GUÍA PARA EL CONTROL DE DIVULGACIÓN ESTADÍSTICA EN MICRODATOS

INSTITUTO NACIONAL DE ESTADÍSTICAS

Diciembre 2021

DIRECCIÓN NACIONAL
DEPARTAMENTO DE METODOLOGÍAS E INNOVACIÓN ESTADÍSTICA
SUBDEPARTAMENTO DE INVESTIGACIÓN ESTADÍSTICA

Instituto Nacional de Estadísticas
31/ Diciembre/ 2021

ÍNDICE

Reconocimientos

Este documento es el resultado del trabajo colaborativo realizado por diversas áreas de la institución, quienes participaron en las distintas temáticas abarcadas para la elaboración de esta *guía para el control de divulgación estadística en microdatos.* Este trabajo se inicia en 2019, pero es durante el 2021 que esta tarea se consolida y concreta mediante un estándar que permite resguardar la seguridad de la información recolectada y procesada por la institución. Este trabajo será un eje para continuar cultivando la confianza institucional, que históricamente se caracteriza por utilizar las mejores prácticas disponibles que provean un resguardo de la calidad estadística.

Como agradecimiento a los distintos profesionales que trabajaron en la ejecución y elaboración de esta guía, comenzaremos mencionando a las áreas de trabajo pertenecientes a la *Subdirección Técnica*, representado por sus cuatro representantes, iniciando por el Departamento de Estadísticas de Precios, para continuar con el Departamento de Estadísticas del Trabajo con representación de las áreas de estadísticas estructurales, así como los de Ingresos del Trabajo, como tercer participante el Departamento de Estadísticas Económicas, abarcando las áreas de Industrias, Comercio y Servicios, para finalizar con el Departamento de Estadísticas Demográficas y Sociales abarcando las estadísticas socioeconómicas, así como las de condiciones de vida. A continuación de los colaboradores mencionados, debemos señalar a los proyectos censales, que también fueron parte del proceso, en donde encontramos al Censo Agropecuario y Forestal, y al Censo de Población y Viviendas.

En aquellas áreas dependientes directamente de *Dirección Nacional*, debemos mencionar la gran labor realizada por las Direcciones Regionales, que abarcan todo el país, por medio de las macrozonas norte, centro y sur; no se puede dejar de mencionar a Fiscalía y el Departamento de Análisis Estadístico, ambos participantes activos en la mesa de trabajo.

Otro equipo participante en la ejecución de esta guía fue el *Departamento de Infraestructura Estadística*, que estuvo representado por los equipos de Geografía y Marcos Maestros, y por último el equipo de Gobierno y Administración de Datos.

Finalmente, se agradece al equipo organizador de la mesa de trabajo desde el *Departamento de Metodologías e Innovación Estadística* representado por el equipo del Subdepartamento de Investigación Estadística, el que ha sido acompañado desde sus inicios por el Subdepartamento de Calidad y Estándares.

1. INTRODUCCIÓN

1.1. Antecedentes

La recopilación y difusión de estadísticas oficiales es una de las principales responsabilidades del Estado. Las estadísticas constituyen una base fundamental para la toma de decisiones basadas en evidencia y por tanto una administración pública eficaz es posible bajo la concepción de estadísticas oportunas y precisas. Tradicionalmente, para compilar estadísticas oficiales, los sistemas estadísticos nacionales recopilan microdatos que se refieren a datos a nivel de unidad (menor nivel de desagregación) que generalmente se recopilan a través de encuestas, recolección de datos digitales y registros administrativos. Estos datos proporcionan información sobre personas naturales y/o jurídicas, unidades tales como hogares, empresas, organizaciones sin fines de lucro, instalaciones, zonas productivas o incluso áreas geográficas. En gran proporción, los microdatos contienen información de identificación personal y/o confidencial de las unidades de información.

La comunidad estadística ha reconocido la importancia de asegurar esta información para mantener la confianza de las poblaciones a las que sirven. En este sentido, el Código Nacional de Buenas Prácticas Estadísticas del Instituto Nacional de Estadísticas (en adelante, INE), en su principio 4 sobre confidencialidad estadística, establece que el "INE y los demás miembros del Sistema Estadístico Nacional (SEN) deben garantizar la protección y confidencialidad de la información con la que se producen las estadísticas oficiales, así como evitar la identificación de las fuentes" (INE, 2015, pág. 8).

Sin embargo, aunque se reconoce la importancia de proteger los datos individuales, las Naciones Unidas también abogan por la libre difusión de los microdatos. Lo que permite a los usuarios contribuir con investigación, aumenta la transparencia y la responsabilidad de los institutos nacionales de estadística y permite mejoras en la calidad a través de la retroalimentación de los usuarios (United Nations, 2014).

Los principios en competencia de la seguridad de los datos y la difusión de microdatos se someten a arbitraje a través de un dominio de estadísticas llamado Control de Divulgación Estadística (SDC[1], por sus siglas en inglés). Los métodos SDC permiten proteger un conjunto de datos mediante la aplicación de herramientas estadísticas, lo que posibilita a la institución difundir de manera segura el conjunto de datos.

En Chile cabe señalar que, en los últimos años, el aumento constante de la disponibilidad de información que se libera mediante instituciones privadas condujo a concientizar la importancia que tiene la protección de la confidencialidad de la información. También, este escenario presionó a las entidades públicas a aumentar el acceso a la información por parte de la ciudadanía, mediante la creación de la "ley de transparencia" (Ley 20.285), promulgada en el 2008. Sin embargo, continúa siendo esencial dar un tratamiento adecuado a los datos, previo a su publicación, para cumplir con la ley de secreto estadístico

[1] En inglés, *Statistical Disclosure Control*.

(Art.29, Ley 17.374), ley sobre protección de la vida privada (Art.2e, Ley 19.628) y la legislación propia de las entidades públicas.

La experiencia del INE en términos de control de divulgación estadística ha sido dispar y, en general, no se cuenta con un consenso formalizado acerca de los métodos de anonimización o los parámetros de riesgo que se deben cumplir para difundir un conjunto de microdatos. No obstante, se pueden citar algunos esfuerzos institucionales por normalizar este ámbito.

En junio 2009, la Resolución exenta N° 1918, emitida en Santiago el 10 de junio de 2009, expone acerca de una experiencia localizada, sobre el tratamiento que se buscaba dar a datos económicos (INE, Resolución exenta 1918, 2009), fue creada con esos fines, y para eso definió lo que era información privilegiada y las sanciones en caso de que el INE revelara o divulgara algún dato que no estuviera en condiciones de ser publicado. Sin embargo, esta resolución solo protegía los datos económicos, y no se cubren datos de otra naturaleza que necesitaran control a la divulgación.

En 2019, un equipo multidisciplinario compuesto por representantes[2] de la producción estadística institucional, se conformó con el objeto de definir lineamientos para desarrollar un proceso estandarizado de control de divulgación en las operaciones estadísticas que desarrolla el INE, entregando como resultado la primera versión de la "Guía para el control de divulgación estadística en microdatos". Esta guía estableció un proceso estandarizado de doce pasos basados en la guía práctica publicada por el Banco Mundial (Benschop, Machingauta, & Welch, 2021). Además, el desarrollo de los pasos fue seguido de estudios de caso o aplicaciones basadas en operaciones estadísticas INE, mediante el uso del entorno y lenguaje de programación con enfoque al análisis estadístico, R (R Core Team, 2019), particularmente a través de la librería `sdcMicro` (Templ, Kowarik, & Meindl, 2015). El resultado de ese trabajo no se formalizó como un estándar debido a la falta de definiciones de umbrales y parámetros, especialmente en la evaluación de riesgos de divulgación. El trabajo desarrollado por la Mesa de Anonimización hasta el segundo semestre de 2021, haciendo modificaciones y adaptaciones a la realidad del INE, ha buscado establecer acuerdos para implementar como un estándar institucional.

En consecuencia, a nivel institucional se exige normar el subproceso de control a la divulgación estadística o anonimización, a fin de responder de manera adecuada, oportuna y segura a los usuarios que requieren información desagregada y que solicitan las bases de microdatos, al mismo tiempo de tener procedimientos estandarizados en la producción de estadísticas oficiales.

[2] **Subdirección Técnica:** Departamento de Estadísticas de Precios, Subdepartamento de Estadísticas Continuas del Trabajo, Departamento de Estadísticas del Trabajo, Subdepartamento de Estadísticas Estructurales del Trabajo, Subdepartamento de Estadísticas de Industrias, Departamento de Estadísticas Económicas, Subdepartamento de Estadísticas de Comercio y Servicios, Subdepartamento de Estadísticas Socioeconómicas, Departamento de Estadísticas Demográficas y Sociales, Subdepartamento de Censos de Población, Subdepartamento de Condiciones de Vida.
Departamento de Metodologías e Innovación Estadística: Subdepartamento de Investigación Estadística, Subdepartamento de Calidad y Estándares.

1.2. Objetivo del documento

El mapa de procesos del INE corresponde a una guía de navegación donde se representan de manera secuencial todos los procesos de la institución, organizados en tres grandes segmentos: Dirección, Negocio[3] y Soporte. Dentro del segmento de Negocio, se describe el ciclo completo de producción estadística basado en el Modelo Genérico del Proceso Estadístico (GSBPM[4], por sus siglas en inglés) adaptado a la realidad institucional.

A partir del mapa de procesos del INE y de acuerdo con el segmento de Negocio, este documento tiene como objetivo normar y detallar el procedimiento y las principales actividades que deben llevarse a cabo en el subproceso 6.4 "Aplicar control a la divulgación", correspondiente al cuarto eslabón del proceso "Análisis de resultados" centrado en una etapa, la que corresponde 6.4 "Aplicar control a la divulgación" en el mapa de procesos institucional (ver anexo **10.1. Mapa de procesos-Segmento de Negocio**), y aplica para todas las operaciones estadísticas y productos relacionados.

Este subproceso tiene por objetivo "garantizar que los datos y metadatos que se difunden no infrinjan las normas de confidencialidad, de acuerdo con las políticas y normas del INE, o con la metodología especifica diseñada en el subproceso 2.5 "Diseñar el procesamiento y análisis". Esto puede incluir comprobaciones de la divulgación primaria y secundaria, así como la aplicación de técnicas de supresión o perturbación de datos y comprobación de los resultados" (UNECE, 2019, pág. 23).

Es importante señalar que, todos los productos estadísticos desarrollados por el INE basan sus mecanismos de confidencialidad en la "Política de privacidad y protección de la información de identificación personal institucional". Esta política establece que los datos producidos por el INE no pueden hacer referencia expresa a personas que participen en sus estudios. En consecuencia, se excluyen de las bases de datos finales aquellas variables cuya información permita identificar unidades de análisis (personas, viviendas, empresas, etc.) que participaron en el estudio o que implicarán un riesgo elevado de identificación de las mismas (por ejemplo, variables de ubicación de manzana).

1.3. Alcance del documento

El subproceso 6.4 "Aplicar control a la divulgación" se inicia desde la etapa 6.4.1 "Realizar definiciones previas al proceso de anonimización" con la recepción y revisión de productos estadísticos proveniente del subproceso 6.3 "Interpretar y explicar los resultados", hasta la etapa 6.4.5 "Generar reportes y liberar datos", con la disposición de los productos estadísticos anonimizados, con el reporte de anonimización y metadatos actualizados correspondientes. Los procedimientos relacionados con este subproceso que se

[3]Ver https://inechile.sharepoint.com/sites/Intranet/departamentodegestionestrategica/Mprocesos/Paginas/Segmento-Negocio.aspx

[4] En inglés, *Generic Statistical Business Process Model*.

describen en este documento aplican para todas las operaciones estadísticas y productos relacionados cuyo levantamiento de información y/o publicación sea realizado por el INE.

Por tanto, este documento se centra en los métodos y procesos para la liberación de microdatos, ya sea que estos provengan de muestreo, censos, procesos estadísticos provenientes de múltiples fuentes, o registros estadísticos generados por el INE. En consecuencia, el alcance de los procesos que se describen en esta guía se ciñe a proveer directrices circunscritas al campo de los microdatos, por lo que no se encuentran cubiertos los procesos de control de divulgación estadística orientados a tabulados, estadísticas geoespaciales, publicaciones web o visualizaciones de mapas, etc., los que no son cubiertos en esta versión del estándar.

1.3.1. Exclusiones al alcance

Se describe, adicionalmente, el procedimiento frente a los requerimientos de información del Banco Central de Chile (en adelante, BCCh), considerando que constituye una excepción reglada y normada al secreto estadístico. En este sentido, el BCCh puede requerir aquella información que sea indispensable y pertinente para el cumplimiento de sus funciones específicas, expresamente encomendadas en virtud de la Constitución Política de la República (en adelante, CPR), y su Ley Orgánica regulatoria. Por lo anterior, la particular excepción tiene su fundamento y límite en los siguientes antecedentes:

1. **Naturaleza Constitucional:**
    - ✓ Ley N° 18.840, de 1989. El art. 108 de la CPR crea al Banco Central de Chile como un organismo autónomo, técnico, con patrimonio propio y cuyas funciones y atribuciones estarían establecidas por una Ley Orgánica Constitucional (en adelante, LOC).
    - ✓ Esa ley, en su artículo 2° establece que el Banco **se regirá exclusivamente por las normas de esta ley orgánica y no le serán aplicables, para ningún efecto legal, las disposiciones generales o especiales, dictadas o que se dicten para el sector público**.

2. **Atribuciones Estadísticas:**
    - ✓ El artículo 53 de la LOC del BCCh establece su obligación de elaborar las cuentas nacionales: "El Banco **deberá compilar y publicar, oportunamente**, **las principales estadísticas macroeconómicas nacionales**, incluyendo aquellas de carácter monetario y cambiario, de balanza de pagos y las cuentas nacionales u otros sistemas globales de contabilidad económica y social".
    - ✓ Para el ejercicio de dicha función, le entrega la facultad de **exigir a los diversos servicios o reparticiones de la Administración Pública, instituciones descentralizadas y, en general, al sector público, la información que estime necesaria.**

Considerando la normativa previa, el INE puede y debe remitir al BCCh, aquella información solicitada, en la medida que el requerimiento fundado sea pertinente y guarde relación directa con el ejercicio de las

facultades estadísticas individualizadas en el artículo 53 de su Ley Orgánica Constitucional, esto es, aquella información necesaria para la elaboración de estadísticas macroeconómicas nacionales.

Sin perjuicio de lo anterior, es importante destacar que la información que el INE ponga a disposición del Banco Central debe ser de naturaleza indispensable y necesaria para la elaboración de Cuentas Nacionales de la Nación, cuestión que se debe verificar previamente mediante la entrega de un informe técnico por parte del Banco Central al INE, el cual dé cuenta de los argumentos de carácter técnico que fundamenten dicha entrega, el cual por otra parte, debe ser visado y aprobado por la Subdirección Técnica del INE.

La remisión de esta información debe ser efectuada únicamente al jefe de Departamento de Cuentas Nacionales Anuales del Banco Central de Chile, o quien se designe en su reemplazo, previa notificación al INE, y debe ser utilizada exclusivamente en la elaboración de Cuentas Nacionales de la Nación, es decir para fines estadísticos.

1.4. Estructura del documento

Esta guía está dividida en las siguientes secciones principales:

1. La sección **Glosario** proporciona definiciones, conceptos o categorías utilizadas en esta guía, para una comprensión acabada del subproceso.
2. La sección **Control de Divulgación Estadística: Una introducción** provee una descripción general del subproceso de control de divulgación estadística (SDC), justificando su necesidad y el *trade-off (o balance)* entre riesgo y utilidad que caracteriza el subproceso.
3. La sección **Documentos aplicables** entrega un listado de documentos que permiten crear un marco de referencia para el desarrollo del subproceso, tales como: documentación de referencia internacional, manuales, instructivos, políticas o procedimientos, entre otros.
4. La sección **Control normativo** proporciona el marco legal del subproceso, correspondiente a las leyes, resoluciones o decretos que se deben tener en cuenta para su implementación y ejecución.
5. La sección **Liberación de microdatos** aborda los tipos de liberación de conjuntos de microdatos y su aplicabilidad dado el marco legal vigente en Chile.
6. La sección

7. **Lista de** registros cubre el listado de medios de verificación que contribuyen y facilitan el desarrollo del subproceso.
8. La sección **Descripción del subproceso** define el flujo de trabajo del proceso SDC describiendo los pasos que se deben ejecutar para proteger un conjunto de microdatos antes de su difusión.

2. GLOSARIO

Respecto a los términos, conceptos o categorías utilizadas en esta guía se detallan aquellos que son relevantes para la comprensión del subproceso.

Tabla 1: Glosario de términos y conceptos

Término	Definición	Referencia
Adición de ruido	Método basado en agregar o multiplicar un número aleatorio a los valores originales para proteger los datos de la coincidencia exacta con archivos externos. La adición de ruido se aplica típicamente a variables continuas.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Anonimización	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 características estadísticas en los resultados.	(DANE, 2018, pág. 9)
Archivo de datos para uso científico	Archivo de uso científico (SUF[5], por su sigla en inglés), es un tipo de publicación del archivo de microdatos, que solo está disponible para investigadores seleccionados bajo un acuerdo. También conocido como "archivo con licencia", "microdatos bajo contrato" o "archivo de investigación".	(Benschop, Machingauta, & Welch, 2021, pág. 10)
Archivo de datos para uso en centro de datos de investigación controlado o enclave	Son los archivos que pueden ofrecerse a los usuarios bajo condiciones estrictas en un enclave de datos. Se trata de una sala equipada con computadores que no están conectados a Internet ni a una red externa, y del que no se puede descargar información a través de puertos USB u otras unidades. Los enclaves de datos contienen datos que son particularmente sensibles o permiten la identificación directa o fácil de los informantes. Los ejemplos incluyen conjuntos de datos completos de censos de población, encuestas empresariales, etc.	Adaptado de (Benschop, Machingauta, & Welch, 2021, pág. 20)

[5] En inglés, *Scientific Use File*.

Término	Definición	Referencia
Archivo de datos para uso público	Archivo de uso público (PUF[6], por sus siglas en inglés), es un tipo de publicación del archivo de microdatos, que está disponible gratuitamente para cualquier usuario, por ejemplo, en el sitio web del INE.	(Benschop, Machingauta, & Welch, 2021, pág. 10)
Barajado[7]	Método que consiste en enmascarar una variable considerada confidencial mediante la generación de una distribución condicional.	(Benschop, Machingauta, & Welch, 2021, pág. 74)
Base de datos	Una colección lógica de información que está interrelacionada y que se gestiona y almacena como una unidad, por ejemplo, en el mismo archivo informático.	(OCDE, s.f.).
Celdas confidenciales	Las celdas de una tabla que no son publicables debido al riesgo de divulgación estadística se denominan celdas confidenciales.	(OCDE, s.f.)
Clave	Combinación o patrón de variables clave o cuasi – identificadores. También, es usado el término llave.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Codificación superior o inferior	Corresponde a la agrupación de una variable continua en una categoría en los extremos de los valores posibles que agrupa todos los valores mayores o menores a un número (por ejemplo: valores mayores o iguales a 5 quedarán en la categoría "5 o más" mientras que el resto conserva su valor).	Adaptado de (Benschop, Machingauta, & Welch, 2021, pág. 52)
Confidencialidad de los datos	Es una propiedad de los datos, generalmente como resultado de medidas legislativas, que previenen su divulgación no autorizada.	(OCDE, s.f.)
Control de Divulgación Estadística (SDC)	Proceso que busca tratar y alterar los datos para que puedan publicarse o difundirse sin revelar la información confidencial que contiene, mientras que, al mismo tiempo, limitan la pérdida de información debido al anonimato de los datos. En el GSBPM, estos métodos están relacionados con la etapa de difusión y generalmente se basan en restringir la cantidad o modificar los datos publicados.	(Australian Bureau of Statistics, s.f.)

[6] En inglés, *Public Use File*.
[7] En inglés, *shuffling*.

Término	Definición	Referencia
Convenio	▪ Contrato, convención o acuerdo que se desarrolla en función de un asunto específico destinado a crear, transferir, modificar o extinguir una obligación. ▪ Es un acuerdo de voluntades entre dos o más organismos públicos con personalidad jurídica, sobre cualquier cuestión pendiente de resolver. ▪ Son instrumentos jurídicos, suscritos por dos o más organismos de la Administración del Estado, que tienen por finalidad, comprometer la colaboración mutua entre ellos, dentro de las facultades que la ley les confiere para satisfacer necesidades actuales o futuras y que requieren de su formalización, mediante actos administrativos, para producir efectos jurídicos. **Tipos de convenios:** 1. Marco: establece las bases para el intercambio de información, mediante convenios específicos. 2. Específico: consiste en la materialización de un convenio marco, y tiene por objeto señalar específicamente las obligaciones de cada parte, detallando los compromisos que adquiere cada institución.	Fiscalía INE
Datos personales	Son datos de carácter personal o datos personales, "los relativos a cualquier información concerniente a personas naturales, identificadas o identificables".	(Ministerio Secretaría General de la Presidencia, 2020)
Datos originales	Datos a los que no se les aplica algún método de anonimización. También se denominan "datos brutos" o "datos no tratados".	(Benschop, Machingauta, & Welch, 2021, pág. 11)
Divulgación	Se produce cuando una persona u organización reconoce o aprende algo que no sabía sobre otra persona u organización a través de los datos divulgados. Ver también Divulgación de	(Benschop, Machingauta, & Welch, 2021, pág. 9)

Término	Definición	Referencia
	identidad, Divulgación de atributos y Divulgación inferencial.	
Divulgación de atributos	La divulgación de atributos ocurre cuando un usuario puede determinar nuevas características de un individuo u organización con base en la información disponible en los datos publicados. A este usuario se le denominará intruso, ver intruso.	(Benschop, Machingauta, & Welch, 2021, pág. 8)
Divulgación de identidad	La divulgación de identidad ocurre cuando un intruso asocia a un individuo (o grupo) u organización conocida, con un registro de datos publicado.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Divulgación inferencial	La divulgación inferencial ocurre si un intruso puede determinar, a partir de los datos publicados, el valor de alguna característica de un individuo u organización con mayor precisión que lo pretendido.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Encuesta	Investigación sobre las características de una población particular, que utiliza procedimientos estandarizados para recopilar información de la población de estudio (incluidos censos, encuestas de muestra, la recopilación de datos de registros administrativos y actividades estadísticas derivadas) para estimar sus características mediante el uso sistemático de la metodología estadística.	(INE, 2021)[8]
Escenario de divulgación	Describe la información potencialmente disponible para un tercero (por ejemplo: datos del censo, padrones electorales, registro de población, datos recopilados por empresas privadas o incluso datos de encuestas publicadas por el INE), para identificar a los encuestados y las formas en que dicha información se puede combinar con los microdatos establecidos para ser publicados y utilizados para la re-identificación de registros en el conjunto de datos.	(Benschop, Machingauta, & Welch, 2021, págs. 25-26)

[8] Propuesta de conceptos estadísticos para la clasificación de la producción estadística elaborada por el Subdepartamento de Calidad y Estándares, versión 26 de marzo 2021. Actualmente está en proceso de validación institucional y proyección a la construcción de un glosario de conceptos.

