Guía para el Control de Divulgación Estadística en Microdatos

Subdepartamento de Investigación Estadística

2023-09-07

Índice

Capítulo 1

Prefacio

Los datos son un recurso valioso que proporciona información crítica para estadísticos, científicos sociales y científicos de datos. Estos datos se utilizan para generar perspectivas detalladas y oportunas que responden a las necesidades de información de una amplia gama de partes interesadas.

En un mundo donde cada vez más grandes volúmenes de datos provienen de un número creciente de proveedores, las Oficinas Nacionales de Estadística (ONE) están utilizando enfoques innovadores para mantener estándares y definiciones de datos, sistemas de gestión de privacidad y confidencialidad, e intercambio responsable de datos.

Las ONE tienen un papel de liderazgo que desempeñar en el establecimiento de formas seguras y transparentes de compartir datos, experiencias y mejores prácticas para respaldar el uso de datos con fines de prueba, evaluación, educación y desarrollo. Con la integridad y la confidencialidad de los datos a la vanguardia, las ONE están posicionandose cada vez más para proporcionar herramientas, métodos y enfoques para promover el intercambio responsable de datos, a fin de satisfacer las necesidades de un número creciente de partes interesadas en este ámbito que está en constante cambio.

Las ONE reconocen que se debe cumplir con el llamado a una mayor apertura y transparencia de los datos. Sin embargo, también se comprometen a proteger la confidencialidad y la privacidad integradas en sus tenencias de datos.

Se reconoce ampliamente que la difusión de información que es a la vez útil y completamente segura, no puede lograrse en su totalidad. Por lo tanto, la seguridad es un concepto relativo, no absoluto, y se debe entender como una métrica, no un estado.

Es dentro de este contexto que las ONE deben establecer protocolos para la difusión segura de datos y métricas para medir la utilidad de la información

estadística publicada y el grado de protección de las unidades, ya sean personas naturales o jurídicas, e información recopilada de la cual se deriva.

Esta guía es para aquellos que trabajan en una ONE u oficina estatal que están involucrados en la gestión del acceso a datos estadísticos, y que deseen explorar herramientas de protección de datos para que los usuarios accedan a ellos. La guía destaca algunas aplicaciones exitosas recientes de control a la divulgación estadística en microdatos en el Instituto Nacional de Estadísticas de Chile (INE), y presenta un marco general sobre medición y evaluación de riesgos, técnicas para generar datos anonimizados, y sobre las medidas de utilidad que se pueden usar para evaluar qué tan bien los datos anonimizados satisfacen las necesidades analíticas de los usuarios. La guía también incluye recomendaciones sobre qué enfoques utilizar en diferentes situaciones, así como consejos prácticos y recursos para que los profesionales comiencen su experiencia en la implementación de proceso de control a la divulgación estadística.

Esta guía se basa en Statistical Disclosure Control: A Practice Guide (Benschop, Machingauta, y Welch, 2021) y en la Guía para el control de divulgación estadística en microdatos, elaborada por el INE en 2021, que es el primer esfuerzo en esta materia en instituciones del Estado en Chile. Sin embargo, aún puede enriquecerse a partir de nuevos conocimientos y experiencias que tanto investigadores como profesionales puedan aportar.

¡Esperamos que esta guía lo ayude en su viaje hacia la implementación de procesos de control a la divulgación estadística en su organización!

1.1 Autores de esta guía

Jonathan González Mejías, Subdepartamento de Estadísticas Socioeconómicas, Subdirección Técnica en Instituto Nacional de Estadísticas de Chile.

José Bustos Melo, Subdepartamento de Investigación Estadística, Departamento de Metodologías e Innovación Estadística en Instituto Nacional de Estadísticas de Chile.

Julio Guerrero Rojas, Subdepartamento de Investigación Estadística, Departamento de Metodologías e Innovación Estadística en Instituto Nacional de Estadísticas de Chile.

Lissette Bastías Navarro, Subdepartamento de Calidad y Estándares, Departamento de Metodologías e Innovación Estadística en Instituto Nacional de Estadísticas de Chile.

Nicolás Berhó Montalvo, Subdepartamento de Estadísticas de Condiciones de Vida, Subdirección Técnica en Instituto Nacional de Estadísticas de Chile.

1.2 Reconocimientos

Aquí escribir si se quiere agradecer o reconocer la colaboración de alguna persona o equipo.

