Identifica bases de datos externas, que al combinar algunas de sus variables, con lasvariables de la base de datos a anonimizar, podrían conducir a un riesgo.
- Identifica y lista las unidades de observación con riesgo de identificación a partir de larealización de las anteriores actividades y cálculos.

2.3. Fase III: Selección de técnicas y viabilidad de anonimización

La anonimización de datos se define: proceso para impedir que, a partir de un dato o de una combinación de datos de una misma fuente o de diferentes fuentes de datos, se logre identificarsujetos individuales ya sean individuos, empresas o establecimientos, u otro tipo de unidades deobservación. (DANE, 2018).

En esta etapa el equipo define las técnicas de anonimización que van a aplicar sobre las variablesidentificadas como riesgosas.

Para la selección de la técnica de anonimización se debe tener en cuenta:

- El tipo de variable: categórica o cuantitativa;
- Que la técnica minimice la ocurrencia de los riesgos identificados;
- Que el procedimiento minimice la perdida de información o de perturbación de los datos
- Que se conserven las propiedades estadísticas
- Que se respeten los resultados globales a los niveles de desagregación requeridos;
- Que no se pierda la utilidad de reutilizar de la información.

Planteadas las técnicas de anonimización, se realiza el informe final del proceso de anonimización,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 labase de datos original.

El informe debe contener:

- La categorización de los riesgos que se presentan con las variables involucradas
- Las variables con identificadores directos que deben ser eliminadas
- Las variables y los registros considerados riesgosas porque tienen categorías con bajoporcentaje a un nivel requerido
- Las variables consideradas pseudoidentificadoras que al combinarlas con otras son consideradas riesgosas
- La descripción del riesgo que se ocasiona sobre cuales registros en las variables
- El porcentaje de registros que se ven involucrados en cada riesgo
- La técnica de anonimización a aplica.

A partir de los insumos anteriores el equipo elabora la propuesta de viabilidad de la anonimización, para lo cual analiza y determina que se cumplan los siguientes aspectos a fin de emitir un concepto favorable:

- Que la base de datos anonimizada a entregar pueda cumplir con las necesidades

de losusuarios.

- Que no existan alguna norma, ley, políticas en la entidad o aspectos temáticos sobre laentrega de la base de datos
- Que las técnicas de anonimización a aplicar no permitan identificar unidades de observación
- Que las propiedades estadísticas se conserven a los niveles de desagregación temáticas ygeográficos requeridos por los temáticos
- Conservar y garantizar que la información anonimizada sigue siendo de utilidad para losusuarios.

a- Técnicas de anonimización:

Las técnicas de anonimización se refieren a procedimientos que modifican en forma sistemática los datos, de modo que se minimiza el riesgo de identificar las unidades de observación. Consiste en incluir, suprimir y/o modificar algunos datos que se considera pueden llevar a deducir la persona, establecimiento o en general a la unidad de observación. Con la aplicación de estas técnicas se producen supresiones parciales o totales en los resultados de los diferentes tipos de variables y reducen el nivel de detalle de la base 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 de las personas, establecimientos u otra unidad de observación. Dentro de este tipo de técnicas se encuentra:

Tabla 5. Técnicas basadas en la en la no perturbación de datos según el tipo de variable

TÉCNICAS	DESCRIPCIÓN	TIPO DE VARIABLE	EJEMPLO VARIABLES
ELIMINACION DE VARIABLES	Consiste en eliminar una variable con toda su información, debido a que contiene información sensible o datos personales y con ella se puede identificar de forma directa la unidad de observación.	En variablescategóricas	CEDULA; TIPO DE IDENTIFICACION; NOMBRE; APELLIDOS; DIRECCIÓN;BARRIO; MUNICIPIO; FECHA DE NACIMIENTO; RH Estas variables son eliminadas porque debido ala información contenida, esposible identificar directamente a las unidadesde observación. Algunas deellas contienen información sensible, por lo tanto, permitirían reconocer las unidades de observación.
RECODIFICACIÓN GLOBAL	Esta técnica consiste en combinar diversas categorías de las variables categóricas en una más general, Es decir, se recodifican los valores de la variable para tener una nueva categoría.	En variables cuantitativaso categóricas	GRUPO ÉTNICO; NÚMERO DE HABITACIONES DE LACASA; NUMERO DE VIAJES REALIZADOS FUERA DEL PAÍS; NIVEL DE ESCOLARIDAD Estas variables son recodificadas para combinarlas categorías de las variables y poder tener en cada nueva categoría una mayor frecuencia de unidades de observación.
CODIFICACIÓN SUPERIOR E INFERIOR	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 presentanvalores máximos y mínimos en el nivel de desagregación geográfico o temático que son de fácil identificación.	En variables continuas o categóricas	EJEMPLO. En una localidad de Medellín es reducido el número de supermercados “grandes” y pueden ser fácilmente identificados por la variable “ventas mensuales”, al presentar valores externos altos, por lo que se utiliza una codificación. por el número reducido de ellos que tienen presencia; por lo que se asigna una codificados a estos supermercados.