Término	Definición	Referencia
Estructura jerárquica	Datos que se componen de colecciones de registros que están interconectados a través de enlaces, por ejemplo, individuos que pertenecen a grupos/hogares o empleados que pertenecen a empresas.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Identificador	Variable/información (o grupo de variables) que puede utilizarse para establecer la identidad de un individuo u organización. Los identificadores pueden conducir a una identificación directa o indirecta.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Identificadores directos	Son variables que identifican inequívocamente unidades estadísticas, como, RUT, ROL, número de seguro social, o nombres y direcciones de empresas o personas. Los identificadores directos deben eliminarse como primer paso del proceso de anonimización.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Identificadores indirectos	Son variables que, si bien no identifican inequívocamente unidades estadísticas, en combinación se pueden vincular a información externa para re-identificar a los informantes en el conjunto de datos publicado. También se les denomina “cuasi-identificadores” o “variables clave”.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Informante	Empresas, autoridades, personas individuales, etc., de quienes se recopilan datos e información asociada para su uso en la compilación de estadísticas.	Adaptado de (OCDE, s.f.)
Intervalo	Un conjunto de números entre dos cotas designadas que pueden o no estar incluidos (abiertos, semiabiertos o cerrados). Los corchetes (por ejemplo, [0, 1]) denotan un intervalo cerrado, que incluye los puntos finales 0 y 1. Los paréntesis, por ejemplo, (0, 1) denotan un intervalo abierto, que no incluye los puntos finales.	Adaptado de (Benschop, Machingauta, & Welch, 2021, pág. 9)
Intruso	Usuario que hace mal uso de los datos publicados al tratar de identificar y divulgar información sobre un individuo u organización, utilizando un conjunto de características conocidas por el usuario.	(Benschop, Machingauta, & Welch, 2021, pág. 9)

Término	Definición	Referencia
K-anonimato	La medida de riesgo k-anonimato se basa en el principio de que, en un conjunto de datos seguro, el número de individuos que comparten la misma combinación de valores (claves) de identificadores indirectos categóricos debe ser superior a un umbral especificado k. Es una medida de riesgo basada en los microdatos que se liberarán, ya que solo tiene en consideración la muestra.	(Benschop, Machingauta, & Welch, 2021, pág. 28)
Metadatos	Datos que entregan la información necesaria para el uso e interpretación adecuada de las estadísticas por parte de las personas usuarias. Los metadatos describen los datos producidos, por medio de la documentación de contenidos relacionados, por ejemplo, con la metodología; el trabajo de campo; el procesamiento; análisis y la calidad; entre otros, de una operación estadística particular. Contexto: Generalmente se hace una distinción entre metadatos estructurales y de referencia. Los metadatos estructurales se utilizan para identificar, describir formalmente, como nombres de dimensiones, diccionarios de variables, descripciones técnicas de conjuntos de datos, ubicaciones de conjuntos de datos, palabras clave para buscar datos, etc. Por ejemplo, los metadatos estructurales incluyen los títulos de las variables y dimensiones de conjuntos de datos estadísticos, así como las unidades empleadas, listas de códigos (por ejemplo, para codificación territorial), formatos de datos, rangos de valores potenciales, dimensiones de tiempo, clasificaciones utilizadas, etc. Los metadatos de referencia (a veces llamados metadatos explicativos) describen los contenidos y la calidad de los datos estadísticos.	(INE, 2021)[9]

Término	Definición	Referencia
	Incluye textos explicativos sobre el contexto de los datos estadísticos, metodologías para la recopilación y agregación de datos, así como características de calidad y difusión.	
Métodos determinísticos	Métodos que siguen cierto algoritmo y producen los mismos resultados si se aplican repetidamente a los mismos datos con el mismo conjunto de parámetros.	(Benschop, Machingauta, & Welch, 2021, pág. 8)
Métodos no perturbativos	Métodos que reducen los detalles en los datos o suprimen ciertos valores (enmascaramiento) sin distorsionar la estructura de datos.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Métodos perturbativos	Métodos que alteran los valores para limitar el riesgo de divulgación al crear incertidumbre en torno a los valores verdaderos, al tiempo que conservan la mayor cantidad de contenido y estructura posible, por ejemplo, microagregación y adición de ruido.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Métodos probabilísticos	Métodos que dependen de un mecanismo de probabilidad o un mecanismo de generación de números aleatorios. Cada vez que se utiliza un método probabilístico se genera un resultado diferente.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Microagregación	Método que se basa en la sustitución de valores para una determinada variable con un valor común para un grupo de registros. La agrupación de registros se basa en una medida de proximidad de variables de interés. Los grupos de registros también se utilizan para calcular el valor de reemplazo.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Microdatos	Corresponde a los datos sobre las características asociadas a las unidades de observación que se encuentran consolidadas en una base de datos. Son observaciones no agregadas o mediciones de las características de la o las unidades de observación, siendo la forma primaria en la que se almacenan los datos y que a partir de esta se derivan los resultados. El conjunto de microdatos es uno de los resultados de la recolección de datos, después de la edición a nivel de unidad e imputación.	(INE, 2021)[9]

Término	Definición	Referencia
Muestra única	Un registro de la muestra con un conjunto particular de características que no se repite en otras observaciones, de modo que el individuo u organización se puede distinguir de otras unidades de la muestra en función de ese conjunto de características.	(Benschop, Machingauta, & Welch, 2021, pág. 10)
Operación estadística	Corresponde a la aplicación de un proceso estadístico sobre una temática de estudio específica, el cual conduce a la producción de información estadística oficial.	(INE, 2021)[9]
Pérdida de información	Se refiere a la reducción del contenido de información en los datos liberados en relación con el contenido de información en los datos sin procesar. A menudo se mide con el uso de medidas analíticas comunes, como regresiones e indicadores. Ver también Utilidad de los datos.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Población única	Un registro en la población con un conjunto particular de características que no se repite en la población, de modo que el individuo u organización puede distinguirse de otras unidades de la población en función de ese conjunto de características.	(Benschop, Machingauta, & Welch, 2021, pág. 10)
Post Randomization Method (PRAM)	Método en el que los puntajes de una variable categórica se alteran de acuerdo con ciertas probabilidades. Por lo tanto, es una clasificación errónea intencional con probabilidades de clasificación errónea conocidas	(Benschop, Machingauta, & Welch, 2021, pág. 10)
Privacidad	Es un concepto que se aplica a las unidades, mientras que la confidencialidad se aplica a los datos. El concepto se define de la siguiente manera: "Es el estatus otorgado a los datos que ha sido acordado entre la persona u organización que proporciona los datos y la organización que los recibe y que describe el grado de protección que se brindará".	(OCDE, s.f.)
Productos estadísticos	Resultado físico o digital producido por una operación estadística que, mediante la presentación de microdatos y metadatos,	(INE, 2021)[9]

Término	Definición	Referencia
	buscan satisfacer las necesidades de los usuarios. Algunos ejemplos de productos estadísticos son: base de datos, publicaciones, mapas, servicios electrónicos, estadísticas agregadas, tabulados y cuadro estadístico, entre otros.	
Protección de datos	Se refiere al conjunto de leyes, políticas y procedimientos motivados por la privacidad que tienen como objetivo minimizar la intrusión en la privacidad de los informantes causada por la recopilación, el almacenamiento y la difusión de datos personales.	(OCDE, s.f.)
Recodificación	Método en el que se agrupan categorías o valores existentes y se reemplazan con nuevos valores, por ejemplo, las categorías "protestante" y "católico" se reemplazan por "cristiano". La recodificación reduce los detalles en los datos y, para las variables continuas, conduce a una transformación de continua a categórica, por ejemplo, creando bandas de ingresos.	(Benschop, Machingauta, & Welch, 2021, pág. 10)
Registro	Un conjunto de datos derivados de un objeto/unidad de estudio, por ejemplo, un individuo (en datos a nivel individual), un hogar (en datos a nivel de hogar) o una empresa (en datos de la empresa). Los registros también se denominan "observaciones".	(Benschop, Machingauta, & Welch, 2021, pág. 10)
Registro administrativo	Registro usado para fines administrativos en un sistema de información administrativa. Contendrá todos los objetos por administrar, sus objetos serán identificables y sus variables se usarán para propósitos administrativos.	(INE, 2021)[9]
Regresión	Proceso estadístico para medir la relación entre el valor medio de una variable y los valores correspondientes de otras variables.	(Benschop, Machingauta, & Welch, 2021, pág. 10)
Riesgo de divulgación	Se refiere a la probabilidad de que ocurra efectivamente una divulgación de la información confidencial de un informante, o una divulgación exacta con un alto nivel de confianza.	Adaptado de (Benschop, Machingauta, & Welch, 2021, pág. 9)

Término	Definición	Referencia
Riesgo global	Es una medida sobre todo el conjunto de datos que agrega los riesgos individuales como la proporción esperada de individuos en una muestra que pueden ser correctamente re-identificados por un intruso. Hay que utilizar con cuidado esta medida, ya que puede esconder altos riesgos individuales con un riesgo global aceptable.	(Benschop, Machingauta, & Welch, 2021, pág. 40)
Riesgo individual	Es la probabilidad de una correcta re-identificación de individuos en los datos divulgados.	Adaptado de (Benschop, Machingauta, & Welch, 2021, pág. 28)
Riesgo jerárquico	Es la probabilidad de una correcta re-identificación de unidades tomando en cuenta la estructura jerárquica de los datos. La estructura jerárquica de un conjunto de datos puede estar dado por ser miembros de un hogar, trabajadores de una empresa o alumnos de un colegio, entre otros ejemplos, el riesgo entonces tomará en cuenta que si se identifica algún miembro de este hogar, empresa o colegio puede que se identifique al resto de sus miembros.	Adaptado de (Benschop, Machingauta, & Welch, 2021, pág. 41)
sdcMicro	Un paquete basado en R creado por Templ, M., Kowarik, A. y Meindl, B. con herramientas para la anonimización de microdatos, es decir, para la creación de archivos de uso público y científico con cierto estándar de anonimato en las observaciones.	(Benschop, Machingauta, & Welch, 2021, pág. 11)
Supresión de datos	La supresión de datos implica no divulgar información que se considera insegura porque no se aplican las reglas de confidencialidad. A veces esto se hace reemplazando valores que significan atributos individuales con valores faltantes (por ejemplo, pasando del nivel de ingresos de un hogar a un "*missing*" o "sin dato" para proteger la identidad del hogar). En el contexto de esta guía, generalmente para lograr el nivel deseado de *k* – anonimato.	(Benschop, Machingauta, & Welch, 2021, pág. 11)

Término	Definición	Referencia
Tabulados	Expresión gráfica que sintetiza un valor o estimación producto del cruce entre dos o más variables.	(INE, 2020, pág. 60)
Técnicas de control de divulgación estadística	Se pueden definir como el conjunto de métodos para reducir el riesgo de divulgar información sobre personas, empresas u otras organizaciones. Dichos métodos solo están relacionados con el paso de difusión y generalmente se basan en restringir la cantidad o modificar los datos publicados.	(OCDE, s.f.)
Umbral de riesgo	Nivel, valor, margen o punto establecido a partir del cual se produce la identificación de unidades. Si no es seguro, se deberán tomar medidas adicionales para reducir el riesgo de identificación.	(Benschop, Machingauta, & Welch, 2021, pág. 11)
Unidades de observación	Unidad identificable sobre la que se obtiene información (o son informados), registran y compilan datos estadísticos.	(INE, 2021)[9]
Usuario final	El usuario del archivo de microdatos liberado después de la anonimización.	Adaptado de (Benschop, Machingauta, & Welch, 2021, pág. 9)
Utilidad de los datos	Describe el valor de una publicación de datos determinada como recurso analítico. Esto comprende la integridad analítica de los datos y su validez analítica. Los métodos de control de divulgación suelen tener un efecto adverso en la utilidad de los datos. Idealmente, el objetivo de cualquier régimen de control de divulgación debería ser maximizar la utilidad de los datos al tiempo que se minimiza el riesgo de divulgación. En la práctica, las decisiones de control de divulgación son una compensación entre la utilidad y el riesgo de divulgación.	(OCDE, s.f.)
Valor atípico	Un valor inusual que se informa correctamente pero que no es típico del resto de la población. Los valores atípicos (*outliers*, en inglés) también pueden ser observaciones con una combinación inusual de valores para variables, como la viuda de 20 años. En su propia edad, 20 y viuda no son	(Benschop, Machingauta, & Welch, 2021, pág. 10)

Término	Definición	Referencia
	valores inusuales, pero su combinación puede serlo	
Variable	Cualquier característica, número o cantidad que se puede medir o contar para cada unidad de observación.	(Benschop, Machingauta, & Welch, 2021, pág. 11)
Variable categórica	Una variable discreta que toma valores sobre un conjunto finito, por ejemplo, sexo representado por los números 1 o 0 para hombre y mujer. También llamado factor en R.	(Benschop, Machingauta, & Welch, 2021, pág. 8)
Variable continua	Una variable que puede tomar valores sobre un conjunto denso. Ejemplos son los ingresos, la altura del cuerpo y el tamaño de la parcela.	Adaptado de (Benschop, Machingauta, & Welch, 2021, pág. 8)
Variables de no identificación	Son variables que no pueden utilizarse para la re-identificación de los informantes o fuentes. Esto podría deberse a que estas variables no están contenidas en ningún otro archivo de datos u otra fuente externa. Estas variables son importantes en el procedimiento del control a la divulgación, ya que pueden contener variables sensibles.	(Benschop, Machingauta, & Welch, 2021, pág. 24)
Variable factor	Son una forma de clasificar variables categóricas en factores, que pueden ser ordenadas o no.	(Benschop, Machingauta, & Welch, 2021, pág. 9)
Variable semicontinua (discreta)	Es una variable que toma valores contenidos en un conjunto discreto. Un ejemplo es la edad medida en años, que podría tomar valores en el conjunto {0, 1, . . ., 100}. La naturaleza finita de los valores para estas variables significa que pueden tratarse como variables categóricas a los efectos de SDC.	Adaptado de (Benschop, Machingauta, & Welch, 2021, pág. 24)
Variable sensible	Variable contenida en un registro de datos, además de las variables clave, que pertenecen al dominio privado de los informantes que no quisieran que se divulgaran. Algunos datos son claramente sensibles, como la posesión de antecedentes penales o la condición médica, pero hay otros casos en los que la distinción depende de las circunstancias, por ejemplo, los ingresos de una persona pueden considerarse como una variable sensible en algunos países.	(OCDE, s.f.)

Término	Definición	Referencia
	La determinación de variables sensibles a menudo está sujeta a preocupaciones legales y éticas.	

Fuente: Instituto Nacional de Estadísticas (INE).

3. CONTROL DE DIVULGACIÓN ESTADÍSTICA: UNA INTRODUCCIÓN

3.1. Necesidad por control de divulgación estadística (proceso SDC)

La protección de la confidencialidad ha sido una preocupación de las Oficinas Nacionales de Estadísticas (ONE), lo que ha sido foco de atención recientemente, esto debido a que en las últimas décadas se ha experimentado un avance tecnológico importante, junto con el desarrollo de técnicas de re-identificación, por ejemplo, basado en *machine learning*. Por lo tanto, proteger los datos personales de los informantes y resguardar la vida personal se hace un imperativo (Yazdani, 2015). Por esta razón, hoy en día, resolver la tensión entre la protección de la información personal y el suministro de datos es realmente un desafío que deben asumir las ONE. En esta situación, tres motivaciones empujan a las ONE a preservar la confidencialidad.

El primer motivo para mantener la confidencialidad proviene del cumplimiento del marco normativo entre los cuales se establecen las funciones de la ONE. Existe una obligación legal y ética de los productores para garantizar que los datos proporcionados por los informantes se utilicen únicamente con fines estadísticos. La ONE debe respetar la confianza de los informantes, cuidar su privacidad y mantenerlos alejados de cualquier daño que pueda surgir de la información que han proporcionado. La ONE debe velar por resguardad el cumplimiento del marco normativo y las normas éticas.

El segundo motivo subyace en el deseo de la ONE de obtener la cooperación de los informantes y obtener datos más precisos. Los informantes que confían que su información permanecerá confidencial tienen más probabilidades de participar en la encuesta y reportar con precisión su información privada. Cualquier duda sobre la confidencialidad puede reducir la disposición de los posibles informantes a cooperar en una encuesta y puede afectar la calidad de las respuestas (Yazdani, 2015).

El último motivo es la obligación impuesta a la ONE por la legislación vigente, así como por compromisos internacionales. La fuerza de la sociedad sobre los gobiernos ha llevado al establecimiento de entornos legales para salvaguardar la privacidad y la ONE está mandada a respetar estas restricciones legales (Duncan, Elliot, & Salzar-González, 2011). Además, como lo aprobó por unanimidad la Asamblea General de las Naciones Unidas en enero de 2014, el principio 6 de los Principios Fundamentales de las Estadísticas Oficiales postula que "Los datos individuales que reúnan los organismos de estadística para la compilación estadística, se refieran a personas naturales o jurídicas, deben ser estrictamente confidenciales y utilizarse exclusivamente para fines estadísticos".

Los motivos señalados anteriormente son de naturaleza moral, ética y legal. El proceso SDC busca tratar y procesar los datos individuales para que cumplan el marco normativo y así, puedan publicarse o difundirse respetando el secreto estadístico, pero al mismo tiempo, controlar la pérdida de información debido al tratamiento de los datos.

El objetivo de anonimizar los microdatos es transformar los conjuntos de datos para lograr un "nivel aceptable" de riesgo de divulgación. El nivel de aceptabilidad del riesgo de divulgación y la necesidad de anonimización generalmente quedan a discreción del productor de datos y guiado por la legislación. Estos se formulan en las políticas y programas de difusión de los proveedores de datos y se basan en consideraciones que incluyen "[. . .] los costos y la experiencia involucrados; cuestiones de calidad de los datos, posible uso indebido y malentendidos de los datos por parte de los usuarios; asuntos legales y éticos; y mantener la confianza y el apoyo de los encuestados " (Benschop, Machingauta, & Welch, 2021, pág. 13).

3.2. Trade-off riesgo-utilidad en el proceso SDC

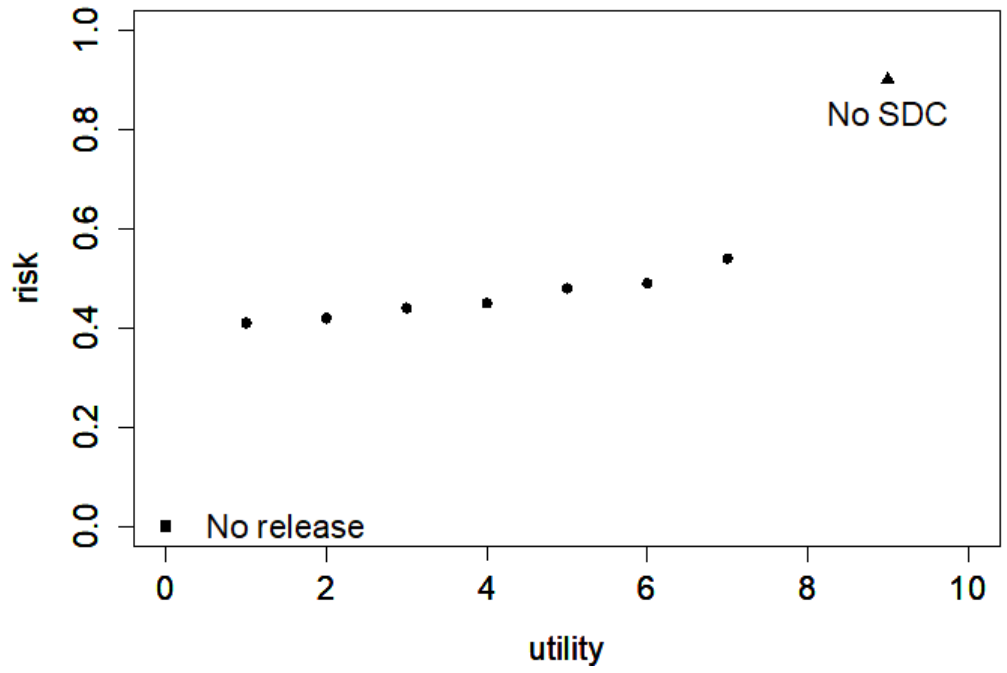
Por otra parte, el proceso SDC se caracteriza por el balance entre el riesgo de divulgación y la utilidad de los datos para los usuarios finales. La escala riesgo-utilidad se extiende entre dos extremos:

i. No se difunden datos (riesgo cero de divulgación) y, por lo tanto, los usuarios no obtienen ninguna utilidad de los datos,
ii. Los datos se difunden sin ningún tratamiento y, por lo tanto, con el máximo riesgo de divulgación, pero con la máxima utilidad para el usuario (es decir, sin pérdida de información).

El objetivo de un proceso SDC bien implementado es encontrar el punto óptimo en el que la utilidad para los usuarios finales se maximice a un nivel de riesgo aceptable.

En el balance entre Riesgo y Utilidad que se muestra en La **Figura 1**, por un extremo, el triángulo corresponde a los datos sin procesar, los que no tienen pérdida de información, pero generalmente tienen un riesgo de divulgación más alto que el nivel aceptable. El otro extremo es el cuadrado, que corresponde a la no publicación de datos. En ese caso, no hay riesgo de divulgación, pero tampoco hay utilidad de los datos para los usuarios. Los puntos intermedios corresponden a diferentes opciones de métodos SDC y/o parámetros aplicados a diferentes variables. El proceso SDC busca métodos y parámetros, que son aplicados de una manera que produce una reducción del riesgo de forma muchas veces satisfactoria, minimizándose generalmente la pérdida de información.

Figura 1: Balance Riesgo – Utilidad



Fuente: Imagen extraída de (Benschop, Machingauta, & Welch, 2021, pág. 15).

El proceso SDC no puede lograr la eliminación total del riesgo, pero puede reducir el riesgo a un nivel aceptable. Cualquier aplicación de métodos SDC suprimirá o alterará los valores en los datos y, como tal, disminuirá la utilidad (es decir, dará como resultado una pérdida de información) en comparación con los datos originales. Un hilo común que se enfatizará a lo largo de esta guía será que el proceso SDC debe priorizar el objetivo de proteger a los informantes y, al mismo tiempo, tener en cuenta a los usuarios de datos para limitar la pérdida de información. En general, cuanto menor es el riesgo de divulgación, mayor es la pérdida de información y menor es la utilidad de datos para los usuarios finales.

En la práctica, la elección de métodos SDC es un proceso iterativo: después de aplicar los métodos, el riesgo de divulgación y la utilidad de datos se vuelven a medir y se comparan con los resultados de otros métodos SDC y parámetros aplicados. Si el resultado es satisfactorio, los datos pueden ser liberados. Como se verá más adelante, a menudo el primer intento no será el óptimo. El riesgo puede no ser reducido lo suficiente o la pérdida de información puede ser demasiado alta y el proceso debe repetirse con diferentes métodos o parámetros hasta que se encuentre una solución satisfactoria. El riesgo de divulgación, la utilidad de los datos y la pérdida de información en el contexto de proceso SDC y cómo medirlos se analizan en capítulos posteriores de esta guía (ver **Etapa 6.4.4.2: Evaluar proceso SDC)**.

Nuevamente, debe enfatizarse que el nivel de SDC y los métodos aplicados dependen en gran medida de todo el marco de publicación de datos. Por ejemplo, una consideración clave es a quién y bajo qué condiciones se liberarán los datos (ver **Liberación de microdatos**). Si los datos se van a difundir como datos de uso público, entonces el nivel de SDC aplicado solo tendrá que ser mayor que en los casos en que los

datos se difundan bajo condiciones de licencia a usuarios confiables, después de un examen cuidadoso. Se discutirá cómo se podría lograr esto más adelante en la guía.
Esto ha dispuesto que entidades internacionales desarrollen diferentes técnicas de anonimización, que se ajustan a diferentes tipos de datos, consiguiendo de mejor manera resguardar la calidad de ellos.
El INE, igualmente deberá tener en cuenta este balance al publicar sus datos, velando porque se ponga a disposición de la ciudadanía información de la mayor calidad posible, cumpliendo el marco normativo relativo a la protección de datos, manteniendo así la confianza de los informantes.

4. DOCUMENTOS APLICABLES

A continuación, se especifican los documentos de referencia para el óptimo desarrollo del subproceso, tales como: documentación internacional, manuales, instructivos, políticas o procedimientos, entre otros.