Capítulo 2

Introducción

Como Instituto Nacional de Estadística (en adelante, INE) tenemos la responsabilidad de la recopilación y difusión de estadísticas oficiales, tomando resguardos para cumplir con la Ley de Secreto Estadístico (Art.29, Ley 17.374 (**?**)), la Ley sobre Protección de la Vida Privada (Art.2e, Ley 19.628 (**?**)) y la legislación propia de las entidades públicas, todas en la línea de la protección y privacidad de la información difundida. Por otro lado, a nivel país, en los últimos años, ha existido un aumento constante en transparentar y disponer información tanto a nivel privado como público, mediante la "ley de transparencia" (Ley 20.285 (**?**)), promulgada en año 2008.

Es en esta misma línea que 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 (**?**).

En paralelo la comunidad estadística ha reconocido la importancia de asegurar la información para mantener la confianza de las poblaciones a las que servimos. En este sentido, el Código Nacional de Buenas Prácticas Estadísticas del 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" (**?**).

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, por su sigla 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.

La experiencia del INE en términos de control de divulgación estadística ha ido

avanzando, iniciando en junio del año 2009, bajo 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, luego en 2019, un equipo multidisciplinario de la producción estadística institucional, define los lineamientos para desarrollar un proceso estandarizado de control de divulgación en las operaciones estadísticas que desarrolla el INE, entregando como resultado una primera versión de la "Guía para el control de divulgación estadística en microdatos". En diciembre del 2021 se transforma en un estándar institucional disponible en la página web institucional https://www.ine.gob.cl/calidad-estadistica/directrices-metodologicas (**?**).

Este documento 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 de interés y que solicitan las bases de microdatos, al mismo tiempo de tener procedimientos estandarizados en la producción de estadísticas oficiales.

Esta guía busca brindar pasos prácticos bajo lineamientos institucionales para aquellas operaciones estadísticas que requieran desbloquear el acceso a sus datos de manera segura y garantizar que los datos sigan siendo aptos para su propósito.

2.1 Estableciendo una base de conocimiento

La publicación de datos es importante, ya que permite a los investigadores y responsables políticos replicar los resultados publicados oficialmente, generar nuevos conocimientos sobre los problemas, evitar la duplicación de encuestas y proporcionar mayores retornos a la inversión en el proceso de encuesta.

Tanto la producción de informes, con tablas agregadas de indicadores y estadísticas, como la publicación de microdatos resultan en desafíos de privacidad para el productor. En el pasado, para muchas ONE, el único requisito era publicar un informe y algunos indicadores clave. El reciente movimiento en torno a los datos abiertos, el gobierno abierto y la transparencia significa que las ONE están bajo una mayor presión para liberar sus microdatos, para permitir un uso más amplio de los datos recopilados a través de fondos públicos. Esta guía se centra en los métodos y procesos para la liberación de microdatos, ya sea que estos provengan de encuestas, censos o registros estadísticos generados por el INE. Por tanto, el alcance de los procesos que se describen en esta guía se ciñe a proveer directriz circunscrita al campo de los microdatos, por lo que se excluyen los procesos de control de divulgación estadística orientados a tabulados, estadísticas geoespaciales, publicaciones web o visualizaciones de mapas, etc., que requieren enfoques diferentes al propuesto en esta guía. Asimismo, se distingue la necesidad de establecer lineamientos para el control de divulgación estadística en la publicación de tablas y publicaciones web, con el fin de cubrir más ámbitos de la producción estadística del INE.

Se requiere la difusión de datos de manera segura para proteger la integridad del sistema estadístico, al garantizar que el INE cumpla con su compromiso con los encuestados de proteger su identidad. Las ONE no comparten ampliamente, en detalle sustancial, su conocimiento y experiencia usando SDC y los procesos para crear datos seguros con otras ONE. Esto lo hace difícil para las instituciones nuevas en el proceso para implementar soluciones. Para llenar esta brecha de experiencia y conocimiento, el equipo de la mesa de trabajo INE (en adelante mesa) evaluó el uso de un amplio conjunto de métodos de SDC en una gama de microdatos de encuestas que cubren importantes temas de desarrollo relacionados con trabajo, seguridad ciudadana, empresas de ferrocarriles, trámites de circulación. Dado que sus productores ya habían tratado estos datos, no era posible, ni era objetivo de la mesa, emitir un juicio sobre la seguridad de estos datos, los cuales son de dominio público. El enfoque se centró más bien en medir los efectos que varios de los métodos tendrían que ver con la relación riesgo – utilidad para los microdatos producidos para medir indicadores comunes de desarrollo. La experiencia de esta experimentación es útil para informar la discusión de los procesos y métodos en esta guía.