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 pseudo identificadoras permitala identificación de las unidades de observación.	En variables categóricas	EJEMPLO. Al cruzar las variables "orientación sexual y "discapacidad", se obtiene una frecuencia de pocos individuos para alguna categoría, lo que hace fácil su identificación, por lo que se reemplazan estas categorías como "datos faltantes".

- 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 suficientementeprecisas como para revelar información sobre casos individuales. Pueden incluirse nuevos datos,suprimir y/o modificar los existentes beneficiando la confidencialidad estadística. Dentro de estetipo de técnicas se encuentra:

Tabla 6. Técnicas basadas en la perturbación de datos según el tipo de variable

TÉCNICAS	DESCRIPCIÓN	TIPO DE VARIABLE	EJEMPLO VARIABLES(BASE COL20)
MICRO AGREGACIÓN	Con esta técnica se reemplazan algunos datos de una variable, por el promedio de ese subconjunto de datos, debido a que, por presentar valores extremos o únicos para un nivel de desagregación geográfica o temática, resulta fácil la identificación de las unidades de observación. 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 VIAJESFUERA DEL PAÍS Estas variables son micro agregadas porque se busca proteger la identificaciónde las unidades de observaciones con los ingresos anualesmás altos por departamento.
	Consiste en sustituir losvalores de	En variables	

REDONDEO	aquellas variables que tienen decimales por valores redondeados (cero decimales), en aquellas unidades de observación considerada con riesgo. Suele aplicarse esta técnica luego de realizar la micro agregación para las variables que no admiten valores con decimales. Ej.Número de hijos.	cuantitativas	EDAD; NÚMERO DE HIJOS NACIDOS VIVOS; NUMERO DE HABITACIONES DE LA CASA; NÚMERO DE BIENES RAÍCES; NUMERO DE VIAJESFUERA DEL PAÍS. Estas variables son redondeadas para mantener lainformación de las variables enunidades enteras después de su micro agregación.
INTERCAMBIODE 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 deidentificación. Este intercambio de datos se realizade manera aleatoria entre pares de observaciones (con riesgo de identificación y sin riesgo).	En variables cuantitativas o categóricas	

Otra de las técnicas poco utilizadas en los procesos de anonimización de bases de datos, son **los métodos de datos sintéticos**, que no son considerados los más adecuados, ya que modifican todos los registros en la base de datos, aunque conservan las mismas tendencias y correlaciones.La técnica consiste en disponer al usuario de una base con datos simulados, a partir del diseño de algoritmos de simulación que utiliza modelos de regresión cuantílica, imputación adicional y datoscombinados. (Hundepool, et al., 2010: 58)

b- Propiedades estadísticas a conservar en la base de datos

El equipo de trabajo deberá establecer las propiedades estadísticas que se deben mantener en labase de datos anonimizada[4] por lo tanto, los encargados de realizar la anonimización de la base de datos debe garantizar la conservación de las siguientes propiedades estadísticas:

- *Mantener tendencias en las variables a través del tiempo***.** Por ejemplo, si la base de

[4] DANE. Guía para la anonimización de bases de datos en el Sistema Estadístico Nacional. 2018. Página 24. En: http://www.dane.gov.co/_les/sen/lineamientos/DSO-020-LIN-08.pdf

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 2020, al publicar la base de datos anonimizada el equipo de trabajo desea garantizar que esta tendencia se conserve.

- *Mantener propiedades globales de las variables.* Definir cuáles de las medidas estadísticasse 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 basede 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 procesode 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.* 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, el equipo requiere que se continúe presentando información sobre la variable pertenencia a un grupo étnico para cada categoría a nivel localidad.

- *Mantener correlaciones entre variables.* Definir 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. Por ejemplo, los temáticos han definido que los coeficientes de correlación entre las variables nivel de concentración de sustancias contaminantes y las mediciones sobre la calidad del aire, se mantenga con el fin de no distorsionar los análisis.

Referencias

DANE. (2017). *Código Nacional de Buenas Prácticas para las Estadísticas Oficiales.* Obtenido de https://www.dane.gov.co/files/sen/bp/Codigo_nal_buenas_practicas.pdf

DANE. (2018). *Guía para la anonimización de bases de datos.* Obtenido de https://www.dane.gov.co/files/sen/registros-administrativos/guia-metadatos.pdf

DANE. (2020). *Lineamientos para el proceso estadístico en el Sistema Estadístico Nacional.* Recuperado el 3 de Septiembre de 2021, dehttps://www.dane.gov.co/index.php/lineamientos-para-el-proceso-estadistico

(DANE, Código Nacional de Buenas Prácticas para las Estadísticas Oficiales,

2017)(DANE, Guía para la anonimización de bases de datos, 2018)

Elaboró Maira Alejandra Santofimio Contratista	Revisó: John Fredy López Ossa Norha Esneida Leon Henao	Aprobó: Jasblleidy Pirazán García Cargo: Subdirectora Administrativa Subdirección de Prospectiva, Información y Evaluación Estratégica Departamento Administrativo de Planeación