Tabla 2: Documentos aplicables

Tipo	Nombre	Referencia
Estándar internacional	Generic Statistical Business Process Model (GSBPM), versión 5.1, enero 2019, UNECE.	Apartado fase "Análisis de resultados".
Estándar internacional	Código Regional de Buenas Prácticas en Estadísticas para América Latina y el Caribe, elaborado por CEPAL, versión 2011.	Documento completo.
Estándar internacional	Quality Indicators for the Generic Statistical Business Process Model (GSBPM) - For Statistics derived from Surveys and Administrative Data Sources, version 2.0, 2017.	Apartado II: Analyse Phase (Unece, 2017).
Estándar Nacional	Código de Buenas Prácticas para las Estadísticas Chilenas, elaborado por el INE, versión 2015.	Documento completo (2015).
Estándar Nacional	Manual de Implementación Norma de Documentación y Gestión de Metadatos (NDGM).	Documento completo (2020).
Resolución[9]	Manual de procedimiento subproceso "Aplicar control a la divulgación"[10]	Documento completo (2021).
Resolución[10]	Manual de procedimiento subproceso "Diseño y planificación"	Documento completo (2021).

Fuente: Instituto Nacional de Estadísticas (INE).

[9] Pendiente el número de resolución asociado al manual del proceso 6.4 y 2.

[10] Para ver los indicadores, revisar manual del subproceso 6.4.

5. CONTROL NORMATIVO

Es primordial la existencia de leyes y otros instrumentos normativos y técnicos que establezcan principios básicos de acceso a la información y a la protección de la confidencialidad. A continuación, se muestra el marco legal relacionado de manera directa o indirecta con el subproceso 6.4 "Aplicar control a la divulgación", así como leyes, resoluciones o decretos que deben ser considerados para su desarrollo.

- Ley N°17.374, de 10 de diciembre de 1970, del entonces Ministerio de Economía, Fomento y Reconstrucción, hoy Ministerio de Economía, Fomento y Turismo, que fija nuevo texto refundido, coordinado y actualizado del DFL. N° 313 de 1960, que aprueba la Ley Orgánica Dirección Estadística y Censos y crea el Instituto Nacional de Estadísticas. En especial, el artículo 29, lo que se refiere a que el INE y todos los organismos públicos y cada uno de sus respectivos funcionarios, no podrán divulgar los hechos que se refieren a personas o entidades determinadas de que hayan tomado conocimiento, lo que constituye el "Secreto Estadístico". También el artículo 30, sobre la prohibición de publicar o difundir datos estadísticos con referencia a las personas o entidades a quienes directa o indirectamente se refieran.
- Ley N° 19.628, de 28 de agosto de 1999, Sobre protección de la vida privada.
- Ley N° 19.653, de 14 de diciembre de 1999, Sobre probidad administrativa aplicable a los órganos de la Administración del Estado.
- Ley N° 20.285, de 20 de agosto de 2008, Sobre Acceso a la Información Pública.
- Ley N° 18.840, de 04 de octubre de 1989, Ley Orgánica Constitucional del Banco Central de Chile.
- Decreto N°1.062, de 17 de diciembre de 1970, del entonces Ministerio de Economía, Fomento y Reconstrucción, hoy Ministerio de Economía, Fomento y Turismo, que aprueba reglamento del Instituto Nacional de Estadísticas.
- Decreto N° 305, de 17 de febrero de 2010, del entonces Ministerio de Economía Fomento y Reconstrucción, hoy Ministerio de Economía, Fomento y Turismo, que ordena la incorporación de la variable sexo en la producción de estadísticas y generación de registros administrativos[11].
- Resolución Exenta N° 5.393, del 30 de diciembre de 2011, del Instituto Nacional de Estadísticas, y sus modificaciones, que Delega Facultades en Niveles Jerárquicos Señalados.
- Resolución Exenta N° 1.753, de 03 de junio de 2019, que Aprueba nueva estructura orgánica del INE y deja sin efecto Resoluciones Exentas N° 1.188, 3.676, 3.967 y 4.402, todas de 2018.
- Resolución Exenta N° 1.098, del 24 de junio de 2021, del INE, que Modifica Resolución Exenta 1.753, de 2019, en el sentido que indica.
- Resolución Exenta N° 917 de 21 de marzo de 2018, que aprueba convenio marco de colaboración para la gestión de información estadística entre el Banco Central de Chile y el Instituto Nacional de Estadísticas, y sus anexos N°1 y N°2 que se indican.
- Resolución Exenta N° 1649, de 28 de mayo de 2020, que aprueba modificación al convenio marco de colaboración gestión de información estadística, suscrito entre el Banco Central de Chile y el Instituto Nacional de Estadísticas.

[11] Para mayor información revisar el decreto N°305 dispuesto en la biblioteca del Congreso Nacional de Chile disponible en el siguiente link: https://www.bcn.cl/leychile/navegar?idNorma=1011115

6. LIBERACIÓN DE MICRODATOS

Esta sección expone sobre la liberación de microdatos, cuyos lineamientos se extrajeron de la guía elaborada por el Banco Mundial (Benschop, Machingauta, & Welch, 2021), que a su vez recoge el trabajo conjunto realizado por el Banco Mundial y sus socios en la Red Internacional de Encuestas de Hogares IHSN12 (Dupriez & Boyko, 2010).

El balance entre riesgo y utilidad en el proceso SDC depende en gran medida de quiénes son los usuarios y bajo qué condiciones se difunde o libera un archivo de microdatos.

En general, se practican tres tipos de métodos de liberación de datos para diferentes grupos objetivo, a saber: archivo de uso público (PUF), archivo de uso científico (SUF) y enclave de datos. En la **Tabla 3** se resumen los tipos de liberación y su aplicabilidad en el INE, dado el marco legal vigente en Chile. Como se podrá observar, **el tipo PUF es el único tipo de liberación de microdatos que es aplicable para el INE** dado el marco legal vigente en Chile.

[12] En inglés, *International Household Survey Network*.

Tabla 3: Resumen de tipos de liberación de microdatos

Tipo	Descripción	Aplicabilidad de acuerdo con el marco legal vigente
Archivo de Uso Público (PUF)	Los datos están disponibles directamente para cualquier persona interesada, por ejemplo, en el sitio web del INE. Estos datos se hacen fácilmente accesibles debido a que los riesgos de identificar a las unidades individuales se consideran mínimos. En el contexto INE, el PUF se puede entregar a nivel de microdatos mediante las siguientes formas: i. Base de datos publicadas (BP) que se dispone en la página web del INE y en la página web de la institución demandante, según corresponda. ii. Base de datos a solicitar por transparencia (BST) que se entrega directamente al usuario responsable de la solicitud.	Aplicable.
Archivo de Uso Científico (SUF)	La difusión está restringida a los usuarios que han recibido autorización para acceder a ellos después de enviar una solicitud documentada y firmar un acuerdo que rige el uso de los datos. Si bien los archivos con licencia general también se anonimizan para garantizar que el riesgo de identificar a las unidades (personas, hogares o establecimientos) se minimice cuando se usan de forma aislada, aún pueden (potencialmente) contener datos identificables si se vinculan con otros archivos de datos. Este tipo de liberación de datos también es conocido como archivo con licencia, microdatos bajo contrato o archivo de investigación.	No aplicable.
Enclave de datos o centro de datos de investigación controlado	Algunos archivos pueden ofrecerse a los usuarios bajo condiciones estrictas en un enclave de datos. Esta es una instalación (puede ser una instalación al interior del INE) equipada con computadoras que no están conectadas a Internet o una red externa y desde las cuales no se puede descargar información a través de puertos USB, CD – DVD u otras unidades. Los enclaves de datos contienen datos que son particularmente sensibles o permiten la identificación directa o fácil de los informantes.	No aplicable.

Fuente: Instituto Nacional de Estadísticas (INE).

6.1. Condiciones para la liberación de datos bajo versión PUF

En general, los datos que se consideran públicos están abiertos a cualquier persona con acceso al sitio web del INE. Sin embargo, es una buena práctica incluir declaraciones de principios que definan los usos adecuados y las precauciones que se adoptarán utilizando los datos. Si bien estos pueden no ser legalmente vinculantes, sirven para sensibilizar al usuario. Prohibiciones como intentos de vincular los datos a otras fuentes puede ser parte de la "declaración de uso", requerida para el uso de datos. La difusión de archivos de microdatos implica necesariamente la aplicación de reglas o principios.

A continuación, se listan principios básicos o “declaraciones de uso” aplicables a una liberación PUF:

1. Los datos y otros materiales proporcionados por el INE no serán redistribuidos o vendidos a otras personas, instituciones u organizaciones sin el acuerdo por escrito del INE.
2. Los datos se usarán solo para fines de investigación estadística y científica. Serán empleados únicamente para reportar información agregada, incluido el modelado, y no para investigar individuos u organizaciones específicos.
3. No se intentará volver a identificar a los informantes, y no se usará la identidad de ninguna persona o establecimiento descubierto inadvertidamente. Cualquier descubrimiento de este tipo se informará inmediatamente al INE.
4. No se intentará crear enlaces entre conjuntos de datos proporcionados por el INE o entre datos del INE y otros conjuntos de datos que podrían identificar individuos u organizaciones.
5. Libros, artículos, documentos de conferencias, tesis, disertaciones, informes u otras publicaciones que empleen datos obtenidos del INE citará la fuente, de acuerdo con el requisito de cita provisto con el conjunto de datos, en caso de no haber sido proporcionado, se debe citar de acuerdo a la norma APA más actualizada.
6. Se enviará al INE una copia electrónica de todas las publicaciones basadas en los datos descargados.
7. El recolector original de los datos, el INE y las agencias de financiamiento relevantes no tienen responsabilidad por el uso o interpretación de los datos o inferencias basadas en ellos.

Nota: Los puntos 3 y 6 de la lista requieren que los usuarios reciban una manera fácil de comunicarse con el INE. Es una buena práctica proporcionar un número de contacto, una dirección de correo electrónico y, posiblemente, un sistema de "suministro de comentarios" en línea.

7. LISTA DE REGISTROS

El listado que se presenta en este apartado se relaciona con los medios de verificación que contribuyen y facilitan el desarrollo del subproceso.

Tabla 4: Lista de registros

Registro	Sigla	Descripción
Reporte de medición y evaluación de los riesgos de divulgación	No aplica	Documento que contiene los cálculos de los riesgos de divulgación de la operación estadística, según los lineamientos establecidos en el proceso 2. "Diseño y planificación". Este informe es un insumo para el subproceso 6.4.4.1 "seleccionar y aplicar métodos SDC".
script. R	No aplica	Documento de texto con la extensión de archivo. R que contiene una secuencia de comandos o códigos que se deben leer y ejecutar en R para realizar los cálculos de los riesgos y, posteriormente, aplicar el método SDC, según corresponda.
Reportes del subproceso aplicar control a la divulgación	No aplica	Documentos que contienen la descripción del proceso SDC desarrollado: la descripción de las características estadísticas priorizadas, los alcances de los resultados obtenidos a partir de la base anonimizadas y la descripción de los métodos SDC empleados, entre otros.

Fuente: Instituto Nacional de Estadísticas (INE).

8. DESCRIPCIÓN DEL SUBPROCESO

Esta sección describe el procedimiento del subproceso 6.4 “Aplicar control a la divulgación”, centrado en una etapa: control a la divulgación. En esta etapa se describen las actividades, roles y cargos responsables, sistemas tecnológicos posibles de utilizar, potenciales contingencias y los mecanismos de control que se utilizan durante la ejecución de cada actividad. Este subproceso está ubicado dentro del proceso denominado 6. “Análisis de resultados” del mapa de procesos institucional (ver anexo **10.1. Mapa de procesos-Segmento de Negocio**), adaptado desde el Modelo Genérico del Proceso Estadístico - GSBPM (Versión 5.1 enero 2019 y anteriores).

Este subproceso tiene por objetivo garantizar que los datos y metadatos que se difunden cumplan el marco normativo de confidencialidad, de acuerdo con las políticas y normas del INE, o con la metodología específicamente diseñada en el subproceso 2.5 “Diseñar el procesamiento y análisis”, según los lineamientos establecidos para cada operación estadística.

8.1. Cuadro de Roles

A continuación, se detallan los cargos involucrados en el subproceso, de acuerdo con la familia de cargos dispuesta por la institución, señalando la descripción del rol dentro del flujo. Para el registro de familia de cargo (ver el anexo **10.3. Familias de cargo**).

Tabla 5: Cuadro de roles

Cargo	Familia de cargo	Descripción del Rol
Analista (socioeconómico/ económico/ estadístico/ metodólogo/calidad) / Analista Geoespacial / Gestor(a) Técnico Análisis de Datos / Gestor(a) Técnico de Proyectos Geoespaciales / Coordinador(a) / Jefatura de unidad	Analistas especialistas/ Coordinadores(as).	- Pertenecen a esta familia de cargo, lo que más adelante en el documento serán referidos como analista temático y analista de anonimización. - Es (son) la(s) persona(s) responsable(s) de revisar los insumos provenientes de los subprocesos 2. "Diseño y Planificación", 5. "Procesamiento" y 6.3 "Interpretar y explicar los resultados", ejecutar las actividades para aplicar el procedimiento que busca garantizar que los productos estadísticos que se difunden no infrinjan las normas de confidencialidad, de acuerdo con las políticas y normas del INE. - Entre las principales competencias de estos roles destacan: • Conocimiento temático del contenido de los productos estadísticos a anonimizar. Esto es clave para establecer escenarios de divulgación. • Manejo de herramientas que permitan realizar los análisis exploratorios de los datos para ejecutar correctamente las actividades que buscan preparar y explorar los datos, así como el cálculo de los riesgos de divulgación. • Conocimiento técnico sobre los métodos SDC (anonimización) que le permita seleccionar apropiadamente los métodos más adecuados según las características de los datos y tipos de variables. Además de conocimientos para la implementación de los métodos y su evaluación.

Cargo	Familia de cargo	Descripción del Rol
Analista Geoespacial / Gestor(a) Técnico Análisis de Datos / Gestor(a) Técnico de Proyectos Geoespaciales [13]/ Supervisor / Coordinador(a)/ Jefatura de subdepartamento/ Jefatura de departamento[14].	Analistas especialistas / Coordinadores(as)/ Jefatura de subdepartamento/ Jefatura de departamento.	- Pertenecen a esta familia de cargo, lo que más adelante en el documento serán referidos como analista temático y jefe de área o proyecto. - Es la persona encargada de realizar los diferentes controles, actividades o procedimientos de calidad desde el inicio del subproceso 6.4 "Aplicar control a la divulgación" hasta su término, establecidos en el proceso de 2. "Diseño y planificación", para validar los resultados generados del subproceso, según corresponda a las características de cada operación estadística. Es el responsable por validar y asegurar la correcta aplicación del proceso SDC. - Este rol de revisor debe ser ejecutado por un cargo distinto al responsable de ejecutar el control a la divulgación.

Fuente: Instituto Nacional de Estadísticas (INE).

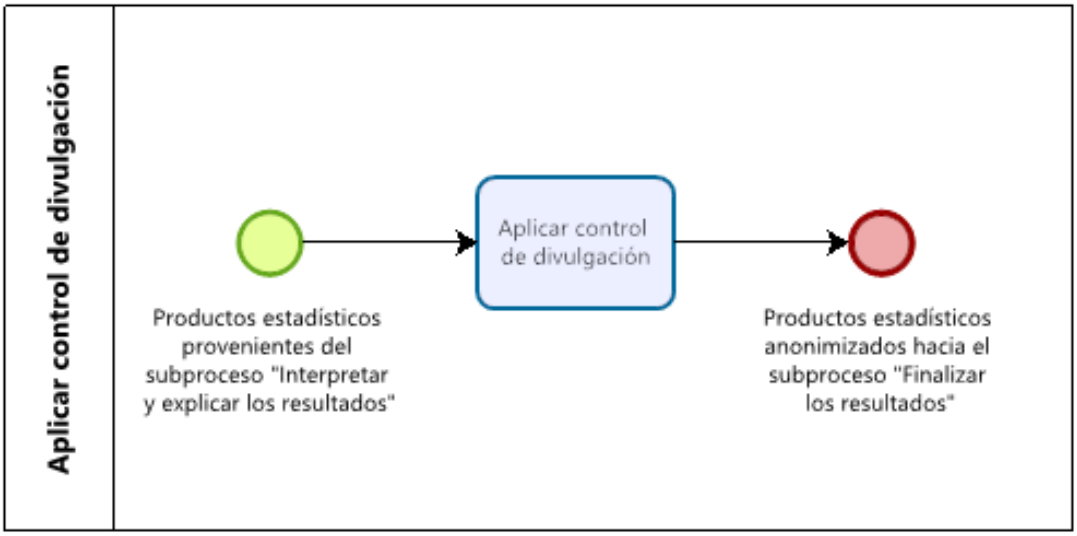
[13] Por la naturaleza y organización del trabajo, en el proceso de control a la divulgación estadística de microdatos que acompañan la publicación de datos georreferenciados (a cargo del Subdepartamento de Geografía), los roles de analista de anonimización y analista temático, pueden ser desempeñados por profesionales con los siguientes cargos: Analista Geoespacial / Gestor(a) Técnico Análisis de Datos / Gestor(a) Técnico de Proyectos Geoespaciales.

[14] Dependiendo de la temática o foco de estudio de la operación estadística, será posible que en algunos casos la ejecución de la(s) actividad(es) asociados al cargo, podrían ejecutarse por la jefatura de departamento o jefatura de subdepartamento según corresponda a la organización interna de cada equipo al interior de la Subdirección Técnica.

8.2. Diagrama de etapas

En este apartado se describen cada etapa, su objetivo y su alcance dentro del subproceso 6.4 “Aplicar control a la divulgación”.

Figura 2: Etapa subproceso 6.4. Aplicar control a la divulgación



Fuente: Instituto Nacional de Estadísticas (INE).

Tabla 6: Descripción de las etapas

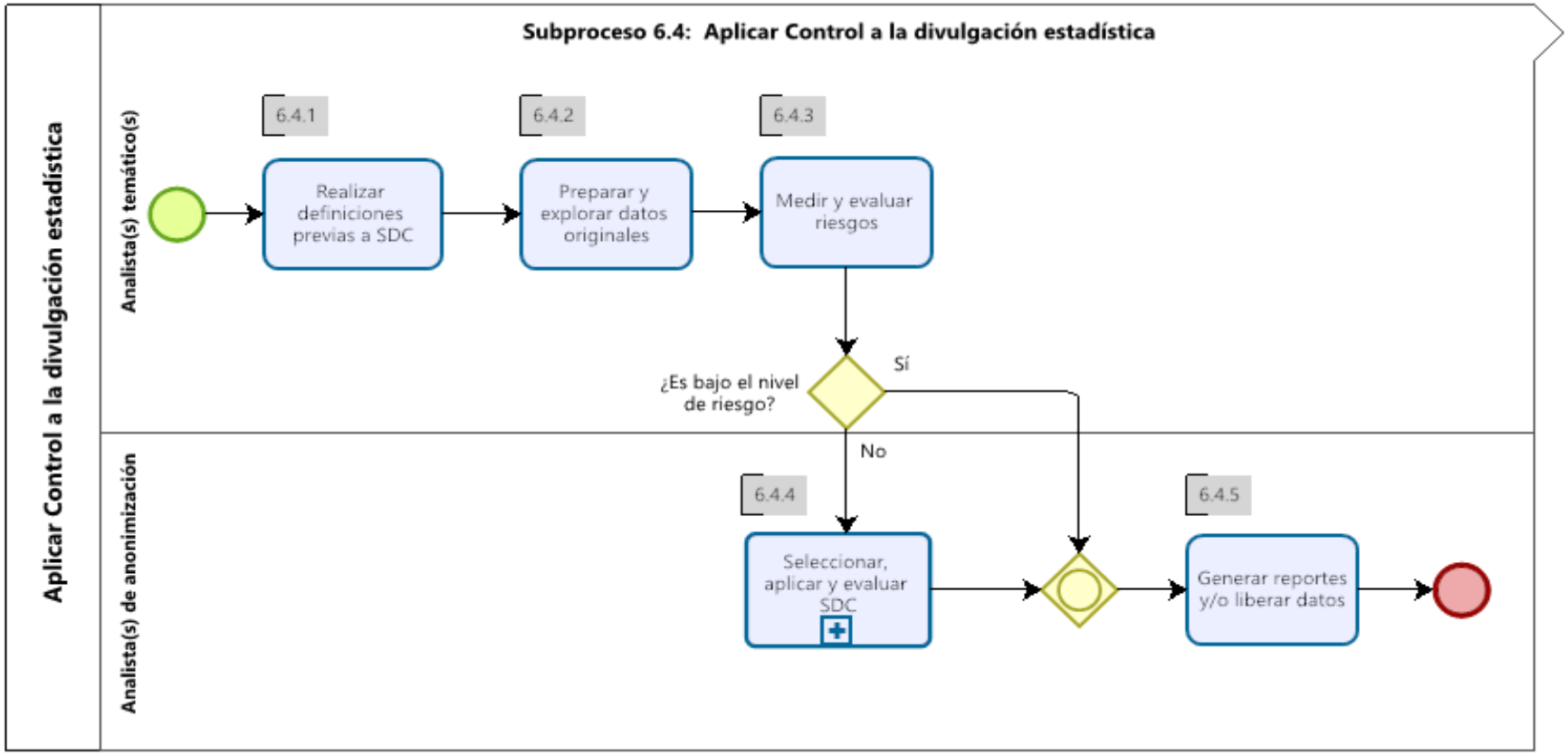
Etapa	Objetivo	Alcance
Aplicar control a la divulgación	Garantizar que los productos estadísticos que se difunden no infrinjan las normas de confidencialidad, de acuerdo con las políticas y normas del INE, en términos de resguardar el secreto estadístico y estándares de calidad establecidos en el proceso de 2. “Diseño y planificación”.	Desde el subproceso 6.4.1 “Realizar definiciones previas al proceso de anonimización” con la recepción y revisión de productos estadísticos provenientes del subproceso 6.3 “Interpretar y explicar los resultados”, hasta el subproceso 6.4.5 “Generar reportes y liberar datos”, con la disposición de los productos estadísticos anonimizados, con el reporte de anonimización y metadatos actualizados correspondientes.

Fuente: Instituto Nacional de Estadísticas (INE).

8.3. Aplicar control a la divulgación

Esta etapa contempla el desarrollo de actividades orientadas a revisar y garantizar que los productos estadísticos que se difunden no infrinjan las normas de confidencialidad, de acuerdo con las políticas y normas del INE.

Figura 3: Esquema general de Aplicar control a la divulgación



Fuente: Instituto Nacional de Estadísticas (INE).

El subproceso de control a la divulgación de microdatos se encuentra compuesto por las cinco etapas en la **Figura 3**. Sin embargo, debe tenerse en cuenta que a menudo se requiere saltar entre los pasos y volver a los pasos anteriores durante el proceso SDC real, ya que no es necesariamente un proceso lineal paso a paso, pues debe entenderse como un proceso iterativo.

Notas:

(1) Si el conjunto de datos posee una estructura jerárquica, los pasos 6.4.3 y 6.4.4 en la **Figura 3** se deben repetir si los identificadores indirectos se encuentran en los diferentes niveles jerárquicos, por ejemplo, hogar y personas. En ese caso, el proceso de anonimización es como sigue:

1. Las variables en el nivel jerárquico superior (hogar) deben anonimizarse primero y luego fusionarse con las variables no tratadas de nivel inferior (personas).
2. Posteriormente, el conjunto de datos combinado debe anonimizarse.

Este enfoque garantiza la coherencia en los datos tratados. Si se descuida este procedimiento, los valores de las variables medidas en el nivel jerárquico superior podrían tratarse de manera diferente para las observaciones de la misma unidad. Por ejemplo, la variable "región" es la misma para todos los miembros del hogar. Si el hogar lo conforman cinco personas y se suprimiera el valor "Valparaíso" para dos miembros, pero no para los tres restantes, se produciría una divulgación no intencionada; con el ID del hogar, la región de la variable sería fácil de reconstruir para los dos valores suprimidos.