2.2 Propósito de esta guía

Esta guía tiene como propósito presentar los lineamientos para la aplicación del control de divulgación estadística en microdatos derivados de censos, registros estadísticos y encuestas por muestreo desarrollados por el INE, permitiendo establecer qué microdatos pueden ser liberados y bajo qué condiciones.

Esta guía no pretende prescribir o abogar por cambios en los métodos que los productores de datos específicos ya están utilizando y que han diseñado para ajustarse y cumplir con sus políticas de difusión de datos existentes, empero, ordenarlos. Los métodos discutidos en esta guía provienen de una gran cantidad de literatura sobre SDC. 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 ONE con experiencia en la preparación de microdatos para su publicación.

Siempre que sea posible, para cada método y tema, se proporciona ejemplos elaborados, referencias al trabajo original o seminal que describe los métodos y algoritmos en detalle y las lecturas recomendadas. Esto, cuando se combina con la discusión del método y las consideraciones prácticas en esta guía, debería permitir al lector comprender los métodos y sus fortalezas y debilidades. También proporciona suficientes detalles para que los lectores usen una solución de *software* adecuada para implementar los métodos.

Para los ejercicios de esta guía, se ha utilizado el paquete de código abierto y gratuito para SDC llamado `sdcMicro`, así como el lenguaje y entorno de programación estadístico `R`. `sdcMicro` es un paquete adicional para el lenguaje `R`. El paquete fue desarrollado y es mantenido por Matthias Templ, Alexander Kowarik

y Bernhard Meindl[1]. El lenguaje estadístico `R` y el paquete `sdcMicro`, así como cualquier otro paquete necesario para el proceso SDC, están disponibles gratuitamente en los `mirrors` de la Red Integral de Archivos `R` (CRAN[2]) (http://cran.r-project.org/). El lenguaje está disponible para los sistemas operativos Linux, Windows y Macintosh. Se ha elegido usar `R` y `sdcMicro` porque está disponible gratuitamente, admite todos los formatos de datos principales y es fácil de adaptar por el usuario. El Banco Mundial, a través de IHSN[3], también ha proporcionado fondos para el desarrollo del paquete `sdcMicro` para garantizar que cumpla con los requisitos de las ONE.

Esta guía no proporciona una revisión de todos los demás paquetes disponibles para implementar el proceso SDC, pues se trata más de proporcionar información práctica sobre la aplicación de los métodos. Sin embargo, cabe destacar otro paquete de *software* en particular que las ONE utilizan comúnmente: `-ARGUS`[4]. `-ARGUS` es desarrollado por Statistics Netherlands. `sdcMicro` y `-ARGUS` son ampliamente utilizados en oficinas de estadística en la Unión Europea e implementan muchos de los mismos métodos.

Las necesidades de usuario acerca de algún conocimiento de `R` para usar `sdcMicro` está más allá del alcance de esta guía, así como enseñar el uso de `R`, pero se presenta una serie de estudios de casos que incluyen el código para el anonimato de una serie de conjuntos de datos de demostración con `R`. A través de estos estudios de caso, se demuestra una serie de enfoques para el proceso de anonimización en `R`.

2.3 Esquema de esta guía

Esta guía está dividida en las siguientes secciones principales:

1. **Introducción a sdcMicro**: donde se visualiza la necesidad de aplicar los métodos SDC y el trade off que se produce entre el riesgo versus la utilidad.

2. **Tipos de liberación de datos**: en este apartado encontrarán los tres tipos de métodos de divulgación, archivos de uso público (PUF, por sus siglas en inglés), archivos de uso científico (SUF, por sus siglas en inglés) y microdatos disponibles en un centro de datos de investigación controlado.

3. **Medición de riesgos**: las medidas de riesgo que se utilizan y la determinación si un archivo de datos es lo suficientemente seguro para su divulgación.

4. **Métodos SDC**: una descripción de los métodos más utilizados para anonimizar.

5. **Medición de utilidad y pérdida de información**: en este apartado se profundiza acerca del trade off entre la medición de la utilidad y la pérdida de información.

6. **Procesos SDC INE 2021**: caso práctico implementado en el INE en la mesa de anonimización institucional.

7. **Caso de estudio: Enusc**: caso práctico para aplicar el método SDC en la Encuesta Nacional Urbana de Seguridad Ciudadana (ENUSC) con datos sintéticos.

[1] Matthias Templ, Alexander Kowarik, Bernhard Meindl (2015). Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. Journal of Statistical Software 67 (October): 1–36. https://doi.org/10.18637/jss.v067.i04.

[2] En inglés, Comprehensive R Archive Network.

[3] En inglés, International Household Survey Network.

[4] μ- ARGUS está disponible en: https://research.cbs.nl/casc/mu.htm

Capítulo 3

Acrónimos y glosario

3.1 Acrónimos

[1] En inglés, *Generic Statistical Business Process Model*.
[2] En inglés, *International Household Survey Network*.
[3] En inglés, *Post Randomization Method*.
[4] En inglés, *Public Use File*.
[5] En inglés, *Statistical Disclosure Control*.
[6] En inglés, *Statistics Canada*.
[7] En inglés, *Scientific Use File*.
[8] En inglés, *United Nations Economic Commission for Europe*.

3.2 Glosario

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

Tabla 3.1: Lista de acrónimos

Acrónimo	Descripción
AEPD	Agencia Española de Protección de Datos
Bloque	Trozo de código en R que permite cargar y procesar datos, realizar los análisis estad
CEPAL	Comisión Económica para América Latina y el Caribe
DANE	Departamento Administrativo Nacional de Estadística
ENUSC	Encuesta Nacional Urbana de Seguridad Ciudadana
FCYTE	Fundación Española de Ciencia y Tecnología
GSBPM [1]	Modelo Genérico del Proceso Estadístico
IHSN [2]	Red Internacional de Encuestas de Hogares
INE	Instituto Nacional de Estadísticas
INEGI	Instituto Nacional de Estadística y Geografía
MINSEGPRES	Ministerio Secretaría General de la Presidencia
OCDE	Organización para la Cooperación y el Desarrollo Económicos
ONE	Oficina Nacional de Estadística
PITEC	Panel de Innovación tecnológica
PRAM [3]	Método de Post-Aleatorización
PUF [4]	Archivo de Uso Público
RUT	Rol Único Tributario
ROL	Identifica a una propiedad o bien raíz
SEN	Sistema Estadístico Nacional
SDC [5]	Control de Divulgación Estadística
sdcMicro	Paquete de implementación bajo el *software* R
STATCAN [6]	Estadísticas de Canadá
SUF [7]	Archivo de Uso Científico
UNECE [8]	Comisión Económica de las Naciones Unidas para Europa

Tabla 3.2: 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.	[@benschop2021], 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 propiedades estadísticas en los resultados.	[@institutonacionaldeestadisticas2022]
Archivo de datos para uso científico	Archivo de uso científico (SUF, por su sigla en inglés, Scientific Use File), 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".	[@benschop2021], 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	Adaptado de [@benschop2021], pág. 20

Capítulo 4

Tipos de liberación de datos

Esta sección expone sobre la liberación de microdatos, cuyos lineamientos se extrajeron de la guía elaborada por el Banco Mundial (**?**), que a su vez recoge el trabajo conjunto realizado por el Banco Mundial y sus socios en la Red Internacional de Encuestas de Hogares IHSN[1] (**?**).

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 **??** 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.

4.1 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",

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

Tabla 4.1: Resumen de tipos de liberación de microdatos

Tipo	Descripción	Aplicabilidad 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	No aplicable.

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.

Capítulo 5

Proceso SDC: Una introducción

5.1 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 (**?**). 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 resguardar 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 (**?**).

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 (**?**). 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"(**?**, p. 33).

5.2 Balance 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 **??**, 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.