(2) El proceso de anonimización es un proceso iterativo donde los pasos pueden revisarse, mientras que la publicación de los microdatos anonimizados es un proceso que se ejecuta solamente una vez para cada versión de la operación estadística. Por ejemplo, para la versión 2021 de la Encuesta Nacional Urbana de Seguridad Ciudadana (ENUSC), solo se debe generar y publicar un conjunto de datos anonimizados correspondientes a dicho periodo.

(3) Es posible que la aplicación del proceso obtenga como resultado la imposibilidad de publicar el conjunto de datos (Para más detalles, ver **Etapa 6.4.4.2: Evaluar proceso SDC**).

Por lo tanto, es imperativo que la lógica del proceso se respete a fin de obtener un subproceso de control a la divulgación de microdatos exitoso.

Cada una de las cinco etapas está compuesta por una serie de actividades que busca un desarrollo organizado de un procedimiento paso a paso, al mismo tiempo que permite optimizar la distribución de cargas de trabajo y evitar la duplicación de tareas.

A lo largo de cada etapa, siempre que sea posible, se introducen ejemplos o ejercicios para una mejor comprensión. Para estos fines, se ha utilizado la librería `sdcMicro` (Templ, Kowarik, & Meindl, 2015) del

entorno y lenguaje de programación con enfoque al análisis estadístico, R (R Core Team, 2019). Además, se realizan indicaciones y recomendaciones de documentación que permiten conocer la trazabilidad del proceso SDC dentro de la institución, al mismo tiempo que permite cimentar el conocimiento y transmitir la experiencia temática de los equipos.

8.3.1. Etapa 6.4.1: Realizar definiciones previas al proceso de anonimización

- **Objetivo:** El propósito de esta etapa es establecer los requerimientos necesarios para iniciar el subproceso de control a la divulgación, que incluye la revisión de insumos, revisión de las necesidades de los usuarios y características estadísticas prioritarias, y la determinación de la necesidad de protección de confidencialidad. Esto último, está estrechamente relacionado con la interpretación de las leyes y normas sobre este tema en Chile (ver **Control normativo**).

- **Alcance:** Los procedimientos descritos para esta etapa aplican para todas las operaciones estadísticas y productos relacionados cuyo levantamiento de información y/o publicación sea realizado por el INE (muestras, censos, procesos de múltiples fuentes y registros estadísticos) que darán a conocer información al público general u otros usuarios.
  En esta etapa se debe revisar que no existan restricciones legales que impidan la publicación de los microdatos.
  Por otra parte, si el conjunto de datos no posee variables sensibles o variables de identificación (directa o indirecta), se pasa a la **Etapa 6.4.5: Generar reportes y liberar datos**.

- **Exclusiones:** Las exclusiones generales se encuentran listadas en la sección **Exclusiones al alcance**. Además, se excluyen las operaciones estadísticas cuyos conjuntos de microdatos no son levantados ni publicados por el INE.

- **Palabras claves:** Necesidades de protección de la confidencialidad, variables sensibles, unidades estadísticas, y definiciones previas.

En la **Tabla 7** se resume responsabilidades, *inputs* y *outputs* relacionados con la etapa Realizar definiciones previas al proceso de anonimización.

Tabla 7: Responsables, *inputs* y *outputs* para la etapa Realizar definiciones previas al proceso de anonimización

Responsable(s)	Analista(s) temático y analista(s) de anonimización.
Input(s)	- Normativa institucional vigente "Política de Privacidad y Protección de la información de Identificación Personal". - Documentos del subproceso 2.5 "Diseñar el procesamiento y análisis", con la especificación del marco y diseño muestral, y creación de variables, además del procedimiento de control a la divulgación adoptado por la operación estadística. - Reporte de consultas y respuestas a solicitudes de información de usuarios u organización externa. - Convenios con fuentes externa, proveniente del subproceso 1.4 "Formalizar requerimiento". - Ficha metodológica, con el diseño temático, estadístico, operativo y TIC proveniente del proceso 2. "Diseño y planificación". - Manual de usuario de la base de datos. - Archivos de datos originales (entendido como el resultado del proceso 5. "Procesamiento"). - Archivos de datos originales de correspondientes a versiones anteriores de la operación estadística. - Base cartográfica (Para productos que impliquen información geográfica). - Diccionario y/o definición de variables proveniente del proceso 2. "Diseño y planificación". - Productos estadísticos, provenientes del subproceso 6.3 "Interpretar y explicar los resultados". - Infraestructura tecnológica y mecanismos de seguridad sobre la(s) base(s) de datos a anonimizar.
Output(s)	- Reporte sobre definiciones previas al subproceso de anonimización. - Base de datos final, objeto de anonimizar. - Diccionario y/o definición de variables proveniente del proceso 2. "Diseño y planificación". - Infraestructura tecnológica y mecanismos de seguridad definidos.
Control o supervisión	Jefe de área o jefe de proyecto.
Contingencias	Los productos estadísticos provenientes del subproceso 6.3 "Interpretar y explicar los resultados" que no cumplan con los criterios de coherencia y consistencia estipulado en el proceso 2. "Diseño y planificación", por lo tanto, deben regresar al proceso anterior con su respectivo reporte o minuta de observaciones. El responsable de la actividad debe enviar un correo al responsable de disponer de los resultados del subproceso 6.3 "Interpretar y explicar los resultados", adjunta las observaciones y plazos para subsanar las observaciones.

Fuente: Instituto Nacional de Estadísticas (INE).

Actividad 1-1. Organización del proceso y equipo de trabajo: El trabajo del proceso SDC y los roles del equipo a cargo se deben organizar considerando las siguientes dos fases:

1. **Diagnóstico:** Esta etapa se vincula con los procesos definidos en el GSBPM 2. "Diseño y planificación" y 6.3 "Interpretar y explicar los resultados". Esta etapa cubre los pasos 6.4.1 al 6.4.3 en la **Figura 3**. El trabajo destinado a esta etapa tiene como principales responsables a los **analistas temáticos**, pues se requiere conocimiento sobre el fenómeno medido, necesidades de información de los usuarios, propósito de la operación estadística, estimadores e indicadores, diseño muestral y representatividad de los datos, vinculación con datos anteriores y fuentes de información externa que puede ser utilizada por un intruso para la re-identificación de informantes, además de conocimientos sobre anonimización a nivel conceptual y manejo de procesamiento a nivel intermedio.
2. **Implementación:** Esta etapa tiene el objetivo de anonimizar las bases de datos y finalizar con la documentación del proceso (pasos 6.4.4 al 6.4.5 en la **Figura 3**). El trabajo destinado a esta etapa tiene como principales responsables a los **analistas de anonimización**, que requieren de los mismos conocimientos señalados para los analistas temáticos, estadística a nivel inferencial y manejo en el procesamiento de datos, con miras a mantener la trazabilidad de las transformaciones aplicadas a los datos y a resguardar que no se introduzca un error de procesamiento a través de estos procedimientos.

Notas:

1. La organización de los analistas a través de las fases no se debe entender como excluyente, pues en la fase de diagnóstico también tienen participación los analistas de anonimización; y en la de implementación también tienen participación los analistas temáticos, ya sea en el desarrollo de actividades específicas o bajo un rol de supervisión o control.
2. Adicional a los analistas temáticos y de anonimización, como se verá más adelante, algunas actividades y tareas requieren del control o supervisión del jefe de área o jefe de proyecto.

Actividad 1-2. Revisión de insumos o productos estadísticos necesarios que permiten la ejecución del proceso: Se debe revisar que los insumos o productos estadísticos provenientes de los procesos 2. "Diseño y planificación", 5. "Procesamiento" y 6.3 "Interpretar y explicar los resultados", necesarios para la ejecución de este subproceso, se encuentren completos y actualizados. Se incluyen entre estos insumos los *inputs* listados en la **Tabla 7**.

Si los insumos o productos estadísticos necesarios para la adecuada ejecución del proceso no cumplen con los requerimientos para su uso, se debe volver al subproceso 2.5 "Diseño de procesamiento y análisis", y subsecuentemente al 5. "Procesamiento" y 6.3 "Interpretar y explicar los resultados".

Notas:
En caso de encontrar información atingente a la operación estadística, se debe hacer minuta informativa que sirva de base para actualizar el subproceso 2.5 “Diseño de procesamiento y análisis” que contenga los ajustes realizados en el proceso.

Tarea 1-2.1. Revisión de bases de datos y diccionario de variables: La revisión de bases de datos y diccionario de variables debe abordar los siguientes puntos:

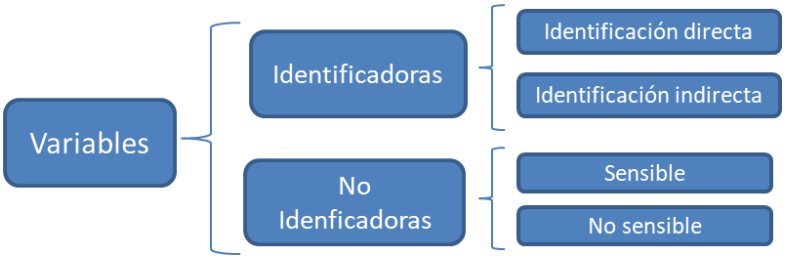
1. **Se debe asegurar que la base de datos esté en su versión final** procedente del proceso 5. “Procesamiento”, es decir, un archivo sin inconsistencias, depurado, etiquetado e informáticamente compatible con el lenguaje R.
2. Se debe contar con una descripción de la estructura de las bases de datos (dimensiones, clasificación de las variables, etc.), así como las relaciones entre variables, sobre todo, las más complejas para no inducir al error. En la **Tabla 8** se presenta un formato de tabla para realizar una descripción de las variables en el conjunto de datos.
3. Se debe asegurar que el diccionario de variables incluya la clasificación de todas las variables en el conjunto de datos. En la **Figura 4** se presenta una clasificación de las variables que debe ser considerado en el diccionario de variables. En complemento, en la **Tabla 9** se presenta un formato de tabla para el proceso de generación del “libro de códigos”, que incorpora tres dimensiones según identificación, sensibilidad y nivel de medida.

Tabla 8: Descripción de las variables en el conjunto de datos

Variable	Nombre	Tipo de variable	Categorías de respuesta (categóricas) y Min. y Máx. en cuantitativa.	Comprobación de saltos	Comentarios
Var1	Nombre de la variable	Definición a qué tipo de variable corresponde (Continua, Categórica, texto, etc.)	Completar con el recorrido de la variable	Señalar si la variable no es aplicable en relación con otras variables, es decir, si se incluye filtros o saltos. Por ejemplo, filtros de edad	Comentarios de la variable de ser necesarios

Fuente: Instituto Nacional de Estadísticas (INE).

Figura 4: Clasificación de las variables



Fuente: Instituto Nacional de Estadísticas (INE).

Tabla 9: Descripción de las variables en el conjunto de datos según nivel de medida, identificación y sensibilidad

Nombre variable	Etiqueta variable	Nivel de medida	Identificación	Variable sensible
nombre	Nombre informante	Categórica (abierta)	Directa	Sí
edad	Edad	Discreta	Indirecta o cuasi	No
ingreso_tp	Ingreso Trabajo principal	Continua	Indirecta o cuasi	Sí
religión	Religión	Categórica	No identificadora	Sí
antigüedad	Antigüedad laboral en años	Discreta	No identificadora	No

Fuente: Instituto Nacional de Estadísticas (INE).

Tarea 1-2.2. Revisión de paquetes estadísticos: Para efectos de la implementación del proceso SDC, el lenguaje R es el único que debe ser estandarizado. Temas por revisar:

1. El lenguaje R debe estar instalado en el servidor[15] donde procederá la anonimización de la base de datos. Tanto el *software* como las librerías necesarias para la implementación del subproceso deben estar actualizadas a la versión de esta guía para evitar errores en sus funciones o argumentos.
2. Los productores deben utilizar paquetes estables y validados por la comunidad R. Para efectos del proceso SDC, el paquete `sdcMicro` debe ser utilizado en la versión indicada.

[15] En los casos que no esté disponible el servidor se deberá proveer de un espacio seguro, ya que como INE se debe propender a espacios colaborativos y no el almacenamiento en equipos personales.

3. Adicionalmente, debe existir una instancia de validación de librerías. Para ello, los equipos deben documentar en el procesamiento y proceso de anonimización todas las librerías que se han utilizado y sus versiones.

Notas:

1. Algunas librerías recomendadas: `dplyr` (Wickham & Francois, 2015) y `tidyverse` (Wickham et al., 2019) para la manipulación y procesamiento de datos; `Matrix` (Bates et al., 2021) para trabajar de forma eficiente en memoria con matrices dispersas (es decir, con mayoría de registros 0) de alta dimensionalidad; `codebook` (Arslan, 2020) para extraer la metadata de una base de datos de manera automática; `Rmarkdown` (Allaire et al., 2021) para redactar informes.
2. El control de versiones de las librerías de R puede ser realizado mediante el uso de la librería `renv` (Ushey, 2021) con Docker[16].

Tarea 1-2.3. Revisión de infraestructura y seguridad: Se debe revisar la infraestructura para el manejo de datos, además de 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 SDC. Entre los temas relacionados con la seguridad, el equipo debe:

1. Indicar los requerimientos mínimos para soportar *software* y librerías definidos en la **Tarea 1-2.2**.
2. Salvaguardar los servidores en donde se aloja la información sensible, es debido a esto, que estos equipos deben contar con contraseñas o técnicas de cifrado necesarios para el correcto resguardo de esta información. Así mismo, habilitar contraseña a microdatos no anonimizados que circulen entre los equipos.
3. Evaluar el uso e implicancias de vulnerabilidad con el fin de mantener estándares válidos para la seguridad de la información en servidores al utilizar grandes volúmenes de datos.
4. Considerar además el otorgamiento de permisos de descarga y actualizaciones de paquetes en R (configuración del cortafuego).
5. Se requiere contar con soporte especializado en R desde el equipo de TI.

[16] Para más información, ver https://cran.r-project.org/web/packages/renv/vignettes/docker.html

Notas:

1. Se recomienda en los casos en que se deban utilizar equipos personales[17], estos sean provistos por la institución a cargo y que cuenten con la configuración necesaria para llevar a cabo el proceso SDC de manera segura.
2. Se recomienda para la circulación de microdatos entre equipos del INE, exista un tratamiento de los datos. Solo pueden circular libremente aquellos microdatos publicados.
3. No existe un control de versión de los microdatos que circulan al interior de la institución. Una forma es alojar en un servidor los microdatos a compartir al interior de la institución, con una contraseña cambiante que se otorgue restrictivamente a quienes se les autorice su uso.

Actividad 1.3 Determinación de la necesidad de protección de la confidencialidad: Se debe determinar la necesidad de protección de la confidencialidad de las unidades estadísticas contenidas en el conjunto de datos, lo cual está estrechamente relacionado con el marco legal vigente en Chile en relación con la función del INE, además de aspectos éticos, morales y de protección de la vida privada de las personas.

Tarea 1-3.1. Revisión del marco normativo: Se debe realizar una revisión normativa que pueda afectar o impedir la publicación de la información sujeta a anonimizar, esto incluye:

1. Identificar las restricciones actuales de publicación de la información.
2. Identificar acuerdos y restricciones de las particularidades establecidas en los convenios del INE con fuentes externas.

Tarea 1-3.2. Determinar las unidades estadísticas en el conjunto de datos: Se deben determinar todas las unidades estadísticas presentes en el conjunto de datos. Si se trata de individuos, hogares o entidades jurídicas, como empresas, es probable que sea necesario controlar la divulgación.

Nota:
Todos los tipos de unidades estadísticas presentes en el conjunto de datos deben considerarse para la necesidad de control de divulgación. Esto es especialmente importante en caso de que los datos tengan una estructura jerárquica, como, por ejemplo, individuos en hogares (encuestas de hogares) o empleados en empresas (encuestas económicas).

Tarea 1-3.3. Determinar si el conjunto de datos tiene variables sensibles o variables de identificación: Se debe determinar si el conjunto de datos posee variables sensibles o de identificación. La definición de variables sensibles está supeditada al marco legal normativo, ético y de la protección de la vida privada de las personas.

[17] En situaciones de fuerza mayor, se aplica la orden y legislación institucional.

Notas:

1. También hay ejemplos de microdatos para los cuales, no existe la necesidad de control de divulgación. Por ejemplo, datos con observaciones climáticas y meteorológicas, o datos de sismología.
2. Incluso, si las unidades estadísticas primarias no son personas naturales o jurídicas, los datos pueden contener información confidencial sobre personas naturales o jurídicas. Por ejemplo, un conjunto de datos con viviendas como unidades estadísticas primarias, también puede contener información sobre las personas que viven en estas viviendas y sus ingresos, o un conjunto de datos sobre hospitalizaciones, estas pueden incluir información sobre los pacientes hospitalizados. En estos casos, es probable que aún sea necesaria la protección de la confidencialidad. Una opción para resolver esto es eliminar la información sobre las personas naturales y jurídicas en los conjuntos de datos para su publicación.

Si no existen variables sensibles o variables de identificación en el conjunto de datos, o restricciones desde el marco legal normativo, la decisión es pasar a la **Etapa 6.4.5: Generar reportes y liberar datos**, según las características establecidas en la metodología de la operación estadística o convenio institucional, según corresponda a lo establecido en el proceso 2. "Diseño y planificación". Por el contrario, en caso de que el conjunto de datos contenga variables sensibles o variables de identificación, y no haya restricciones desde el marco legal normativo, la decisión es llevar a cabo un proceso SDC.

Una vez que, el equipo de trabajo ha decidido llevar a cabo el proceso SDC, se debe continuar con la siguiente tarea **Etapa 6.4.1: Realizar definiciones previas al proceso de anonimización**:

Actividad 1-4. Definición de las características de las bases de datos a preservar en la base de datos: Se debe definir las características estadísticas a preservar en la base de datos. Esta actividad es crucial, pues es un insumo para evaluar la utilidad de los datos después de la anonimización y la producción de un conjunto de datos anonimizados, que es útil para los usuarios finales (ver **Etapa 6.4.4.2: Evaluar proceso SDC**). Esta actividad incluye las siguientes tareas:

1. Identificación de las necesidades de información que presentan los usuarios sobre la base de datos. Así, el equipo de trabajo podrá identificar las variables requeridas, los niveles de desagregación, los períodos de tiempo de la información de la base de datos que tiene mayor demanda, por consiguiente, le dará información para definir la base de datos y desagregaciones que satisfagan coherentemente las necesidades de los usuarios y que, además, no expongan la identificación de las unidades de observación. Las siguientes son algunas preguntas que se pueden considerar para esta tarea:
   i) ¿Cuál es el objetivo de la operación estadística?
   ii) ¿Qué tipos de usuarios solicitan información (academia y centros de investigación, ciudadanía, empresas, entidades públicas, etc.)? ¿Cuáles son las variables requeridas por los usuarios? ¿A

qué nivel de desagregación geográfica? ¿Con qué frecuencia es requerida la información? ¿Con qué objetivo o finalidad se requiere la información?

iii) ¿Existen otras operaciones estadísticas relacionadas con la temática de la base de datos?

2. Una vez identificadas las necesidades de información de los usuarios, el equipo de trabajo deberá establecer cuáles son las características estadísticas que se deben mantener en la base de datos anonimizada. Entre las características estadísticas a preservar destacan las siguientes:

   i) **Mantener características globales de las variables:** Se debe definir qué medidas estadísticas para las variables categóricas y cuantitativas se deben mantener sin variación y para qué niveles de desagregación geográfica o temática. Así mismo, se debe decidir cuáles de las características globales pueden presentar alguna variación significativa y hasta qué porcentaje de variación es permitido en la base de datos anonimizada. Por ejemplo, se decidió que la propiedad global que se desea mantener es el promedio de la variable "Ingreso por hogar". Además, se aceptará el proceso de anonimización, solamente si el promedio de la variable en la base de datos anonimizada difiere 1% o menos del promedio en la base de datos sin anonimizar.

   ii) **Mantener cifras por nivel de desagregación geográfica o temática:** Se debe definir qué medidas estadísticas se deben conservar sin variación en los niveles de desagregación geográfica o temática, para garantizar a los usuarios el análisis de estadísticas más sectorizados. Por ejemplo, se decidió mantener para la variable grupos étnicos en los totales de cada categoría a nivel regional; esta propiedad permite caracterizar la población étnica en cada región y con esta información los usuarios pueden realizar análisis estadístico por regiones.

   iii) **Mantener correlaciones entre variables:** Esta propiedad busca conservar las posibles relaciones lineales o no lineales que se puedan presentar entre las variables. Se debe definir si se mantienen 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. Esto también aplica en el caso de una regresión, donde la utilidad analítica se puede medir a través de la preservación de los coeficientes betas asociados a los predictores.

   iv) **Mantener tendencias de las variables a través del tiempo:** Esta propiedad hace referencia a que las variables conserven un comportamiento en determinados períodos de tiempo. Por ejemplo, si la base de datos de una operación estadística de temática económica contiene la variable ingreso de los hogares chilenos, y esta variable ha presentado un comportamiento decreciente en el primer trimestre de 2021, al publicar la base de datos anonimizada el equipo de trabajo desea garantizar que esta tendencia se conserve.

Notas:

1. En el caso en que el equipo de trabajo no tenga identificadas las demandas de información y/o desconozca los usuarios de microdatos, el equipo de trabajo podrá tener en cuenta distintos recursos; tales como vacíos de información del SEN y prioridades de política pública.
2. El cálculo de las métricas estadísticas definidas en el punto 2 tiene lugar en el análisis exploratorio alojado en la **Etapa 6.4.2: Preparar y explorar los datos** .

Tarea 1-4.1. Definición de porcentajes de variación permitidos por variable y niveles de desagregación geográfica o temática para las características estadísticas: Las estadísticas calculadas a partir del archivo de microdatos anonimizado y publicado, deben producir resultados analíticos que estén de acuerdo con las estadísticas publicadas previamente de los datos originales. Por tanto, se debe definir cuál es la discrepancia o porcentaje de variación aceptable entre el resultado obtenido con los datos originales versus los anonimizados. Esta definición es clave para medir la utilidad que compara los datos originales y los datos anonimizados, teniendo en cuenta la necesidad del usuario final para su análisis (ver **Etapa 6.4.4.2: Evaluar proceso SDC**).

La definición de estos porcentajes depende de los objetivos y características de cada operación estadística, y concierne al equipo a cargo.

Nota:
Esto es especialmente importante si se utilizan métodos perturbativos (ver **Etapa 6.4.4.1: Seleccionar y aplicar métodos *SDC*)**.

Tarea 1-4.2. Priorización de indicadores o características estadísticas: Es posible que no todas las estadísticas publicadas se puedan generar a partir de los datos publicados. Si este es el caso, el equipo de trabajo debe realizar lo siguiente:

1. Definir en qué indicadores y estadísticas enfocarse.
2. Informar a los usuarios sobre los indicadores que han sido priorizados y por qué. El orden de importancia deberá documentarse claramente para el usuario en el reporte externo (ver **Etapa 6.4.5: Generar reportes y liberar datos**), que la priorización de ciertas métricas, sobre otras, significa que ciertas métricas ya no son válidas. Por ejemplo, un cierto método SDC puede conducir a una menor pérdida de información para las cifras de la fuerza laboral, pero una mayor pérdida de información para las relaciones con niveles de educación.
   Esto es necesario, ya que no es posible liberar varios archivos para diferentes usuarios.

   En relación con las características estadísticas prioritarias, los elementos mínimos que deben ser descritos son los siguientes:

   i) Descripción de las características estadísticas priorizadas para conservar en la base de datos anonimizada.
   ii) Listado de las variables con la respectiva característica estadística a conservar en la base de datos anonimizada.
   iii) Nivel de desagregación geográfica o temática en el que se ha decidido conservar las características estadísticas.
   iv) Porcentajes de variación permitidos por variables y niveles de desagregación geográfica o temática para las características globales en la base de datos anonimizada.