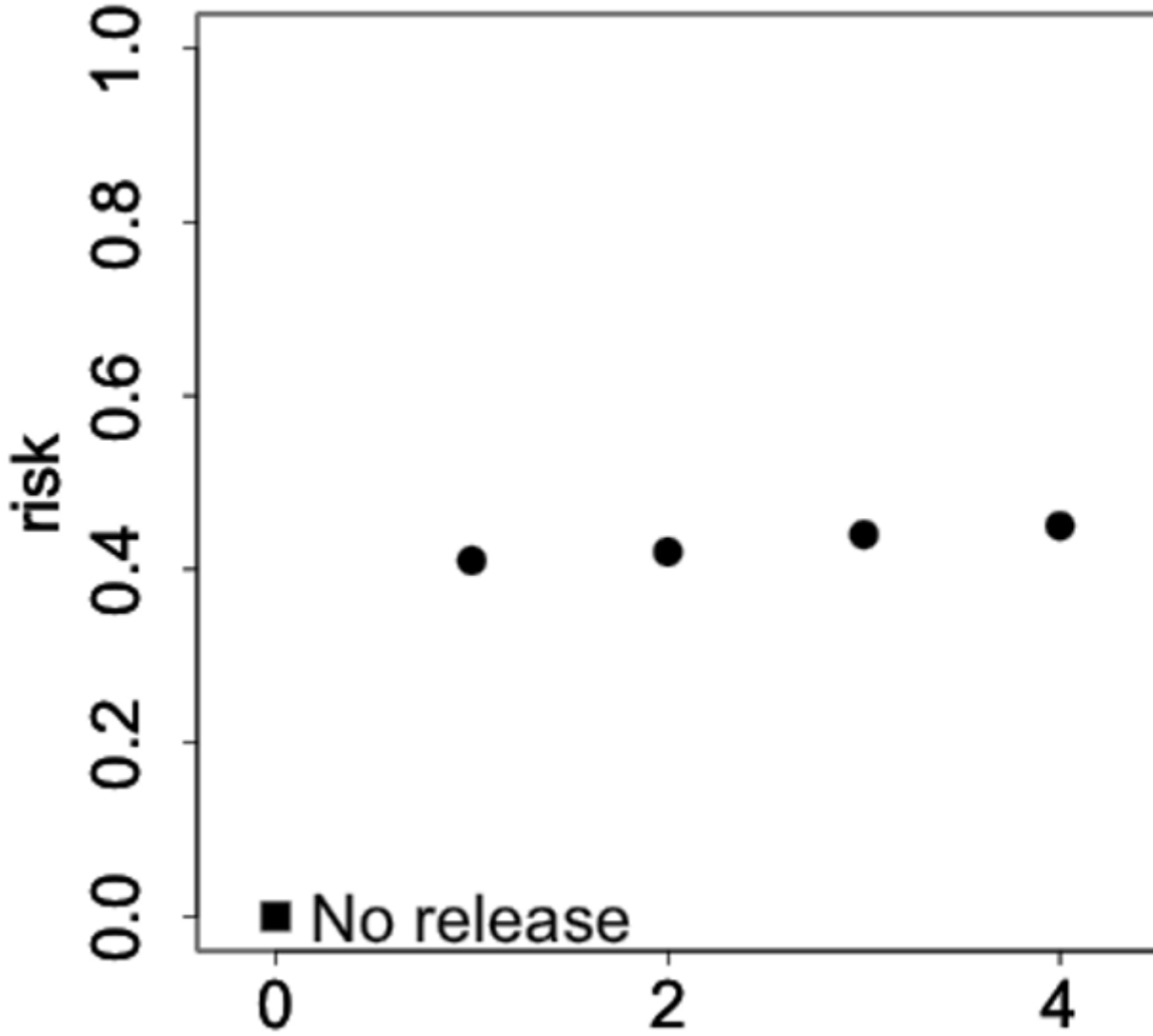

Figura 5.1: Balance Riesgo-Utilidad en un conjunto de datos. Imagen extraída de [@benschop,p.15].

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 los 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.

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 sección Tipos de liberación de datos). 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 [1] . 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.

[1] Esto no aplica en el caso del INE DE chile, pues solo es aplicable, según el marco legal vigente, la difusión de datos mediante formato PUF.

Capítulo 6

Introducción a sdcMicro

6.1 Introducción

El paquete `R sdcMicro ?` sirve para evaluar y anonimizar conjuntos de microdatos confidenciales, facilita el manejo de métodos SDC mediante una implementación de clase S4 orientada a objetos. Incluye todos los métodos populares de perturbación y riesgo de divulgación. El paquete realiza un nuevo cálculo automático de recuentos de frecuencia, medidas de riesgo individuales y globales, pérdida de información y estadísticas de utilidad de datos después de cada paso de anonimización. Todos los métodos están altamente optimizados en términos de costos computacionales para poder trabajar con grandes conjuntos de datos. Los profesionales también pueden utilizar fácilmente las funciones de generación de informes que resumen el proceso de anonimización. Describimos el paquete y demostramos su funcionalidad con un complejo conjunto de datos de prueba procedente de encuestas de hogares, que ha sido distribuido por la Red Internacional de Encuestas de Hogares.

Para más información ver https://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf

6.2 Instalación de `R`, `sdcMicro` y otros paquetes

Esta guía se basa en el paquete de *software* `sdcMicro`, que es un paquete adicional para el lenguaje de programación estadístico `R`. Tanto `R` como `sdcMicro`, así como otros paquetes de `R`, están disponibles gratuitamente en el sitio web de CRAN (Comprehensive R Archive Network) para Linux, Mac y Windows (http://cran.r-project.org). Este sitio web también ofrece descripciones de paquetes. Además de la versión estándar de `R`, existe una interfaz de usuario

más fácil de usar para R: RStudio. RStudio también está disponible gratuitamente para Linux, Mac y Windows (http://www.rstudio.com). El paquete sdcMicro tiene dependencia de otros paquetes R que deben instalarse en su computadora antes de usar sdcMicro. Se instalarán automáticamente al instalar sdcMicro. Para algunas funcionalidades, usamos otros paquetes (como foreign para leer datos y algunos paquetes gráficos). Si es así, esto se indica en la sección correspondiente de esta guía. R, RStudio, el paquete sdcMicro y sus dependencias y otros paquetes tienen actualizaciones periódicas. Se recomienda encarecidamente comprobar periódicamente si hay actualizaciones: esto requiere instalar una nueva versión para una actualización de R; con el comando update.packages() o usando las opciones de menú en R o RStudio se pueden actualizar los paquetes instalados.

Al iniciar R o RStudio, es necesario especificar cada vez qué paquetes se están utilizando cargándolos. Esta carga de paquetes se puede realizar con la función library() o require(). Ambas opciones se ilustran en el Bloque **??**.

Bloque 6.1. Cargando paquetes requeridos

```
library(sdcMicro) # cargando paquete sdcMicro
require(sdcMicro) # cargando paquete sdcMicro
```

Todos los paquetes y funciones están documentados. La forma más fácil de acceder a la documentación de una función específica es usar la ayuda integrada, que generalmente brinda una descripción general de los parámetros de las funciones, así como algunos ejemplos. La ayuda de una función específica se puede llamar con un signo de interrogación seguido del nombre de la función sin ningún argumento. El Bloque **??** muestra cómo llamar al archivo de ayuda para la función microaggregation() del paquete sdcMicro [1]. La página de descarga de cada paquete en el sitio web de CRAN también proporciona un manual de referencia con una descripción completa de las funciones del paquete.

Bloque 6.2. Visualización de ayuda para funciones

```
?microaggregation # ayuda para la función microagregación
```

Cuando se encuentran problemas o errores en el paquete sdcMicro, se pueden publicar comentarios o sugerencias para los desarrolladores de sdcMicro en su GitHub (https://github.com/sdcTools/sdcMicro/issues).

[1] A menudo, también es útil buscar ayuda en Internet sobre funciones específicas en R. Hay muchos foros donde los usuarios de R discuten los problemas que encuentran. Un sitio particularmente útil es *stackoverflow.com*

Tabla 6.1: Paquetes y funciones para lectura de datos en 'R'

Tipo/software	Extensión	Paquete	Función
SPSS	.sav	'haven'	'read_sav()'
STATA (v.5-14)	.dta	'haven'	'read_dta()'
SAS	.sas7bdat	'haven'	'read_sas()'
Excel	.csv	'utils' (paquete base)	'read_csv()'
Excel	.xls/.xlsx	'readxl'	'read_xlsx()'

6.3 Leer funciones en R

El primer paso en el proceso SDC cuando se usa `sdcMicro` es leer los datos en `R` y crear un marco de datos[2] `R` es compatible con la mayoría de los formatos de datos estadísticos y proporciona funciones de lectura para la mayoría de los tipos de datos. Para esas funciones de lectura, a veces es necesario instalar paquetes adicionales y sus dependencias en `R`. En la Tabla **??** se proporciona una descripción general de los formatos de datos, las funciones y los paquetes que contienen estas funciones . Estas funciones también están disponibles como escritura (por ejemplo, `write_dta()`) para guardar los datos anónimos en el formato requerido [3].