Actividad 1-4: Elaboración de reporte sobre definiciones previas al proceso de anonimización: Se debe elaborar un reporte de esta etapa junto con la siguiente ("Preparar y explorar datos originales") que contenga al menos los siguientes elementos:

1. Antecedentes.
2. Control de versiones: versión, fecha del reporte, descripción de cambio(s) realizado(s), nombres de quienes elaboraron el reporte, nombres de quienes supervisaron y aprobaron el reporte.
3. Descripción del equipo de trabajo: nombres de analistas, roles y responsabilidades. En este punto se recomienda precisar un diagnóstico más concreto de capacidades disponibles en el equipo, para roles de implementación y apoyo en pasos subsiguientes del proceso. En este paso, se ponen a prueba estas capacidades, que se establecen previamente "en teoría", por lo que corresponde actualizar este diagnóstico del equipo en caso de percibir discrepancias con lo establecido inicialmente. Adicionalmente, se debe considerar una breve de descripción de cómo se organizará el trabajo durante el proceso de anonimización en función de los distintos roles establecidos.
4. Resumen sobre revisión de insumos y productos estadísticos necesarios para la implementación del proceso de anonimización. Es importante que se describa con qué insumos el equipo de trabajo inicia el proceso SDC: Metodología del producto estadístico, bases de datos, diccionarios de variables, *software* y paquetes estadísticos, infraestructura tecnológica y seguridad de la información, entre otros.
5. Especificación de aspectos clave para establecer necesidades de protección de la confidencialidad, dando cuenta del marco normativo y convenios asociados al producto estadístico, unidades estadísticas y variables sensibles o variables de identificación contenidas en los archivos de datos.
6. A partir de esto, se define un diagnóstico de necesidades de protección de confidencialidad, indicando si corresponde o no llevar a cabo el proceso de anonimización.
7. Definición de las características estadísticas a preservar, indicando los usos claves de datos, una priorización de indicadores y/o características estadísticas y la medición de utilidad.
8. Los apartados posteriores corresponden al subproceso de "Preparar y explorar los datos originales", que se precisarán en la **Actividad 2-3**.
9. Descripción de incidencias encontradas en el desarrollo de este subproceso (información o detalle que permita actualizar el proceso de 2. "Diseño y planificación", contingencias derivadas del proceso de 6.3"Interpretar y explicar los resultados", etc.).
10. Para la elaboración de este documento, debe basarse en el formato de Reporte SDC correspondiente a los subprocesos de "Definiciones previas" y "Preparar y explorar los datos originales".

8.3.2. Etapa 6.4.2: Preparar y explorar los datos originales

- **Objetivo:** El propósito de esta etapa es preparar el conjunto de datos originales y luego explorar las características y estructura de los datos, que son importantes para los usuarios de los datos.

- **Alcance:** Los procedimientos descritos para esta etapa aplican para todas las operaciones estadísticas y productos relacionados cuyo levantamiento de información y/o publicación sea realizado por el INE (encuestas, censos, procesos de múltiples fuentes y registros administrativos) que darán a conocer información al público general u otros usuarios.

- **Exclusiones:** Las exclusiones generales se encuentran listadas en la sección **Exclusiones al alcance**. No aplican exclusiones adicionales para esta etapa.

- **Palabras claves:** Preparación de datos originales, exploración de datos originales.

En la **Tabla 10** se resume responsabilidades, *inputs* y *outputs* relacionados con la etapa Preparar y explorar los datos originales.

Tabla 10: Responsables, *inputs* y *outputs* para la etapa Preparar y explorar los datos originales

Responsable(s)	Analista(s) temático y analista(s) de anonimización.
Input(s)	- Reporte sobre definiciones previas al proceso de anonimización. - Base de datos final, objeto de anonimizar. - Diccionario y/o definición de variables proveniente del proceso 2. “Diseño y planificación”. - Infraestructura tecnológica y mecanismos de seguridad definidos.
Output(s)	- Producto estadístico (base de datos) a anonimizar preparada y caracterizada (análisis exploratorio). - Reporte sobre preparar y explorar los datos originales.
Control o supervisión	Jefe de área o jefe de proyecto.

Fuente: Instituto Nacional de Estadísticas (INE).

Actividad 2-1. Preparación de datos: Se debe preparar el conjunto de datos objeto de anonimización, esto incluye:

1. **Integración de datos**. Si hay varios archivos de datos, combinarlos y considerar todos los archivos de datos relacionados. En la preparación del conjunto de datos, si es que existe más de una base de datos, se deben explorar los datos, transformar a una estructura común, de forma de combinar y considerar

todos los archivos de datos relacionados (integración de datos). Así, obtener una base de datos que contenga toda la información necesaria.

En el criterio de combinación de microdatos: cuando más de un microdato refieran a la misma unidad estadística y ambos microdatos puedan ser vinculables por potenciales intrusos (variables de *match* entre los microdatos). Por ejemplo, esto se da en el caso de la Encuesta Laboral (ENCLA), con cuatro cuestionarios y Encuesta de Microemprendimiento (EME), con 2 cuestionarios.

2. **Eliminación de identificadores directos**. Cualquier identificador obvio de individuo, hogar o grupo, como nombres, números de identificación o direcciones, se suprime del conjunto de datos final, lo que evita que los usuarios de datos vinculen esa información con los individuos en particular. Los identificadores directos revelan directa e inequívocamente la identidad de una unidad o informante. La eliminación de identificadores directos es un proceso sencillo y siempre es el primer paso para producir una publicación segura de datos estadísticos.
3. **Selección de variables** que contienen información relevante para los usuarios finales y que deben incluirse en el conjunto de datos para su publicación.
4. **Consolidación de variables** que proveen la misma información. Cuando dos o más identificadores indirectos entreguen exactamente la misma información para cada unidad del microdato, estas deberán identificarse y reportarse. Esta duplicación de las variables podría deberse a que corresponden a cuestionarios aplicados a distintos informantes sobre las mismas unidades estadísticas, o a que una variable es la consolidación de varias otras (por ejemplo, ver **Tabla 11**). Una de las variables "espejo" deberá ser removida del microdato que entra al proceso de anonimización, y solo podrá ser reintegrada al microdato final si no presenta inconsistencias con la variable considerada en la anonimización. Es útil consolidar variables que proporcionan la misma información cuando sea posible, esto ayuda a:

    - Reducir la probabilidad de inconsistencias.
    - Minimizar las variables que un intruso puede usar para reconstruir los datos.

Tabla 11: Ilustración de consolidación de variables sin pérdida de información para proceso SDC

Antes					Después	
En la fuerza de trabajo	Empleado	Rama A	Rama B	Rama C	En la fuerza de trabajo	Empleado
Sí	Sí		Sí		Sí	B
No	No				No	No
Sí	Sí	Sí			Sí	A
Sí	Sí		Sí		Sí	B
Sí	Sí			Sí	Sí	C
Sí	No				Sí	No

Fuente: Instituto Nacional de Estadísticas (INE).

Notas:
En esta actividad, también puede ser útil eliminar variables distintas de los identificadores directos del conjunto de microdatos que se publicarán. Sin embargo, la decisión de eliminar variables depende de los objetivos y características de la operación estadística, y concierne al equipo a cargo. A continuación, se listan ejemplos de casos extraídos de Benschop, Machingauta, & Welch (2021, pág.110) donde se puede considerar la eliminación de variables:

1. Variables que son demasiado sensibles para ser anonimizadas y publicadas (por ejemplo, afiliación política, religión y variables relacionadas con la salud).
2. Variables de texto y de gestión interna.
3. Aquellas que no son importantes para los usuarios de datos y que podrían aumentar el riesgo de divulgación.

Actividad 2-2. Exploración del conjunto de datos: Se deben explorar las características y estructura del conjunto de datos. La compilación de un inventario de estas características es importante para evaluar la utilidad de los datos después de la anonimización y la producción de un conjunto de datos anonimizados, que es útil para los usuarios finales. Esta actividad incluye las siguientes tareas:

Tarea 2-2.1. Cálculo de porcentaje de valores perdidos en las variables: Es posible distinguir dos situaciones para explicar la presencia de valores faltantes:

i) Variable con muchos valores perdidos, por ejemplo, una variable registrada solo para un grupo selecto de individuos elegibles para un módulo de encuesta en particular, y valores perdidos para el resto (No Aplica). Algunos ejemplos son variables relacionadas con la educación (grado actual), donde un valor faltante indica que la persona no está actualmente en la escuela, o variables relacionadas con el parto, donde un valor faltante indica que la persona no ha dado a luz a un niño en el período de referencia.
ii) Variable con muchos valores faltantes debido a la falta de respuesta (No Sabe/No Responde). Esto aplica en el caso de variables donde debió registrarse respuesta y, por alguna razón, el campo se encuentra vacío, y además métodos de imputación que no son aplicables o no han sido aún aplicados.

Tras la validación, no deberían existir valores perdidos por no respuesta para el caso ii). En caso de que el equipo encargado de la anonimización encuentre valores perdidos por no respuesta, se debería levantar alerta y retroceder a **Etapa 6.4.1: Realizar definiciones previas al proceso de anonimización** (**Actividad 1-2**) hasta que exista una correcta validación del microdato.

La **Tabla 12** presenta un formato para registrar la distribución de valores perdidos por variable.

Tabla 12: Ilustración para registrar distribución de valores perdidos por variable

Nombre variable	Frecuencia Absoluta de valores perdidos por flujo de la encuesta	Frecuencia Relativa de valores perdidos por flujo de la encuesta	Frecuencia Absoluta de valores perdidos por no respuesta	Frecuencia Relativa de valores perdidos por no respuesta
Nombre Variable 1		Escriba en este espacio el porcentaje de valores faltantes observados para la variable 1		Escriba en este espacio el porcentaje de valores faltantes observados para la variable 1
Nombre Variable 2	Escriba en este espacio el número de valores faltantes observados para la variable 2		Escriba en este espacio el número de valores faltantes observados para la variable 2	

Fuente: Instituto Nacional de Estadísticas (INE).

Notas:

1. Las variables con alta proporción de valores perdidos pueden causar un alto nivel de riesgo de divulgación al tiempo que agregan poca información para los usuarios finales.
2. Los valores faltantes en sí mismos pueden ser reveladores, especialmente si indican que la variable no es aplicable.
3. Es importante tener presente la existencia de estos valores en el caso de los identificadores indirectos. Los valores perdidos deben considerarse como una categoría más de respuesta a la hora de hacer los análisis (por defecto, `sdcMicro` las considera). Para la anonimización, el equipo a cargo debería considerar en conjunto a los valores "no sabe" y "no responde" como una categoría única de valores perdidos. Para el análisis de los riesgos de divulgación no es pertinente la diferencia entre estos dos tipos de no respuesta, por lo que deben combinarse en una categoría única, para no sobreestimar los riesgos de divulgación (ver **Etapa 6.4.3: Medir** y **evaluar riesgos**).
4. Según Benschop, Machingauta, & Welch (2021, pág. 110) a menudo, las variables con alta proporción de los valores perdidos se eliminan en esta etapa. Sin embargo, para efectos de esta guía esto se debe tomar como una recomendación, pues la decisión de eliminar variables depende de los objetivos y características de la operación estadística, y concierne al equipo a cargo.

Tarea 2-2.2. Cálculo de estadísticas de resumen: Las estadísticas de resumen son medidas que forman parte de las características globales de las variables definidas en la **Actividad 1-3** y deben estar sujetas a verificación por parte del equipo de trabajo para examinar la validez del resultado del proceso SDC (ver **Etapa 6.4.4.2: Evaluar proceso SDC**). Dependiendo de la naturaleza de las variables, se pueden calcular las siguientes estadísticas de resumen:

i) **Distribución de frecuencias univariadas** para variables categóricas. La **Tabla 13** presenta un formato para registrar distribución de frecuencias para una variable categórica con dos categorías.
ii) **Distribución de frecuencias bivariadas** para variables categóricas.
iii) **Medidas descriptivas** (media, desviación estándar, coeficiente de variación, etc.) para cada una de las variables cuantitativas (o medibles). La **Tabla 14** presenta un formato para registrar medidas descriptivas para una variable cuantitativa.
iv) **Matriz de correlaciones** (para variables categóricas o variables cuantitativas).

Tabla 13: Ilustración para registrar distribución de frecuencias para una variable con dos categorías

Nombre variable categórica	Frecuencia Absoluta	Frecuencia Relativa
Categoría 1		Escriba en este espacio el porcentaje de unidades de observación que corresponden a la categoría 1
Categoría 2	Escriba en este espacio el número de unidades de observación que corresponden a la categoría 2	
Total	Total	

Fuente: Instituto Nacional de Estadísticas (INE).

Tabla 14: Ilustración para registrar medidas descriptivas para una variable cuantitativa

Nombre variable cuantitativa	Media	Mediana	Desviación estándar	Mínimo	Q1	Q3	Máximo
Escriba en este espacio el nombre de la variable medible			Escriba en este espacio la desviación estándar de la variable medible				

Fuente: Instituto Nacional de Estadísticas (INE).

Actividad 2-3: Elaboración de reporte sobre preparación y exploración de los datos originales: El reporte de esta etapa se registra en un mismo documento en conjunto con la etapa previa ("Definiciones Previas"). Debe contener al menos los siguientes elementos:

1. Control de versiones: versión, fecha del reporte, descripción de cambio(s) realizado(s), nombres de quienes elaboraron el reporte, nombres de quienes supervisaron y aprobaron el reporte.
2. Antecedentes.
3. Secciones referidas a "Definiciones Previas" (se precisan en **Actividad 1-5**).
4. Descripción de la preparación de datos (integración de datos, eliminación de variables, fusión o consolidación de variables, etc.), y los criterios empleados para este propósito.

5. Resumen del análisis exploratorio de datos (valores faltantes, tablas de frecuencias, estadísticas de resumen, correlaciones, etc.) basado en las medidas globales priorizadas en la **Actividad 1-4**.
6. Descripción de incidencias encontradas en el desarrollo de este subproceso (información o detalle que permita actualizar el proceso de 2. "Diseño y planificación", contingencias derivadas del proceso de 6.3 "Interpretar y explicar los resultados", etc.).
7. Para la elaboración de este documento, debe basarse en el formato de Reporte SDC correspondiente a los subprocesos de "Definiciones previas" y "Preparar y explorar los datos originales".

8.3.3. Etapa 6.4.3: Medir y evaluar riesgos

- **Objetivo:** El propósito de esta etapa es calcular medidas de riesgo sobre los datos originales o brutos y, en base a estas medidas, juzgar si un archivo de microdatos es lo suficientemente seguro para su publicación.

- **Alcance:** Los procedimientos descritos para esta etapa aplican para todas las operaciones estadísticas y productos relacionados cuyo levantamiento de información y/o publicación sea realizado por el INE (encuestas, censos, procesos de múltiples fuentes y registros administrativos) que darán a conocer información al público general u otros usuarios.

  En esta etapa solo se calculan las medidas de riesgo sobre los datos originales, que permiten decidir si el archivo de microdatos es lo suficientemente seguro para su publicación, o si requiere la aplicación de métodos SDC (ver

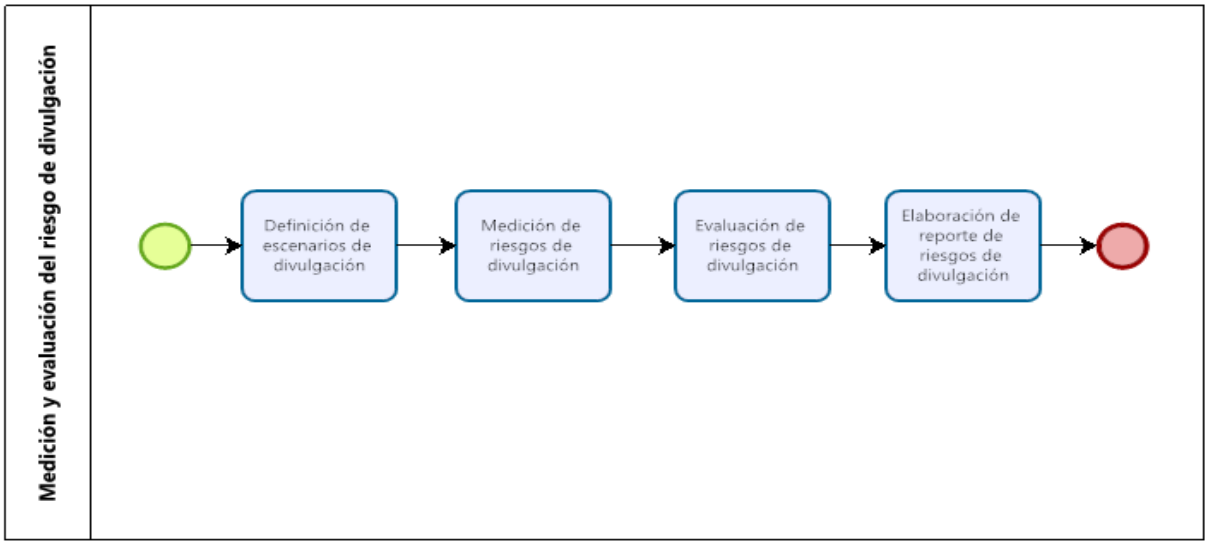
  Etapa 6.4.4.1: Seleccionar y aplicar métodos SDC). La decisión sobre qué hacer con las unidades riesgosas (una vez aplicados los métodos SDC) se aborda en la **Etapa 6.4.4.2: Evaluar proceso SDC**.

- **Exclusiones:** Las exclusiones generales se encuentran listadas en la sección **Exclusiones al alcance**. No aplican exclusiones adicionales para esta etapa.

- **Palabras claves:** Escenarios de divulgación, medición de riesgo de divulgación, riesgo global, riesgo jerárquico, riesgo individual, y evaluación de riesgo de divulgación.

La etapa Medir y evaluar riesgos comprende cuatro actividades que se deben desarrollar de forma secuencial según el orden que se indica en la ¡Error! No se encuentra el origen de la referencia..

Figura 5: Etapa medición y evaluación del riesgo de divulgación



Fuente: Instituto Nacional de Estadísticas (INE).

En la **Tabla 15** se resume responsabilidades, *inputs* y *outputs* relacionados con la actividad definición de escenarios.

Tabla 15: Responsables, *inputs* y *outputs* para la actividad definición de escenarios

Responsable(s)	Analista(s) temático y analista(s) de anonimización.
Input(s)	- Lista previa de identificadores indirectos (variables claves) derivada de las etapas **Etapa 6.4.1: Realizar definiciones previas al proceso de anonimización** y **Etapa 6.4.2: Preparar y explorar los datos** . - Inventario de fuentes de información externa disponibles, tanto públicas como privadas. - Producto estadístico (base de datos) a anonimizar preparada y caracterizada (análisis exploratorio). - Reporte sobre preparar y explorar los datos originales.
Output(s)	Definición de identificadores indirectos y escenarios de divulgación.
Control o supervisión	Jefatura área o jefe de proyecto.

Fuente: Instituto Nacional de Estadísticas (INE).

Actividad 3-1. Definición de escenarios de divulgación: Escenarios de divulgación (con priorización) para la evaluación de los riesgos de divulgación deben ser definidos.

Notas:

1. La evaluación del riesgo de divulgación se basa en los identificadores indirectos, que se identifican en el análisis de los escenarios de divulgación. El riesgo de divulgación depende directamente de la inclusión o exclusión de variables en el conjunto de identificadores indirectos definidos en la **Etapa 6.4.1: Realizar definiciones previas al proceso de anonimización**. Por lo tanto, este paso en el proceso SDC (elección de identificadores indirectos) se debe abordar con gran atención y cuidado.
2. Los escenarios de divulgación incluyen tanto la divulgación de identidad como de atributos.

Tarea 3-1.1. Elaboración de inventario de conjuntos de datos que pueden ser usados para re-identificación: Para la definición de escenarios de divulgación, se debe elaborar un inventario de todos los conjuntos de datos disponibles para los intrusos en el país.

1. Se deben considerar tanto conjuntos de datos publicados por el INE como por otras fuentes.

Ejemplos de fuentes de datos externos para la elaboración de inventarios son:

- Encuestas publicadas por el INE
- Datos censales
- Padrones electorales
- Registros de población
- Registros del Servicio de Impuestos Internos (SII)
- Datos recopilados por bancos y empresas privadas, entre otras.

2. Se deben listar los identificadores indirectos (variables claves) incluidos en los conjuntos de datos compilados en el punto 1, que un intruso puede utilizar para re-identificar unidades de la base de microdatos a publicar (es decir, que permitan realizar *matching* entre los conjuntos de datos externos y el conjunto de microdatos a publicar).
3. Registrar para cada identificador indirecto (variable clave) listado el (los) intruso(s) que tiene(n) acceso a esa variable.

Notas:

1. No todos los datos externos son necesariamente de dominio público. También deben tenerse en cuenta los conjuntos de datos de propiedad privada o los conjuntos de datos que no se publican para la definición de escenarios de divulgación (Benschop, Machingauta, & Welch, 2021, pág. 26).
2. Es esta información la que servirá como métrica clave a la hora de decidir qué variables elegir como identificadores indirectos potenciales, además de dictar el nivel de riesgos y métodos SDC necesarios.

3. A nivel institucional se contará con un inventario centralizado, que contendrá la información acerca de fuente de datos públicos, que proveerá de una base común.

Tarea 3-1.2. Definición de escenarios de divulgación: Se deben definir entre uno y cinco escenarios de divulgación. La definición de cada escenario viene dada por el listado de variables comunes a un mismo intruso.

Notas:

1. Pueden ser especificados tanto identificadores indirectos categóricos como cuantitativos.
2. Considerar como mínimo un identificador indirecto, pero sin número máximo.
3. Considerar en la definición de los escenarios, variables que den cuenta de la desagregación geográfica. Por ejemplo, región, provincia, comuna.
4. Considerar el escenario de reconocimiento espontáneo. Bajo este escenario, se debe verificar combinaciones raras o patrones inusuales en las variables.
   Ejemplos de variables que pueden conducir a reconocimiento espontáneo son:
   - Número de integrantes del hogar
   - El área de un terreno
   - El número de trabajadores de una empresa
   - Ingresos y gastos
   - Enfermedades
   - Profesiones u oficios de baja prevalencia en el área geográfica circunscrita, entre otras.

   Estas variables, son algunos ejemplos que hacen que las unidades sean fácilmente identificables, especialmente cuando se combinan con otras variables de identificación como, por ejemplo, la región. Un ejemplo clásico es un edificio en un área rural.
5. Si el número de identificadores indirectos es alto, se recomienda reducir el conjunto de identificadores indirectos, eliminando algunas variables del conjunto de datos para su publicación. Sin embargo, esta decisión debe estar fundada bajo los siguientes criterios:
   - La variable no posee alto valor analítico y se puede prescindir de ella, o
   - La variable tiene una alta contribución al riesgo de divulgación. Si esta variable no se puede tratar adecuadamente mediante los métodos SDC (es decir, aún se mantienen altos niveles de riesgo), se debe quitar del conjunto de datos.

Tarea 3-1.3. Priorización de un escenario de divulgación: Un escenario de divulgación debe ser priorizado. La definición de este escenario puede responder a los siguientes criterios:

1. Considerar escenarios realistas (probables). Las variables o conjuntos de datos pueden no coincidir perfectamente (por ejemplo, diferentes definiciones, variables más o menos detalladas, diferentes períodos de tiempo, etc.). Los registros externos podrían no estar lo suficientemente actualizados y, por lo tanto, un *matching* exacto con la base de datos a anonimizar, puede ser poco probable.

2. Algunos criterios para priorizar la selección de los escenarios son: Probabilidad de datos disponibles para el intruso con más variables y categorías, y probabilidad de *matching* exitoso, considerando combinación de variables con mayor frecuencia.

Actividad 3-2. Medición de riesgos: Se deben calcular medidas de riesgo sobre el conjunto de datos no tratado, que permitan juzgar si este es lo suficientemente seguro para ser publicado, o si se requiere la aplicación de métodos SDC.

En la **Tabla 16** se resumen responsabilidades, *inputs* y *outputs* relacionados con la actividad cálculo de medidas de riesgo.

Tabla 16: Responsables, *inputs* y *outputs* para la actividad cálculo de riesgos

Responsable(s)	Analista(s) de anonimización.
Input(s)	Identificadores indirectos y escenarios de divulgación.
Output(s)	- Medidas de riesgo: **globales e individuales para cada escenario de divulgación**. - Rutina R con medición de riesgos.
Control o supervisión	Analista(s) temático.

Fuente: Instituto Nacional de Estadísticas (INE).

Notas:

1. Para la medición de los riesgos de divulgación es importante distinguir entre datos de muestra y datos censales. En el caso de los datos censales, es posible calcular directamente las medidas de riesgo asumiendo que el conjunto de datos cubre toda la población objetivo.
2. Para los datos que provienen de una muestra, las medidas de riesgo que se presentan en esta sección se basan en varios supuestos. En general, estas medidas se basan en supuestos bastante restrictivos y, a menudo, conducirán a estimaciones de riesgo conservadoras. Estas medidas de riesgo conservadoras pueden exageran el riesgo ya que asumen el peor de los casos. Sin embargo, deben cumplirse dos supuestos para que las medidas de riesgo sean válidas y significativas; los microdatos deben ser una muestra de una población mayor (sin censos) y deben estar disponibles las ponderaciones muestrales.
3. Las medidas de riesgo difieren para los identificadores indirectos categóricos y continuos. Para variables categóricas, se considerará el concepto de unicidad de combinaciones de valores de identificadores indirectos utilizadas para identificar individuos en riesgo. El concepto de unicidad, sin embargo, no es útil para las variables continuas, ya que es probable que todos o

muchos individuos tendrán valores únicos para esa variable, por definición de una variable continua.

4. Medidas de riesgo para las variables categóricas son generalmente medidas a priori, es decir, se pueden evaluar antes de aplicar los métodos SDC ya que se basan en el principio de unicidad.
5. Las medidas de riesgo para las variables continuas son medidas a posteriori (ver **Etapa 6.4.4.2: Evaluar proceso SDC**); se basan en comparar los microdatos antes y después de la anonimización. Por ejemplo, se basan en la proximidad de las observaciones entre los datos originales y tratados (anonimizados).

Tarea 3-2.1. Medición de riesgo individual para identificadores indirectos categóricos: Se deben calcular medidas de riesgo sobre el conjunto de datos no tratado, basado en identificadores indirectos categóricos, que permitan juzgar si este es lo suficientemente seguro para ser publicado, o si se requiere la aplicación de métodos SDC.

Notas:

1. El enfoque principal de la medición del riesgo para los identificadores indirectos categóricos es la divulgación de identidad.
2. La medición del riesgo de divulgación se basa en la evaluación de la probabilidad de re-identificación correcta de los informantes en los datos divulgados.
3. Se usa medidas basadas en conteos de frecuencias a partir de los microdatos reales que se publicarán.
4. En general, cuanto más rara es una combinación de valores de los identificadores indirectos de una observación en la muestra, mayor es el riesgo de divulgación de identidad.

Como ilustración, la **Tabla 17** muestra valores de 10 informantes para los identificadores indirectos “Área”, “Sexo”, “Nivel educacional” y “Situación laboral”. En este conjunto de datos, se encuentra siete combinaciones únicas de identificadores indirectos (es decir, patrones o llaves) de los cuatro identificadores indirectos que forman este escenario de divulgación. Ejemplos de claves son {“Urbana”, “Mujer”, “Secundaria incompleta”, “Ocupado”} y {“Urbana”, “Mujer”, “Primaria incompleta”,” No Está en la fuerza laboral”}. Sea f_k la frecuencia muestral de la *k-ésima* clave, es decir, el número de individuos en la muestra con valores de los identificadores indirectos que coinciden con la *k-ésima* clave. Esta sería 2 para la clave {“Urbana”, “Mujer”, “Secundaria incompleta”, “Ocupado”} y 1 para la clave {“Urbana”, “Mujer”, “Primaria incompleta”,” No Está en la fuerza laboral”}, la que es única para el registro número 3. Por definición, f_k es el mismo para cada registro que comparte una clave. Los registros en la muestra con f_k igual a 1 se conocen como muestra única. La **Tabla 17** contiene cuatro muestras únicas. Las medidas de riesgo se basan en esta frecuencia de muestreo.

Tabla 17: Ejemplo de conjunto de datos con: frecuencias de muestreo y población, y riesgo de divulgación individual

No	Área	Sexo	Nivel educacional	Situación laboral	w_i	f_k	F_k	r_k
1	Urbana	Mujer	Secundaria incompleta	Ocupado	180	2	360	0.0054
2	Urbana	Mujer	Secundaria incompleta	Ocupado	180	2	360	0.0054
3	Urbana	Mujer	Primaria incompleta	No está en la fuerza laboral	215	1	215	0.0251
4	Urbana	Hombre	Secundaria completa	Ocupado	76	2	152	0.0126
5	Rural	Mujer	Secundaria completa	No ocupado	186	1	186	0.0282
6	Urbana	Hombre	Secundaria completa	Ocupado	76	2	152	0.0126
7	Urbana	Mujer	Primaria completa	No está en la fuerza laboral	180	1	180	0.0290
8	Urbana	Hombre	Postsecundaria	No ocupado	215	1	215	0.0251
9	Urbana	Mujer	Secundaria incompleta	No está en la fuerza laboral	186	2	262	0.0074
10	Urbana	Mujer	Secundaria incompleta	No está en la fuerza laboral	76	2	262	0.0074

Fuente: Instituto Nacional de Estadísticas, Adaptación de Benschop, Machingauta, & Welch (2021, pág. 28).

Con w_i peso muestral; F_k es la suma de todos los pesos muestrales de los registros que comparten la misma clave k, es decir, $F_k = \sum_{i|clave\ individuo\ i\ correspondiente\ a\ la\ clave\ k} w_i$; r_k riesgo individual para la clave k, con $r_k = 1/F_k$.

Notas:

1. Si $F_k = 1$, la clave k es tanto una muestra como una población única y el riesgo de divulgación sería 1. Las poblaciones únicas son un factor importante a considerar al evaluar el riesgo y merecen especial atención.
2. Los valores r_k también se pueden interpretar como la probabilidad de divulgación para los individuos o como la probabilidad de una coincidencia exitosa con individuos elegidos al azar de un archivo de datos externo con los mismos valores de las variables clave.
3. Valores faltantes en los identificadores indirectos se deben tratar con cuidado. El equipo a cargo debería considerar en conjunto a los valores “no sabe” y “no responde” como una categoría única de valores perdidos. Para el análisis de los riesgos de divulgación no es pertinente considerarlos como categorías separadas, pues sobreestimarían el riesgo de divulgación.

Una medida de riesgo que sirve como complemento al riesgo individual es el k-anonimato. La medida de riesgo k-anonimato se basa en el principio de que, en un conjunto de datos seguro, el número de informantes que comparten la misma combinación de valores de identificadores indirectos categóricos debe ser mayor que un umbral especificado k.

La medida de riesgo es el número de observaciones que violan k-anonimato para un cierto valor de k, que es:

$$\sum_i I(f_k < k)$$

Donde I es una función indicativa e i se refiere al *i-ésimo* registro. Es decir, un recuento del número de individuos con una frecuencia de muestreo de su clave inferior a k.
Por ejemplo, en la **Tabla 17**, cuatro registros violan el 2-anonimato y los 10 registros violan el 3-anonimato.

Notas:

1. La medida de riesgo k-anonimato no considera los pesos de la muestra, pero **es importante considerar los pesos de la muestra al determinar el nivel requerido de k-anonimato.**
2. Si los pesos muestrales son grandes, un individuo en el conjunto de datos representa a más individuos en la población objetivo, la probabilidad de una coincidencia correcta es menor y, por lo tanto, el umbral requerido puede ser menor.
3. Los pesos de muestra grandes van de la mano con conjuntos de datos más pequeños.
4. En un conjunto de datos más pequeño, la probabilidad de encontrar otro registro con la misma clave es menor que en un conjunto de datos más grande. Esta probabilidad está relacionada con el número de registros en la población con una clave particular a través de los pesos muestrales.
5. En el marco de censos, esta medida toma mayor relevancia, en el sentido de que toda la población está contenida en el conjunto de datos. Es especialmente importante asegurar que ninguno de los registros viole el 2-anonimato (población única).
6. En el marco de registros administrativos, las notas anteriores aplican de acuerdo a la naturaleza del registro administrativo. Si este posee una estructura de encuesta (considerar las notas 1 a 4). Por el contrario, si el registro administrativo tiene la naturaleza de censo, considerar nota 5.
7. Es importante señalar que los valores perdidos (NA en R) se tratan como si fuera cualquier otro valor. Dos registros con claves {“Hombre”, “NA”, “Ocupado”} y {“Hombre”, “Secundaria completa”, “Ocupado”} comparten la misma clave, y similarmente, {“Hombre”, “NA”, “Ocupado”} y {“Hombre”, “Secundaria incompleta”, “Ocupado”} también comparten la misma clave. Por lo tanto, el valor perdido es interpretado como “Secundaria completa” en la primera clave y “Secundaria incompleta” en la segunda. Esto se ilustra en la
8.
9. **Tabla** 18**.** Note que la frecuencia de muestreo del tercer registro es 3, ya que se considera que comparte clave tanto con el primer registro como con el segundo.

Tabla 18: Ejemplo de conjunto de datos para ilustrar el efecto de valores perdidos sobre el *k*-anonimato

No	Sexo	Nivel educacional	Situación laboral	f_k
1	Hombre	Secundaria completa	Ocupado	2
2	Hombre	Secundaria incompleta	Ocupado	2
3	Hombre	NA	Ocupado	3

Fuente: Instituto Nacional de Estadísticas, Adaptación de Benschop, Machingauta, & Welch (2021, pág. 32).

Tarea 3-2.2. Medición de riesgo global: Se deben calcular medidas de riesgo global sobre el conjunto de datos no tratado, basado en identificadores indirectos categóricos, que permitan juzgar si este es lo suficientemente seguro para ser publicado, o si se requiere la aplicación de métodos SDC.

Notas:

1. Para construir una medida de riesgo agregada a nivel global para el conjunto de datos completo, se puede agregar las medidas de riesgo a nivel individual de varias formas. Las siguientes son las dos principales:
   i) A través de la media de las medidas de riesgo individuales.

$$R_1 = \frac{1}{n}\sum_i r_k = \frac{1}{n}\sum_k f_k r_k$$

   Donde r_k es el riesgo individual para la clave *k* que comparte el *i-ésimo* registro.
   El riesgo global en los datos de ejemplo en la **Tabla 17** es 0.01582, que es la proporción esperada de todos los individuos en la muestra que podrían ser re-identificados por un intruso. Otra forma de expresar el riesgo global es la cantidad de re-identificaciones, $n \cdot R_1$, que en el ejemplo es $10 \cdot 0.01582$.

   ii) A través del recuento de unidades con riesgos superiores a un determinado umbral de riesgo individual.
   Por ejemplo, en los datos de la **Tabla 17**, el 50% de los registros posee riesgos superiores a 0.01 o 1%, y no hay registros con riesgo superior a 0.05 o 5%.

2. **Las medidas de riesgo global deben usarse con precaución:** detrás de un riesgo global aceptable se pueden esconder algunos registros de muy alto riesgo que son compensados por muchos registros de bajo riesgo.

Tarea 3-2.3. Medición de riesgo jerárquico: En el caso de que el conjunto de datos posea una estructura jerárquica, se deben calcular medidas de riesgo individual y global sobre el conjunto de datos no tratado considerando esta estructura, que permitan juzgar si este es lo suficientemente seguro para ser publicado, o si se requiere la aplicación de métodos SDC.

Notas:

1. La re-identificación de una unidad que pertenece a una entidad de nivel superior, puede permitir la re-identificación de las otras unidades pertenecientes a la misma entidad. Por ejemplo, en el contexto de una encuesta a hogares, la re-identificación de un miembro del hogar también puede llevar a la re-identificación de los otros miembros del hogar. Por tanto, se puede ver que, si se tiene en cuenta la estructura del hogar, el riesgo de re-identificación es el riesgo de que al menos uno de los miembros del hogar sea re-identificado. El riesgo global viene dado por:

$$r^h = P(A_1 \cup A_2 \cup \ldots \cup A_J) = 1 - \prod_{j=1}^{J} 1 - P(A_j)$$

Donde A_j es el evento de que el *j-ésimo* miembro de un hogar sea re-identificado y $P(A_j) = r_k$ es el riesgo de divulgación individual del *j-ésimo* miembro.

Por ejemplo, si un hogar tiene tres miembros con riesgos de divulgación individuales basados en sus respectivas claves 0.02, 0.03 y 0.03, respectivamente, el riesgo del hogar es

$$r^h = P(A_1 \cup A_2 \cup \ldots \cup A_J) = 1 - \prod_{j=1}^{J} 1 - P(A_j) = 1 - (1 - 0.02)(1 - 0.03)(1 - 0.03) = 0.078.$$

2. **El riesgo jerárquico no puede ser menor que el riesgo individual, y el riesgo del hogar es siempre el mismo para todos los miembros del hogar.**
3. Si la estructura jerárquica está presente en el conjunto de datos, siempre debe ser considerada. De lo contrario, los riesgos de divulgación pueden ser subestimados, al mismo tiempo que puede provocar inconsistencias en la estructura del conjunto de datos.

$$r^h = P(A_1 \cup A_2 \cup \ldots \cup A_J) = 1 - \prod_{j=1}^{J} 1 - P(A_j)\, 1 - (1 - 0.02)(1 - 0.03)(1 - 0.03) = 0.078.$$

Actividad 3-3. Evaluación de riesgos: Las medidas de riesgos obtenidas sobre el conjunto de microdatos no tratado deben ser evaluadas, a fin de juzgar si este es lo suficientemente seguro para ser publicado, o si se requiere la aplicación de métodos SDC.

En la **Tabla 19** se resume responsabilidades, *inputs* y *outputs* relacionados con la actividad evaluación de riesgos.

Tabla 19: Responsables, *inputs* y *outputs* para la actividad evaluación de riesgos

Responsable(s)	Analista(s) temático y analista(s) de anonimización.
Input(s)	Medidas de riesgo: globales e individuales para cada escenario de divulgación.
Output(s)	Evaluación de riesgos para el escenario de divulgación.
Control o supervisión	Jefatura área o jefe de proyecto.

Fuente: Instituto Nacional de Estadísticas (INE).

En la **Tabla 20** se presenta umbrales de riesgos según operación estadística que se consideran aceptables para decidir la liberación del conjunto de microdatos.

Tabla 20: Umbrales de riesgo aceptables para liberación de microdatos según tipo de operación estadística

Riesgos	Porcentaje de riesgos	Encuestas a hogares	Encuestas económicas	RR.AA.	Censo de Población y Vivienda	Censo Agropecuario y Forestal
Riesgo global	% Riesgo Global	<10%	<5%	<5%	<2%	<2%
Riesgo individual	% de unidades con riesgo sobre el 1%	<20%	<20%	<5%	<1%	<1%
	% de unidades con riesgo sobre el 5%	<15%	<15%	<3%	0%	0%
	% de unidades con riesgo sobre el 25%	0%	<10%	0%	0%	0%
	% de unidades con riesgo sobre el 50%	0%	<5%	0%	0%	0%
	% de unidades con riesgo sobre el 90%	0%	<1%	0%	0%	0%
	% de unidades con riesgo del 100%	0%	0%	0%	0%	0%
k-anonimato	% de observaciones violando 2k	0%	0%	0%	0%	0%
	% de observaciones violando 3k	<5%	0%	<2%	0%	0%

Riesgos	Porcentaje de riesgos	Encuestas a hogares	Encuestas económicas	RR.AA.	Censo de Población y Vivienda	Censo Agropecuario y Forestal
	% de observaciones violando 5k	<10%	<5%	<5%	<5%	<5%

Fuente: Instituto Nacional de Estadísticas (INE).

Notas:

1. Los umbrales en la **Tabla 20** solo establecen límites máximos según tipo de operación estadística, sin embargo, umbrales específicos para cada operación estadística, deben ser decididos por el equipo a cargo de dicha operación estadística. Es decir, para alguna operación estadística en particular, estos umbrales pueden llegar a ser más exigentes.
2. Una vez evaluados los riesgos, se recomienda revisar los patrones de los identificadores indirectos. Esto con el fin de pesquisar qué variables pueden estar contribuyendo de manera importante en el aumento de los riesgos y así, focalizar la atención para un posterior tratamiento. Por ejemplo, en el contexto de encuestas a hogares, la variable tamaño del hogar puede provocar riesgos altos.
3. Las encuestas dirigidas a personas corresponden al caso con conjuntos de datos sin estructura jerárquica. Por el momento, no hay umbrales fijados para estos casos específicos, así que, se recomienda asumir los umbrales propuestos para encuestas a hogares o encuestas económicas, según sea la temática abordada en la encuesta.

Actividad 3-4. Elaboración del informe de riesgos: Se debe elaborar un informe de riesgos que será utilizado en la etapa de

Etapa 6.4.4.1: Seleccionar y aplicar métodos *SDC*, el que describirá los criterios utilizados para la especificación de identificadores indirectos y definición de escenarios de divulgación, el cálculo de las medidas de riesgo y su evaluación, la definición de unidades riesgosas a la hora de publicar la base de datos, y métodos SDC sugeridos para realizar el control a la divulgación.

En la **Tabla 21** se resume responsabilidades, *inputs* y *outputs* relacionados con la actividad elaboración del informe de riesgos.

Tabla 21: Responsables, *inputs* y *outputs* para la actividad elaboración del informe de riesgos

Responsable(s)	Analista(s) temático y analista(s) de anonimización.
Input(s)	Evaluación de riesgos para el escenario de divulgación.
Output(s)	Informe de medidas de riesgos con datos originales o brutos.
Control o supervisión	Jefatura área o jefe de proyecto.

Fuente: Instituto Nacional de Estadísticas (INE).

El reporte de esta etapa se registra en un mismo documento en conjunto con las etapas posteriores ("Seleccionar y aplicar métodos SDC" y "Evaluar proceso SDC"). Debe contener al menos los siguientes elementos:

1. Control de versiones: versión, fecha del reporte, descripción de cambio(s) realizado(s), nombres de quienes elaboraron el reporte, nombres de quienes supervisaron y aprobaron el reporte.
2. Antecedentes.
3. Descripción de base de datos originales.
4. Identificación y descripción de variables sensibles e identificadores indirectos (variables claves).
5. Criterios y aspectos considerados en la definición de escenarios.
6. Criterios y aspectos considerados en la selección de escenarios (priorizados).
7. Tabla resumen de los riesgos calculados (globales e individuales) para cada escenario priorizado.
8. Evaluación de las medidas de riesgo para cada escenario priorizado.
9. Conclusiones sobre necesidad de seleccionar y aplicar métodos de anonimización y recomendaciones de métodos.
10. Secciones referidas a "Seleccionar y aplicar métodos SDC" se precisan en la **Actividad 4-2**.
11. Secciones referidas a "Evaluar proceso SDC" se precisan en la **Actividad 5-5**.

8.3.4. Etapa 6.4.4.1: Seleccionar y aplicar métodos SDC

- **Objetivo:** El propósito de esta etapa es seleccionar y aplicar métodos SDC adecuados para lograr el cumplimiento tanto de los niveles de riesgo definidos en la **Etapa 6.4.3: Medir y evaluar riesgos** como la preservación de las características estadísticas de la base de datos definidas en la **Etapa 6.4.1: Realizar definiciones previas al proceso de anonimización**.

- **Alcance:** Los procedimientos descritos para esta etapa aplican para todas las operaciones estadísticas y productos relacionados cuyo levantamiento de información y/o publicación sea realizado por el INE (encuestas, censos, procesos de múltiples fuentes y registros administrativos) que darán a conocer información al público general u otros usuarios.

  Considera desde la selección de los métodos SDC a aplicar según su factibilidad, prioridad, ventajas y desventajas hasta la aplicación de los mismos. Además, durante el tratamiento de los datos, esta etapa incorpora la verificación del procesamiento mediante validaciones para evitar errores.

- **Exclusiones:** Las exclusiones generales se encuentran listadas en la sección **Exclusiones al alcance**. No aplican exclusiones adicionales para esta etapa.

- **Palabras claves:** Métodos SDC no perturbativos, métodos SDC perturbativos, `sdcMicro`.

En la **Tabla 22** se resumen responsabilidades, *inputs* y *outputs* relacionados con la etapa Seleccionar y aplicar métodos SDC.

Tabla 22: Responsables, *inputs* y *outputs* para la etapa Seleccionar y Aplicar métodos SDC

Responsable(s)	Analista(s) de anonimización.
Input(s)	- Producto estadístico (base de datos) a anonimizar preparada y caracterizada (análisis exploratorio). - Diccionario y/o definición de variables proveniente del proceso 2. “Diseño y planificación”. - Reporte sobre preparar y explorar los datos originales. - Informe de medidas de riesgos con datos originales o brutos. - Rutina R con medición de riesgos.
Output(s)	- Datos tratados (anonimizados), con la documentación de los métodos aplicados, con su orden y especificaciones respectivas. - Verificación de que la base solo fue modificada en la forma pretendida y no se produjeron errores de procesamiento durante el tratamiento de los datos. - Copia de los datos originales o no tratados. - Reporte de incidencias: trazabilidad de métodos, iteraciones y problemas. - Rutina de programación R con la aplicación de métodos. - Informe de riesgos actualizado (para evaluación). - Reporte sobre selección y aplicación de métodos SDC.
Control o supervisión	Analista temático, Jefatura área o jefe de proyecto.

Fuente: Instituto Nacional de Estadísticas (INE).

Actividad 4-1. Selección de métodos SDC: Se debe seleccionar métodos SDC apropiados para aplicar control a la divulgación sobre los identificadores indirectos contenidos en el escenario de divulgación. Para la selección de los métodos SDC, el equipo de trabajo debe considerar los siguientes elementos:

1. La necesidad de protección de datos (medida por el riesgo de divulgación, ver **Etapa 6.4.3: Medir y evaluar riesgos**).
2. La estructura de los datos y el tipo de variables. Los métodos deben elegirse de acuerdo con el tipo de variable, categórica o cuantitativa (continua o semicontinua), los requisitos de los usuarios y el tipo de publicación.
3. La influencia de diferentes métodos sobre las características de los datos. Esto es importante para los usuarios finales o la utilidad de datos.

Los métodos que se presentan en esta sección provienen de una gran cantidad de literatura sobre control de divulgación estadística. Los procesos que subyacen a muchos de los métodos son objeto de una extensa investigación académica y muchos, si no todos, son utilizados ampliamente por las ONE con experiencia en la preparación de microdatos para su publicación. Además, pueden ser implementados en R utilizando el paquete `sdcMicro`. Se discute en cada método para qué tipo de datos es adecuado, tanto en términos de características como: tipo de datos, sus principales ventajas y desventajas, así como la función R para su implementación.

La clasificación de los métodos como se presenta en la **Tabla 23** ofrece una buena visión general para elegir los métodos apropiados, además de su función respectiva del paquete `sdcMicro` de R.