La mayoría de estas funciones tienen opciones que especifican cómo manejar los valores faltantes y las variables con niveles de factor y etiquetas de valor. El Bloque **??**, el Bloque **??** y el Bloque **??** proporcionan código de ejemplo para leer un archivo STATA (.dta), un archivo de valores separados por ; (.csv) y un archivo SPSS (.sav), respectivamente.

Bloque 6.3. Lectura en un archivo STATA

```
setwd("../Capacitación/GitHub") # directorio de trabajo


fname = "data.dta" # nombre del archivo
library(haven) # carga el paquete requerido para la función de lectura/escritura
               # para archivos STATA
file <- read_dta(fname)
# lee los datos en el marco de datos tbl llamado file
```

Bloque 6.4. Lectura en un archivo csv

[2] Un marco de datos es una clase de objeto en `R`, que es similar a una tabla o matriz de datos

[3] No todas las funciones son compatibles con todas las versiones del paquete de *software* respectivo. Nos referimos a los archivos de ayuda de las funciones de lectura y escritura para más información.

```
setwd("../Capacitación/GitHub") # directorio
                                                                    # de
                                                                    # trabaj

fname = "data.csv" # nombre del archivo
file <- read.csv(fname, header = TRUE, sep = ";", dec = ".")
# lee los datos hacia un dataframe llamado file,
# la primera línea contiene los nombres de las variables,
# campos son separados con comas, posiciones decimales se indican con ';'
```

Bloque 6.5. Lectura en un archivo SPSS

```
setwd("../Capacitación/GitHub") # directorio
                                                                    # de
                                                                    # trabaj

fname = "data.sav" # nombre del archivo
library(haven) # carga paquete requerido para la función lectura/escritura
               # para archivos SPSS
file <- read_sav(fname)
# lee los datos hacia un dataframe llamado file
```

El tamaño máximo de datos en R está técnicamente restringido. El tamaño máximo depende de la versión R (32 o 64 bits) y del sistema operativo. Algunos métodos SDC requieren largos tiempos de cálculo para grandes conjuntos de datos (consulte la Sección Tiempo de cómputo).

6.4 Valores faltantes

La forma estándar en que los valores faltantes se representan en R es mediante el símbolo NA, que es diferente a los valores imposibles, como la división por cero o el logaritmo de un número negativo, que se representan con el símbolo NaN. El valor NA se usa tanto para variables numéricas como categóricas[4]. Los valores suprimidos por la rutina `localSuppression()` también se reemplazan por el símbolo NA. Algunos conjuntos de datos y *software* estadístico pueden usar diferentes valores para los valores faltantes, como '999' o cadenas. Es posible incluir argumentos en las funciones de lectura para especificar cómo se deben tratar los valores faltantes en el conjunto de datos y recodificar automáticamente los valores faltantes a NA. Por ejemplo, la función `read.table()` tiene el argumento `na.strings`, que reemplaza las cadenas especificadas con valores NA.

Los valores faltantes también se pueden recodificar después de leer los datos en R. Esto puede ser necesario si hay varios códigos de valores perdidos diferentes

[4]Esto es independientemente de la clase de la variable en R. Consulte la sección Clases en R para obtener más información sobre las clases en R.

en los datos, códigos de valores perdidos diferentes para diferentes variables o la función de lectura para el tipo de datos no permite especificar los códigos de valores perdidos. Al preparar los datos, es importante volver a codificar cualquier valor faltante que no esté codificado como NA a NA en R antes de iniciar el proceso de anonimización para garantizar la medición correcta del riesgo (por ejemplo, k-anonimato), así como para asegurar que muchos de los métodos se aplican correctamente a los datos. El Bloque **??** muestra cómo recodificar el valor '99' a NA para la variable "TOILET".