Tabla 23: Resumen de métodos SDC [18]

Métodos	Técnicas	Descripción	Ventajas	Desventajas	Clasificación de método SDC	Tipo de datos	Función en sdcMicro	Referencias
Perturbativos	**PRAM**	Método en el que los puntajes de una variable categórica se alteran de acuerdo con ciertas probabilidades. Por lo tanto, es una clasificación errónea intencional con probabilidades de clasificación errónea conocida, predefinida a partir de una matriz de transición.	• Aplicar a subgrupos del conjunto de datos. • Probabilidad positiva de un *match* con un individuo erróneo. • Seleccionar la matriz de transición y probabilidad. • Útil con muchas variables y/o alta pérdida de información.	• Modificación de datos originales, se intercambian características. • Mayor complejidad y posibilidad de error (seleccionar estratos). • Aparenta que no hay anonimización. • En el caso de las encuestas por muestreo, cada observación puede tener un peso muestral	Probabilístico	Categórico	`pram`	(Templ, 2018); (Piertzak, 2020)

[18] Para conocimiento de otras opciones de métodos SDC, se puede consultar el documento "Difusión de archivos de microdatos- Principios, procedimientos y prácticas", disponible en www.ihsn.org/sites/default/files/resources/IHSN-WP005_SP.pdf.

Métodos	Técnicas	Descripción	Ventajas	Desventajas	Clasificación de método SDC	Tipo de datos	Función en sdcMicro	Referencias
				diferente, por lo que, después de la generalización, no se garantiza la coherencia.				
	Microagregación	Método que se basa en la sustitución de valores para una determinada variable con un valor común para un grupo de registros. La agrupación de registros se basa en una medida de proximidad de variables de interés. Los grupos de registros también se utilizan para calcular el valor de reemplazo.	• Sencilla de comprender y aplicar. • Versión univariada implica baja pérdida de información. • Versión multivariada permite disminuir bastante el riesgo. • Todos los valores originales están contenidos en la data tratada (distribucione	• *Trade-off* entre pérdida de información y disminución de riesgo según si se aplica de forma univariado o multivariada.	Probabilístico	Continuo	`microaggregation`	(Templ, 2008); (Cano & Torra, 2011)

Métodos	Técnicas	Descripción	Ventajas	Desventajas	Clasificación de método SDC	Tipo de datos	Función en sdcMicro	Referencias
			s univariadas no cambian). • Evita la introducción de inconsistencias que requieran edición.					
	Adición de ruido	Método basado en agregar o multiplicar un número aleatorio a los valores originales para proteger los datos de la coincidencia exacta con archivos externos. La adición de ruido se aplica típicamente a variables continuas.	• Puede añadir ruido de forma univariada o bivariada. • Sencilla de comprender y aplicar.	• Puede mantener niveles altos riesgos, no permite medir los riesgos post-aplicación. • Variable de número enteros puede pasar a tener decimales (por ejemplo, 3,5 hijos). • Puede incrementar rango de	Probabilístico	Continuo	`addNoise`	(Templ, 2008); (Cano & Torra, 2011); (Piertzak, 2020)

Métodos	Técnicas	Descripción	Ventajas	Desventajas	Clasificación de método SDC	Tipo de datos	Función en sdcMicro	Referencias
				valores de variable. • Necesaria edición posterior a la anonimización.				
	Barajado	Similar al *swapping*, pero utiliza un modelo de regresión subyacente para determinar qué variables se intercambian. El barajado mantiene las distribuciones marginales en los datos barajados. Valores generados por regresiones permite ranquear observaciones y cada valor original es reemplazado con otro valor	• Todos los valores originales están contenidos en la data tratada (distribuciones univariadas no cambian). • Permite mantener relaciones multivariadas.	• Requiere una clasificación completa de los datos, que puede ser computacionalmente muy intensiva para grandes conjuntos de datos con varias variables. • Necesita regresión lineal (deben cumplirse supuestos y la recta ajustar bien)	Probabilístico	Continuo	`shuffle`	(Templ, 2008); (Benschop et al., 2017)

Métodos	Técnicas	Descripción	Ventajas	Desventajas	Clasificación de método SDC	Tipo de datos	Función en sdcMicro	Referencias
		original de la variable, según el rango correspondiente.		• La técnica requiere al menos dos variables cuantitativas en la fórmula de la función.				
	Rank swapping	Intercambia valores de la variable entre las observaciones. Los valores de la variable se clasifican en orden ascendente y luego cada valor de la variable se intercambia con otro valor elegido al azardentro de un rango restringido.	• Todos los valores originales están contenidos en la data tratada (distribuciones univariadas no cambian). • Sencilla de comprender y aplicar. • Permite mantener relaciones multivariadas. • Baja pérdida de información.	• Reduce rango de valores mínimos y máximos. • Puede mantener niveles altos de riesgos.	Probabilístico	Continuo	`rankSwap`	(Piertzak, 2020)

Métodos	Técnicas	Descripción	Ventajas	Desventajas	Clasificación de método SDC	Tipo de datos	Función en sdcMicro	Referencias
No perturbativos	**Supresión local**	Suprimir ciertos valores en al menos una variable (agregar *missing* a datos). Los valores suprimidos son los de observaciones que tienen combinaciones de variables llaves que son compartidas por muy pocos individuos. Si se aplica, debe hacerse después de la recodificación global y superior-inferior No confundir con eliminar observación.	• Quitar datos atípicos. • Seleccionar datos a eliminar según importancia.	• No es útil para variables continuas ni con alto número de categorías. • Con alto número de variables o categorías puede no lograr solución. • Limitación de algoritmo en librería `SdcMicro`. • Observaciones con variables suprimidas quedan con valor perdido. ¿Cómo explicar un valor perdido si no es por	Determinístico	Categórico	`localSuppression`	(Templ, 2018)

Métodos	Técnicas	Descripción	Ventajas	Desventajas	Clasificación de método SDC	Tipo de datos	Función en sdcMicro	Referencias
				flujo o no respuesta/no sabe?				
	Recodificación Global	Combinar varias categorías en una nueva categoría menos informativa (para una variable continua, corresponde a su discretizar dicha variable).	• Se disminuye información, pero manteniendo estructura y relaciones.	• Pérdida de información continua al categorizar para estimación de modelos. • Cambia nivel de medición de variables continuas a ordinales o categóricas.	Determinístico	Continuo y categórico	`globalRecode`, `groupVars`	(Templ, 2018)
	Codificación superior e inferior	Se recodifican los valores superiores y/o inferiores de la distribución o de categorías.	• Se pierde menos información que con recodificación global, permite agrupar solo las colas de la distribución.	• Menor nivel de anonimización, se requiere evaluar *trade-off* entre ambos métodos.	Determinístico	Continuo y categórico	`topBotCoding`	(Templ, 2018)

Fuente: Instituto Nacional de Estadísticas (INE).

Notas:

1. En la práctica, la selección de los métodos es parcialmente un proceso constructivista de prueba y error: después de aplicar un método elegido, el riesgo de divulgación de datos y la utilidad se mide y se compara con otras opciones de métodos y parámetros.
2. La selección de los métodos está limitada por la legislación, por un lado, y una compensación entre utilidad y riesgo por el otro.
3. Con respecto al punto 3 (pág. 52), se establece que, para variables específicas y/o conjuntos de datos, como prioridad los métodos no perturbativos deben considerarse primero. Esta instrucción se basa en los siguientes argumentos:

    - Los métodos perturbativos distorsionan la estructura del dato a nivel desconocido para el usuario externo. Estos métodos se pueden usar sin recodificación previa de las variables, sin embargo, conservan la estructura original de la base solo parcialmente.
    - Con los métodos perturbativos puede no haber pérdida importante de información al nivel analizado internamente, pero la variación puede ser importante para el usuario externo que no se limita a replicar los cálculos INE (por ejemplo, estima modelos).
    - Escasa experiencia en métodos perturbativos en ONE de América Latina (por ejemplo, caso DANE).
    - Los métodos perturbativos pueden generar incertidumbre respecto de la recepción de los usuarios externos.
    - Los métodos perturbativos pueden crear inconsistencia en el microdato (por ejemplo, edad puede pasar a menos de 15 en encuesta de fuerza de trabajo).
    - Los métodos perturbativos se basan en introducir incertidumbre en el conjunto de datos y no en aumentar las frecuencias de claves en los datos y, por lo tanto, sobreestimará el riesgo.
4. **Como recomendación, para variables categóricas**, se puede considerar la siguiente priorización en la selección de métodos: Codificación superior e inferior -> Recodificación Global -> PRAM -> Supresión local. No es tan lineal este orden, depende de especificidades de los datos con que se trabaja.
5. **Como recomendación, para variables cuantitativas**, se puede considerar la siguiente priorización en la selección de métodos: Codificación superior e inferior -> Recodificación Global -> Adición de ruido -> Rank Swapping -> Microagregación -> Shuffle. No es tan lineal este orden, depende de especificidades de los datos con que se trabaja.
6. En el caso de las recodificaciones, se debe tener cuidado de generar nuevas categorías en línea con el uso de datos de los usuarios finales con el fin de minimizar la pérdida de información como resultado de la recodificación. Cualquier agrupación debe ser una agrupación lógica relacionada con la naturaleza de la variable tratada y no una unión aleatoria de categorías.
7. En el caso de la supresión local, menos supresiones en una variable aumenta el número de supresiones necesarias en otras variables. En particular, si el porcentaje de unidades riesgosas es alto, conviene aplicar recodificación global antes de una supresión local (ver nota con recomendación para variables categóricas).

8. En el PRAM, las tabulaciones univariantes no se modifican, si se elige una matriz de transición que tenga la propiedad invariante. Además, es posible aplicar PRAM a subgrupos del conjunto de microdatos de forma independiente. En este caso, el usuario necesita seleccionar la variable de estratificación que define los subgrupos. Si se omite la especificación de esta variable, el procedimiento PRAM se aplica a todas las observaciones en el conjunto de datos.
9. La microagregación altera los valores periféricos, esto puede tener un impacto significativo en el cálculo de algunas medidas sensibles a los valores atípicos, como el índice GINI.
10. También es posible realizar una microagregación de forma independiente a grupos predefinidos.
11. Los métodos de recodificación, supresión local y microagregación son muy útiles para alcanzar medidas de k-anonimato.
12. Después de aplicar *shuffling* los valores originales están contenidos en el conjunto de datos tratado, esto implica que las tabulaciones univariantes no se modifican.
13. En los casos en que un modelo de regresión se ajuste bien a los datos, *shuffling* funcionaría muy bien, ya que debería haber suficientes regresores (continuos) disponibles.

Actividad 4-2. Aplicación de métodos SDC: Se debe aplicar los métodos seleccionados en la **Actividad 4-1.** Esto incluye:

1. Implementación de métodos seleccionados según tipo de variables y características de los datos.
2. Documentación de los métodos aplicados y medición de riesgos actualizados (para evaluación).
3. Documentación de las rutinas utilizadas en la programación de la aplicación de los métodos.
4. Reporte de incidencias: errores o fallos, iteraciones, etc.

Actividad 4-3. Elaboración de reporte sobre la selección y aplicación de métodos SDC: El reporte de esta etapa se registra en un mismo documento en conjunto con la etapa anterior ("Medir y evaluar riesgos") y la etapa posterior ("Evaluar proceso SDC"). Debe contener al menos los siguientes elementos:

1. Control de versiones: versión, fecha del reporte, descripción de cambio(s) realizado(s), nombres de quienes elaboraron el reporte, nombres de quienes supervisaron y aprobaron el reporte.
2. Antecedentes.
3. Resumen de los métodos SDC seleccionados para cada identificador indirecto contenido en el escenario de divulgación definido en la **Actividad 3-1**.
4. Documentación de las rutinas utilizadas en la programación de la aplicación de los métodos SDC.
5. Documentación de incidencias (errores o fallos, iteraciones, etc.)
6. Las secciones referidas a "Medir y evaluar riesgos" se precisan en la **Actividad 3-4**.
7. Las secciones referidas a "Evaluar proceso SDC" se precisan en la **Actividad 5-5**.

8.3.5. Etapa 6.4.4.2: Evaluar proceso SDC

- **Objetivo:** El propósito de esta etapa es verificar si la base de datos anonimizada cumple con las condiciones para presentarse como versión final. Estas son: el nivel de riesgo definido en la **Etapa 6.4.3: Medir y evaluar riesgos** y la utilidad esperada definida en la **Etapa 6.4.1: Realizar definiciones previas al proceso de anonimización**.

- **Alcance:** Los procedimientos descritos para esta etapa aplican para todas las operaciones estadísticas y productos relacionados cuyo levantamiento de información y/o publicación sea realizado por el INE (encuestas, censos, procesos de múltiples fuentes y registros administrativos) que darán a conocer información al público general u otros usuarios.

  Su alcance considera la reevaluación del riesgo y la medición de la utilidad, comparar con mediciones de los datos originales y decidir si la base cumple o no con los criterios establecidos.

- **Exclusiones:** Las exclusiones generales se encuentran listadas en la sección **Exclusiones al alcance**. No aplican exclusiones adicionales para esta etapa.

- **Palabras claves:** Reevaluación de riesgos, medición de utilidad, evaluación del proceso SDC.

En la **Tabla 24** se resume responsabilidades, *inputs* y *outputs* relacionados con la etapa Evaluar proceso SDC.

Tabla 24: Responsables, *inputs* y *outputs* para la etapa evaluar proceso SDC

Responsable(s)	Analista(s) de anonimización.
Input(s)	- Producto estadístico (base de datos) a anonimizar preparada y caracterizada (análisis exploratorio). - Diccionario y/o definición de variables proveniente del proceso 2. "Diseño y planificación". - Reporte sobre preparar y explorar los datos originales. - Informe de medidas de riesgos con datos originales o brutos. - Rutina R con medición de riesgos. - Datos tratados (anonimizados), con la documentación de los métodos aplicados, con su orden y especificaciones respectivas. - Verificación de que la base solo fue modificada en la forma pretendida y no se produjeron errores de procesamiento durante el tratamiento de los datos. - Copia de los datos originales o no tratados. - Reporte de incidencias: trazabilidad de métodos, iteraciones y problemas. - Rutina de programación R con la aplicación de métodos. - Informe de riesgos actualizado (para evaluación). - Reporte sobre selección y aplicación de métodos SDC.
Output(s)	- Reporte sobre evaluación del proceso SDC (reevaluación de riesgos y medición de utilidad). - Conclusión sobre si es factible o no la liberación de datos anonimizados. - Datos tratados (anonimizados).
Control o supervisión	Analista temático, jefatura área o jefe de proyecto.
Contingencias	- En caso de que el riesgo no está en un nivel aceptable se debe volver y repetir desde la - - **Etapa 6.4.4.1: Seleccionar y aplicar *métodos SDC***, para utilizar otros métodos y/o parámetros, revisando detalladamente la aplicación de los métodos SDC propuestos con el objetivo de verificar la inexistencia de errores de procedimientos. - En caso de que los datos no cumplen con la utilidad se debe iniciar el procedimiento desde la - - **Etapa 6.4.4.1: Seleccionar y aplicar *métodos SDC***, para utilizar otros métodos y/o parámetros, revisando detalladamente la aplicación de los métodos SDC.

Fuente: Instituto Nacional de Estadísticas (INE).

Actividad 5-1. Reevaluación de los riesgos de divulgación: Se debe reevaluar el riesgo de divulgación con las medidas y umbrales de riesgo elegidas en la **Etapa 6.4.3: Medir y evaluar riesgos (Actividad 3-3)**. Esto incluye los riesgos de divulgación tanto para las variables categóricas como continuas.

Tarea 5-1.1. Reevaluación de los riesgos de divulgación (variables categóricas): Se debe evaluar el riesgo de divulgación con las medidas y umbrales de riesgo elegidas en la **Etapa 6.4.3: Medir y evaluar riesgos (Actividad 3-3)** para los identificadores indirectos categóricos. Esto consiste en verificar que se han cumplido los umbrales establecidos para la operación estadística de acuerdo a los valores en la **Tabla 20**.

Tarea 5-1.2. Medición y evaluación de los riesgos de divulgación (variables continuas): Se debe medir y evaluar el riesgo de divulgación para los identificadores indirectos continuos. Esto incluye las siguientes medidas:

1. Medida de intervalo (Benschop, Machingauta, & Welch, 2021, pág. 39). Esta medida consiste en:
    i) Crear intervalos alrededor de cada valor perturbado y luego se determina si el valor original de esa observación perturbada está contenido en este intervalo.
    ii) Los valores que están dentro del intervalo alrededor del valor inicial después de la perturbación se consideran demasiado cercanos al valor inicial y, por lo tanto, no son seguros y necesitan más perturbaciones.
    iii) Los valores que están fuera de los intervalos se consideran seguros. El tamaño de los intervalos se basa en la desviación estándar de las observaciones y un parámetro de escala.

Notas:
1. Este método se implementa en la función **dRisk** () en `sdcMicro`.
2. Para la mayoría de los valores, este es un enfoque satisfactorio. Sin embargo, no es una medida suficiente para los valores atípicos. Después de la perturbación, los valores atípicos seguirán siendo valores atípicos y serán fácilmente re-identificables, incluso si están lo suficientemente lejos de sus valores iniciales. Por lo tanto, los valores atípicos deben tratarse con precaución.

2. Valores Atípicos. Identificar los valores de una variable continua que son mayores que un percentil $p\%$ predeterminado podría ayudar a identificar valores atípicos y, por lo tanto, unidades con mayor riesgo de identificación. El valor de p depende de la asimetría de los datos. El procedimiento a posteriori consiste en:
    i) Construir un intervalo alrededor de los valores perturbados (como en la medida de intervalo).
    ii) Si los valores originales caen en el intervalo alrededor de los valores perturbados, los valores perturbados se consideran inseguros ya que están demasiado cerca de los valores originales.

Notas:

1. Hay diferentes formas de construir dichos intervalos, como intervalos basados en rangos y basados en desviaciones estándar. Se recomienda utilizar los intervalos basados en la distancia robusta de Mahalanobis (RMD[19], por sus siglas en inglés) al cuadrado de los valores individuales. Los intervalos son escalados por el RMD de modo que los valores atípicos obtengan intervalos más grandes y, por lo tanto, necesitan tener una perturbación mayor para ser considerados seguros sobre los valores que no son valores atípicos.
2. Este método se implementa en sdcMicro en la función **dRiskRMD** (), que es una extensión de la función **dRisk** ().

Actividad 5-2. Evaluación de la utilidad del conjunto de microdatos anonimizados: Se debe verificar que las características estadísticas definidas en la **Etapa 6.4.1: Realizar definiciones previas al proceso de anonimización (Actividad 1-3)** se han preservado en el conjunto de datos anonimizado.

Tarea 5-2.1. Evaluación de la utilidad para las variables categóricas en el conjunto de microdatos anonimizado: Se debe verificar que las características estadísticas definidas en la **Etapa 6.4.1: Realizar definiciones previas al proceso de anonimización (Actividad 1-3)** para las variables categóricas se han preservado en el conjunto de datos anonimizado. Entre las medidas estadísticas de verificación se incluyen:

1. Comparar el número de valores faltantes en los datos.
2. El número de registros modificados por variable. El número de registros cambiados da una buena indicación del impacto de los métodos de anonimización en los datos.
3. Comparar tabulaciones univariadas.
4. Comparar tablas de contingencia entre pares de variables. Para mantener la validez analítica de un conjunto de datos, las tablas de contingencia deben permanecer aproximadamente iguales.

Notas:

1. La visualización de datos es una buena manera de evaluar de un vistazo cuánto han cambiado los datos después de la anonimización.
2. La visualización de datos puede ser una herramienta útil para evaluar el impacto en la utilidad de datos de los métodos de anonimización y ayudan a elegir entre los métodos de anonimización.
3. Se pueden emplear visualizaciones tales como histogramas, gráficos de densidad, gráficos de cajas y mosaicos, entre otros.
4. El lenguaje R proporciona varias funciones y librerías que pueden ayudar a visualizar los resultados de la anonimización. Entre ellos se encuentran las librerías `base` (R Core Team, 2019) y `ggpplot2` (Wickham, 2016).

[19] En inglés, *Robust Mahalanobis Distance*.

Tarea 5-2.2. Evaluación de la utilidad para las variables cuantitativas en el conjunto de microdatos anonimizado: Se debe verificar que las características estadísticas definidas en la **Etapa 6.4.1: Realizar definiciones previas al proceso de anonimización (Actividad 1-3)** para las variables cuantitativas (continuas y semicontinuas) se han preservado en el conjunto de datos anonimizado. Entre las medidas estadísticas de verificación se incluyen:

1. Comparar medias y variaciones (desviación estándar, coeficiente de variación, covarianza).
2. Comparar las distribuciones multivariadas de los datos. Especialmente los cambios en las correlaciones dan información valiosa sobre la validez de los datos para las regresiones.
3. Las regresiones son una herramienta útil para evaluar si la estructura en los datos se mantiene después de la anonimización. Al comparar los parámetros de regresión, también es posible comparar relaciones entre variables no continuas (por ejemplo, mediante la introducción de variables *dummies* o la regresión con variables ordinales). Se pueden usar regresiones comunes para comparar cambios en coeficientes e intervalos de confianza.
   Si las nuevas estimaciones caen dentro del intervalo de confianza original y los intervalos de confianza nuevos y originales se superponen en gran medida, los datos pueden considerarse válidos para este tipo de regresión después de la anonimización.
4. Comparar tendencias.
5. Otras estadísticas que caracterizan los datos incluyen los componentes principales y las cargas.

Nota:
Considerar las notas dadas para la **Tarea 5-2.1**.

Tarea 5-2.3. Evaluación de pérdida de información: En el caso de las variables continuas se debe medir y evaluar la pérdida de información.

La evaluación de pérdida de información se puede realizar utilizando la medida *IL1s*, que es la suma de las distancias absolutas entre las observaciones correspondientes en los conjuntos de datos originales y anonimizados, que están estandarizados por la desviación estándar de las variables en los datos originales. Para las variables continuas en el conjunto de datos, la medida de *IL1* se define como:

$$IL1s = \frac{1}{pn}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\left|x_{ij-z_{ij}}\right|}{\sqrt{2}S_j}$$

Donde p es el número de variables continuas; n es el número de registros en el conjunto de datos; x_{ij} y z_{ij}, respectivamente, son los valores antes y después de la anonimización para la variable j y el registro i; y la S_j es la desviación estándar de la variable j en los datos originales.

Notas:

1. Usando sdcMicro, la medida *IL1* puede ser calculada para todos los identificadores indirectos con la función **dUtility**().
2. La medida *IL1* es útil para comparar diferentes métodos. Cuanto menor sea el valor de la medida, más cercanos son los valores anonimizados a los valores originales y mayor es la utilidad.
3. Esta medida está relacionada con medidas de riesgo basadas en distancias e intervalos.
4. Cuanto mayor sea la distancia entre los valores originales y anonimizados, menor será la utilidad de los datos. Sin embargo, una mayor distancia también reduce el riesgo de re-identificación.

Tarea 5-2.4. Evaluación de la utilidad en el conjunto de microdatos anonimizado a partir de medidas de *"benchmarking"*: En el caso de que las necesidades de los usuarios definidas en la **Etapa 6.4.1: Realizar definiciones previas al proceso de anonimización** establezcan medidas analíticas específicas, como, por ejemplo, "indicadores de *benchmarking*", el equipo debe verificar que los datos anonimizados preservan estos indicadores. Ejemplos de este tipo incluyen: medidas de pobreza para conjuntos de datos de ingresos y relaciones de asistencia escolar, el coeficiente GINI, que es una medida de dispersión estadística, a menudo se usa para medir la desigualdad en el ingreso, entre otros indicadores.
Si las diferencias entre los indicadores (con datos brutos y datos anonimizados) no son demasiado grandes, el conjunto de datos anonimizados puede ser liberado para uso de los investigadores. Debe tenerse en cuenta que los indicadores calculados en las muestras son estimaciones con cierta variación e intervalo de confianza. Por lo tanto, para los datos de la muestra, **es más informativo comparar la superposición de los intervalos de confianza y/o evaluar si la estimación puntual calculada después del anonimato está contenida dentro del intervalo de confianza de la estimación original**.

Nota:
En R existen varios paquetes que tienen funciones para calcular el coeficiente GINI, por ejemplo, el paquete `laeken` (Alfons & Templ, 2013).

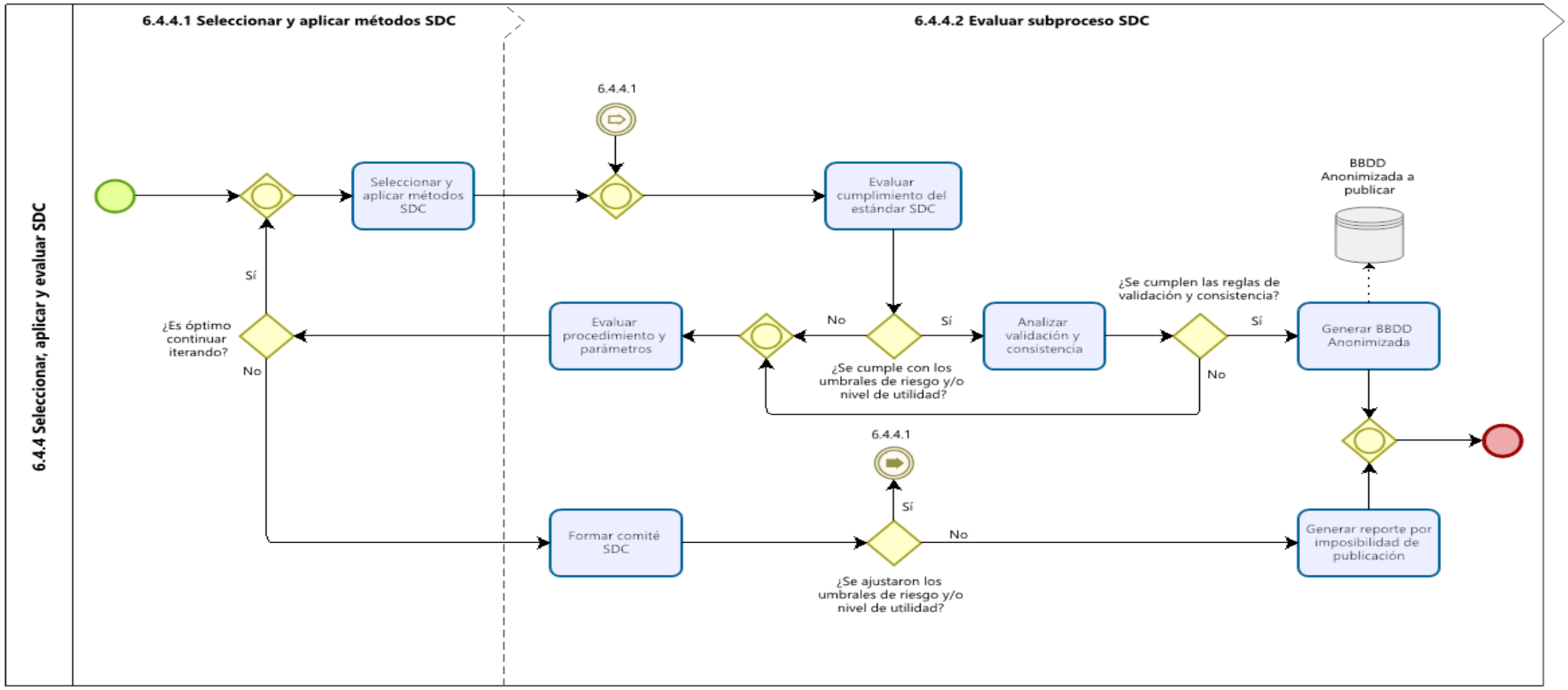
Actividad 5-3. Evaluación de reglas de validación y consistencia: Se debe verificar que todas las relaciones en los datos anonimizados preserven todas las reglas de validación y consistencia propias de la operación estadística. Esto incluye:

1. Variables que son sumas de otras variables o proporciones.
2. Relaciones de orden, por ejemplo, la variable X debe ser siempre menor a la variable Y.
3. Cualquier valor inusual causado por la anonimización debe ser detectado.
   Por ejemplo, los ingresos negativos, una persona de 14 años en la fuerza laboral o un alumno en el vigésimo grado de la escuela. Esto puede suceder después de aplicar métodos perturbativos de SDC.

Nota:
Se debe verificar que los indicadores publicados previamente de los datos originales o brutos son reproducibles a partir de los datos que se van a publicar. Si este no es el caso, los usuarios de datos podrían cuestionar la credibilidad del conjunto de datos anonimizados.

En la **Figura 6** se presenta el flujo que resume las **actividades 5-1 a 5-3** y que, en consecuencia, permite juzgar la efectividad de los métodos SDC aplicados y la factibilidad de liberar el conjunto de microdatos.

Figura 6: Evaluación de los métodos SDC



Fuente: Instituto Nacional de Estadísticas (INE).

Actividad 5-4. Constitución de comité de control a la divulgación: Se debe constituir un comité de control a la divulgación **solo en el caso de que exista un compromiso de entrega activo por convenio con alguna institución** y el proceso sufra un atasco debido a contingencias que incluyen: caso reiterado de unidades con incumplimiento de los umbrales de riesgo, una degradación considerable de las características estadísticas (nivel de utilidad) y la imposibilidad de seguir iterando.

1. **El comité se compone por el equipo a cargo del proceso SDC para** la operación estadística (analista(s) de anonimización, analista(s) temático y jefe de área o de proyecto), y **un equipo externo, conformado por un analista del Departamento de Análisis Estadístico, un analista del Subdepartamento de Metodologías e Innovación Estadística y un analista de auditoría o control de procesos.** Además, **se incluirá al subdirector Técnico** que, además de cumplir un rol de arbitraje en caso de no acuerdo, será el responsable por la decisión tomada en esta instancia.
2. En primer lugar, este comité deberá evaluar si es posible realizar ajustes a definiciones adoptadas en las etapas previas. Específicamente, a las medidas de utilidad definidas en la **Etapa 6.4.1: Realizar definiciones previas al proceso de anonimización**, y a los umbrales de riesgo definidos en la **Etapa 6.4.3: Medir y evaluar riesgos**. En otras palabras, deberá estudiar la viabilidad de relajar condiciones en las medidas de utilidad o umbrales de riesgo, que permitan la liberación de los datos. Lo anterior se debe realizar teniendo en consideración no comprometer el cumplimiento de los objetivos de la operación estadística ni los beneficios esperados de los datos para los usuarios.
3. En segundo lugar, este comité deberá elaborar una breve minuta donde se especifique y justifique la decisión adoptada, ya sea para rechazar posibles cambios, que se traduce en la imposibilidad de liberar los datos; o para proponer nuevas definiciones. En caso de proponer nuevas definiciones para medidas de utilidad o umbrales de riesgo, estas se deben documentar como anexo en los reportes de las respectivas etapas.

Notas:

1. Una vez levantadas las alertas de atasco del proceso, **este comité deberá operar en un plazo que no exceda los cinco días hábiles**, a fin de dar respuesta a los usuarios de forma oportuna.
2. En el caso de imposibilidad de liberación de datos, se puede considerar como alternativa, elaborar un conjunto de datos con información más agregada. Por ejemplo, si la publicación de los datos no es posible debido a la desagregación geográfica requerida, digamos región, se puede considerar publicar agregando a macrozonas. Esta alternativa se debe tomar como una sugerencia, y su puesta en práctica depende, por una parte, de la utilidad del dato agregado para los usuarios y, por otra, de los recursos disponibles en el equipo de trabajo a cargo de la operación estadística, para desarrollar un producto publicable alternativo.
3. Las evaluaciones realizadas en esta instancia pueden servir de insumos para actualizar definiciones en el proceso 2. “Diseño y Planificación”.

Actividad 5-5. Elaboración de reporte sobre la evaluación del proceso de anonimización: El reporte de esta etapa se registra en un mismo documento en conjunto con las etapas previas ("Medir y evaluar riesgos" y "Seleccionar y aplicar métodos SDC"). Debe contener al menos los siguientes elementos:

1. Control de versiones: versión, fecha del reporte, descripción de cambio(s) realizado(s), nombres de quienes elaboraron el reporte, nombres de quienes supervisaron y aprobaron el reporte.
2. Antecedentes.
3. Evaluación del resultado de remedición de los riesgos de divulgación.
4. Conclusión del resultado de la evaluación de utilidad del conjunto de microdatos anonimizado.
5. Resumen de incidencias. Documentar hallazgos de la evaluación relacionados con la necesidad de verificar el procesamiento de los métodos SDC o el replanteamiento de los métodos propuestos. Esto debido al incumplimiento de las características estadísticas definidas en la **Actividad 1-3**, al incumplimiento de los umbrales de riesgo definidos para la operación estadística en la **Actividad 3-3**, o a la identificación de nuevas unidades riesgosas como resultado de la anonimización, entre otros.
6. Conclusión sobre la evaluación del proceso SDC y recomendaciones sobre la liberación de datos anonimizados.
7. Las secciones referidas a "Medir y evaluar riesgos" se precisan en la **Actividad 3-4**.
8. Las secciones referidas a "Seleccionar y aplicar métodos SDC" se precisan en la **Actividad 4-2**.

8.3.6. Etapa 6.4.5: Generar reportes y liberar datos

- **Objetivo:** El propósito de esta etapa es la generación de reportes, tanto interno como externo, que acompañan la liberación de datos.

- **Alcance:** Los procedimientos descritos para esta etapa aplican para todas las operaciones estadísticas y productos relacionados cuyo levantamiento de información y/o publicación sea realizado por el INE (encuestas, censos, procesos de múltiples fuentes y registros administrativos) que darán a conocer información al público general u otros usuarios.

- **Exclusiones:** Las exclusiones generales se encuentran listadas en la sección **Exclusiones al alcance**. Además, esta etapa no aplica para operaciones estadísticas cuyos datos anonimizados no son factibles de publicar, de acuerdo con la **Etapa 6.4.4.2: Evaluar proceso SDC**.

- **Palabras claves:** Reporte interno, reporte externo, liberación de microdatos.

En la **Tabla 25** se resume responsabilidades, *inputs* y *outputs* relacionados con la etapa generar reportes y liberar datos.

Tabla 25: Responsables, *inputs* y *outputs* para la etapa generar reportes y liberar datos

Responsable(s)	Analista(s) de anonimización.
Input(s)	- Reporte sobre evaluación del proceso SDC (reevaluación de riesgos y medición de utilidad). - Conclusión sobre si es factible o no la liberación de datos anonimizados. - Datos tratados (anonimizados).
Output(s)	- Datos tratados (versión final para publicación). - Metadatos actualizados. - Reporte interno de proceso SDC. - Reporte externo de proceso SDC.
Control o supervisión	Analista temático, jefatura área o jefe de proyecto.

Fuente: Instituto Nacional de Estadísticas (INE).

Actividad 6-1. Generación de reporte interno: Se debe generar un reporte que va dirigido a los equipos de trabajo no implicados de forma directa con la operación estadística ni con el proceso de anonimización respectivo, y debe considerar lo siguiente:

1. Control de versiones: versión, fecha del reporte, descripción de cambio(s) realizado(s), nombres de quienes elaboraron el reporte, nombres de quienes supervisaron y aprobaron el reporte.
2. Antecedentes.
3. Descripción de las características estadísticas que fueron priorizadas para el desarrollo del proceso y su justificación.
4. Alcances de los análisis a partir de los datos anonimizados. Proporcionar información para un análisis válido de los datos y explicar las limitaciones de los datos como resultado de la anonimización. La pérdida de información debido al proceso de anonimización debe explicarse en detalle a los usuarios para que conozcan los límites de la validez de los datos y sus análisis.
5. Descripción exacta de los métodos de anonimización usados, los parámetros, pero también las medidas de riesgo antes y después de la anonimización.

Nota:
Este reporte interno debe permitir replicar el proceso de anonimización y es importante para los organismos de supervisión, ya que contribuye a garantizar que el proceso de anonimización es suficiente para lograr el anonimato de acuerdo con la legislación aplicable.

Actividad 6-2. Generación de reporte externo: Se debe generar un reporte que va dirigido al usuario final, donde se informa que los datos han sido anonimizados, y debe considerar lo siguiente:

1. Antecedentes.
2. Descripción de las características estadísticas que fueron priorizadas para el desarrollo del proceso y su justificación.
3. Alcances de los análisis a partir de los datos anonimizados. Proporcionar información para un análisis válido de los datos y explicar las limitaciones de los datos como resultado de la anonimización. La pérdida de información debido al proceso de anonimización debe explicarse en detalle a los usuarios para que conozcan los límites de la validez de los datos y sus análisis.
4. Descripción breve de los métodos usados. Esto permite a los investigadores realizar análisis válidos (por ejemplo, cantidad de ruido agregado, matriz de transición para PRAM).

Nota:
Se debe tener cuidado de que esta información no se pueda usar para el re-identificación (por ejemplo, no se libera semillas aleatorias). La ingeniería inversa del proceso debe ser evitada.

Actividad 6-3. Actualización de los metadatos: Los metadatos habituales de la operación estadística (por ejemplo, fichas técnicas, manuales, pesos de la encuesta, estratos, metodología de la encuesta, etc.) deben ser actualizados para cumplir con los datos anonimizados. Las descripciones de las variables o los valores de las etiquetas pueden haber cambiado como resultado del proceso de anonimización.

Nota:
Para la actualización de metadatos, se recomienda el uso del estándar de implementación de la Norma de Documentación y Gestión de Metadatos (NDGM) el cual orienta sobre elementos mínimos exigibles en la etapa posterior al proceso de anonimización y que contribuyen a la estandarización de metadatos fundamentalmente para operaciones estadísticas que provienen de muestreos.

Actividad 6-4. Liberación de microdatos: El conjunto de microdatos anonimizados deben ser liberados bajo la versión PUF.

Notas:
1. Los cambios en las variables realizados en la **Actividad 2-1**, como las variables de fusión, se pueden deshacer para generar un conjunto de datos útil para los usuarios.
2. La liberación del conjunto de microdatos anonimizados debe ir acompañada de los metadatos habituales actualizados de la operación estadística.
3. Una vez que se publican los datos anonimizados, no es posible revocar y publicar otro conjunto de datos del mismo archivo de microdatos. De hecho, esto significaría publicar más de un archivo anonimizado del mismo conjunto de microdatos, ya que algunos usuarios podrían haber guardado el archivo anterior.

9. BIBLIOGRAFÍA

Alfons, A., & Templ, M. (2013). Estimation of Social Exclusion Indicators from Complex Surveys: The R Package laeken. *Journal of Statistical Software, 54*(15), 1-25. Obtenido de http://www.jstatsoft.org/v54/i15/

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., . . . Iannone, R. (2021). rmarkdown: Dynamic Documents for R. R package version 2.10. Obtenido de https://github.com/rstudio/rmarkdown

Arslan, R. (7 de Junio de 2020). Automatic Codebooks from Metadata Encoded in Dataset Attributes. Obtenido de https://github.com/rubenarslan/codebook

Australian Bureau of Statistics. (s.f.). Obtenido de http://www.nss.gov.au/nss/home.nsf/pages/Confidentiality+-+Glossary

Bates, D., Maechler, M., Davis, T., Oehlschlägel, J., & Riedy, J. (1 de Junio de 2021). Sparse and Dense Matrix Classes and Methods. Obtenido de http://Matrix.R-forge.R-project.org/

Benschop, T., Machingauta, C., & Welch, M. (2021). Statistical Disclosure Control: A Practice Guide. 14. Obtenido de https://buildmedia.readthedocs.org/media/pdf/sdcpractice/latest/sdcpractice.pdf

CEPAL, N.U. (2011). Código Regional de Buenas Prácticas en Estadísticas para América Latina y el Caribe. Obtenido de https://repositorio.cepal.org/bitstream/handle/11362/16422/FILE_148023_es.pdf?sequence=1&isAllowed=y

DANE. (2018). *Guía para la anonimización de bases de datos en el Sistema Estadístico Nacional.*

Duncan, G., Elliot, M., & Salzar-González, J. (2011). *Statistical Confidentiality: Principles and Practice.* New York: Springer.

Dupriez, O., & Boyko, E. (2010). *Dissemination of microdata files: Principles, procedures and practices.* International Household Survey Network (IHSN), IHSN Working Paper No. 005. Obtenido de http://www.ihsn.org

Eurostat. (2020). European Statistical System handbook for quality and metadata reports. Obtenido de https://ec.europa.eu/eurostat/documents/3859598/10501168/KS-GQ-19-006-EN-N.pdf/

INE. (2015). Código de Buenas Prácticas para las Estadísticas Chilenas.

INE. (Marzo de 2020). Fundamentos del Estándar para la evaluación de la calidad de las estimaciones en encuestad de hogares.

Ministerio Secretaría General de la Presidencia. (2020). Ley N° 19628 Artículo 2. *Protección de la Vida Privada Diario Oficial de la Republica de Chile*. Santiago, República de Chile.

OCDE. (s.f.). *Glossary of Statistical Terms*. Recuperado el 22 de Enero de 2021, de https://stats.oecd.org/glossary/search.asp

R Core Team. (2019). *R: A language and environment for statistical computing.* Viena, Austria: R Foundation for Statistical Computing. Obtenido de https://www.R-project.org/

Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. *Journal of Statistical Software, 67*(4), 1-36. doi:10.18637/jss.v067.i04

UNECE. (2017). Quality Indicators for the Generic Statistical Business Process Model (GSBPM) - For Statistics derived from Surveys and Administrative Data Sources, version 2.0.

UNECE. (2019). *Generic Statistical Business Process Model. GSBPM (Versión 5.1).*

United Nations. (2014). *Principios Fundamentales de las Estadísticas Oficiales*. Obtenido de unstats.un.org

Ushey, K. (2021). Project Environments. 1-60. Obtenido de https://rstudio.github.io/renv/

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag. Obtenido de https://ggplot2.tidyverse.org.

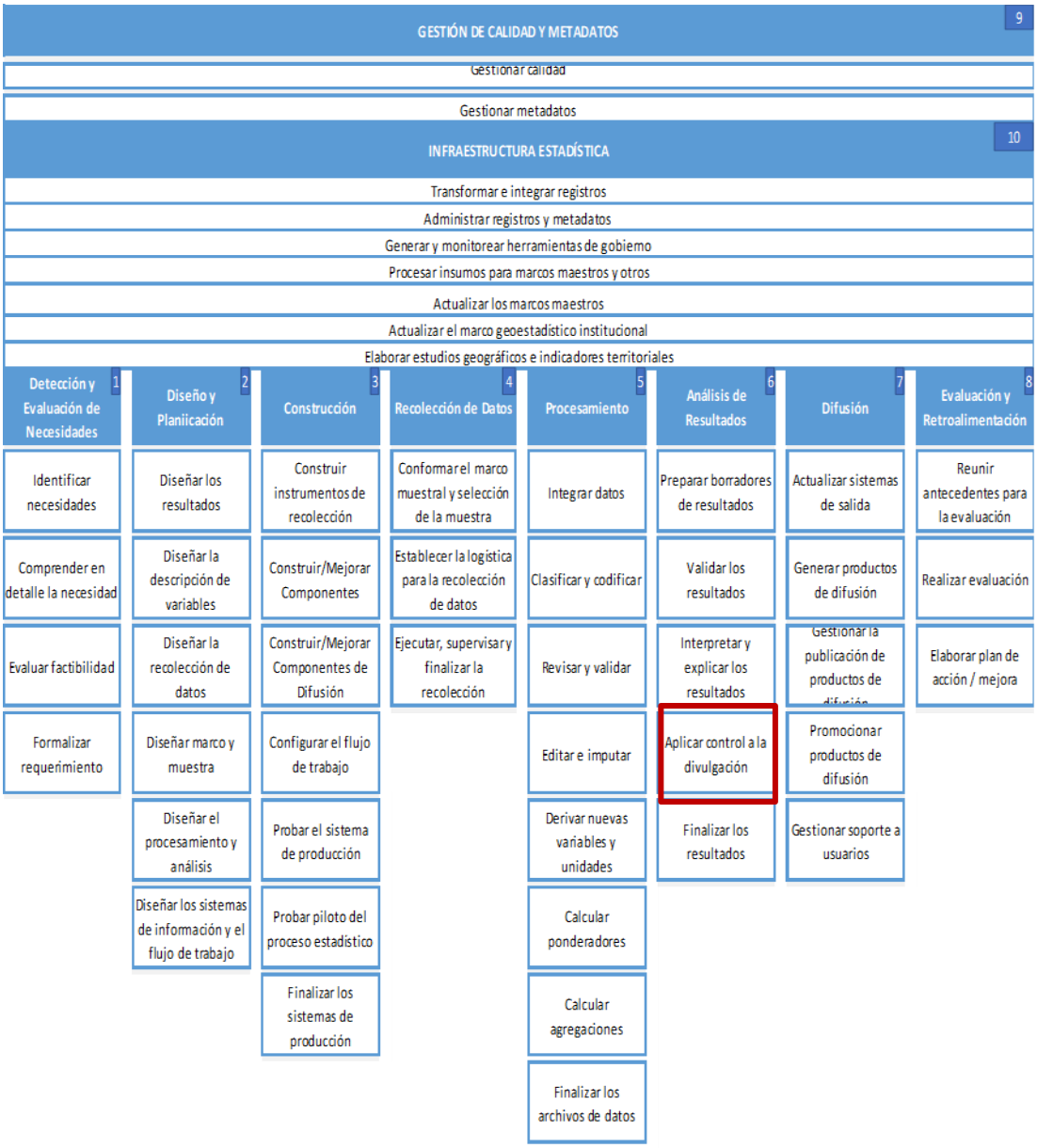
Wickham, H., & Francois, R. (2015). dplyr: A Grammar of Data Manipulation. R Package Version 0.4.3. Obtenido de http://CRAN.R-project.org/package=dplyr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43). doi:10.21105/joss.01686

Yazdani, A. (2015). Statisticial Confidentiality and Disclosure Control Textbook. Organisation of Islamic Cooperation, Statistical Economic and Social Research and Training Centre for Islamic Countries SESRIC.

10. ANEXOS

10.1. Mapa de procesos-Segmento de Negocio



Fuente: Instituto Nacional de Estadísticas (INE), desde **Intranet INE** [20]

[20] Ver https://inechile.sharepoint.com/sites/Intranet/departamentodegestionestrategica/Mprocesos/Paginas/Segmento-Negocio.aspx

10.2. Simbología diagramación en Bizagi

Símbolo	Nombre	Descripción
	Evento de inicio	Representa el inicio de un proceso.
	Evento de fin	Representa el fin de un proceso.
	Evento intermedio	Representa un evento que puede ocurrir durante el proceso.
	Eventos de enlace	Este evento se utiliza para conectar etapas. El que se encuentra arriba es el evento que "lanza" y se le asigna el nombre del lugar de destino. El que se encuentra abajo es el que "recibe" y se le asigna el nombre del lugar de procedencia.
	Compuerta exclusiva	Ocurre cuando en un punto del proceso se escoge un solo camino de varios disponibles. También se utiliza para unir caminos alternativos.
	Compuerta paralela	Ocurre cuando en un punto del proceso se siguen dos o más caminos simultáneamente. También se utiliza para unir caminos alternativos que deben unirse todos antes de continuar.
	Compuerta inclusiva	Ocurre cuando en un punto del proceso se sigue alguno de los caminos disponibles.
	Subproceso	Se utiliza para graficar actividades a través de compuertas, eventos y flujos de secuencia.

Fuente: Instituto Nacional de Estadísticas (INE).

10.3. Familias de cargo

Familias de cargo
Directivos[21]
Jefes de Departamento
Jefes de Subdepartamento
Coordinadores
Supervisores
Analistas Especialistas
Analistas
Supervisores Operativos
Operativos
Asistentes
Auxiliares

Fuente: Instituto Nacional de Estadísticas (INE).

[21] Los Directivos no forman parte de la familia de cargos definida por la institución, sin embargo, con el objetivo de identificar aquellos roles que participan dentro del flujo de este manual de procedimiento se creó esta clasificación.