Bloque 6.6. Recodificación de valores perdidos a NA

```
file[file[,'TOILET'] == 99,'TOILET'] <- NA
# Recodificar el código de valor faltante 99 a NA para la variable TOILET
```

6.5 Clases en R

Todos los objetos en R son de una clase específica, como un número entero, un carácter, una matriz, un factor o un marco de datos. La clase de un objeto es un atributo que hereda de la clase base, haciéndolo miembro e instancia de esta clase. Para averiguar la clase de un objeto, se puede utilizar la función `class()`. Las funciones en R pueden requerir objetos o argumentos de ciertas clases o funciones que pueden tener una funcionalidad diferente según la clase del argumento. Algunos ejemplos son las funciones de escritura que requieren marcos de datos y la mayoría de las funciones en el paquete `sdcMicro` que requieren marcos de datos u objetos `sdcMicro`. La funcionalidad de las funciones en el paquete `sdcMicro` difiere para marcos de datos y objetos `sdcMicro`. Es fácil cambiar el atributo de clase de un objeto con funciones que comienzan con "as.", seguido del nombre de la clase (por ejemplo, `as.factor()`, `as.matrix()`, `as.data.frame()`). El Bloque **??** muestra cómo verificar la clase de un objeto y cambiar la clase a "data.frame". Antes de cambiar el atributo de clase del objeto "file", estaba en la clase "matrix". Una clase importante definida y utilizada en el paquete `sdcMicro` es la clase denominada `sdcMicroObj`. Esta clase se describe en la siguiente sección.

Bloque 6.7. Cambiando la clase de un objeto en R

```
# Averiguar la clase del objeto 'file'
class(file)
"matrix"

# Cambiar la clase al marco de datos (data frame)
file <- as.data.frame(file)

# Comprobando la clase del resultado (file)
"data.frame"
```

6.6 Objetos de la clase `sdcMicroObj`

El paquete `sdcMicro` se basa en objetos [5] de la clase `sdcMicroObj`, una clase especialmente definida para el paquete `sdcMicro`. Cada componente de esta clase tiene una estructura determinada con elementos que contienen información sobre el proceso de anonimización (consulte la Tabla **??** para obtener una descripción de todos los elementos o propiedades (*slots*, en inglés)). Antes de evaluar el riesgo y la utilidad y aplicar métodos SDC, se recomienda crear un objeto de clase `sdcMicro`. Todos los ejemplos de esta guía se basan en estos objetos. La función utilizada para crear un objeto `sdcMicro` es `createSdcObj()`. La mayoría de las funciones en el paquete `sdcMicro`, como `microaggregation()` o `localSuppression()`, usan automáticamente la información requerida (por ejemplo, identificadores indirectos, pesos de muestra) del objeto `sdcMicro` si se aplica a un objeto de clase `sdcMicro`.

Los argumentos de la función `createSdcObj()` permiten especificar el archivo de datos original y categorizar las variables en este archivo de datos antes del inicio del proceso de anonimización.

En el Bloque **??**, mostramos todos los argumentos de la función `createSdcObj()`, y primero definimos vectores con los nombres de las diferentes variables. Esta práctica brinda una mejor visión general y luego permite cambios rápidos en las opciones de variables si es necesario. Elegimos los identificadores indirectos categóricos (keyVars); las variables vinculadas a los identificadores indirectos categóricos que necesitan el mismo patrón de supresión (ghostVars, consulte la sección Supresión local); los identificadores indirectos numéricos (numVars); las variables seleccionadas para aplicar PRAM (pramVars); una variable con pesos muestrales (weightVar); el identificador de agrupación (hhId, por ejemplo, un identificador de hogar, consulte la sección Riesgo jerárquico (o del hogar)); una variable que especifica los estratos (strataVar) y las variables sensibles especificadas para el cálculo de *l-diversity* (sensibleVar, consulte la sección l-diversity).

La mayoría de los métodos SDC en el paquete `sdcMicro` se aplican automáticamente dentro de los estratos, si se especifica el argumento 'strataVar'.

Los ejemplos son la supresión local y PRAM. No se deben especificar todas las variables, por ejemplo, si no hay una estructura jerárquica (hogar), se puede omitir el argumento 'hhId'. Los nombres de las variables corresponden a los nombres de las variables en el marco de datos que contiene los microdatos a anonimizar. La selección de variables es importante para las medidas de riesgo que se calculan automáticamente. Además, varios métodos se aplican por defecto a todas las variables de un tipo, por ejemplo, microagregación a todas las variables clave [6]. Después de seleccionar estas variables, podemos crear el

[5] La clase `sdcMicroObj` tiene objetos S4, que tienen elementos o atributos y permiten la programación orientada a objetos.

[6] A menos que se especifique lo contrario en los argumentos de la función.

objeto sdcMicro. Para obtener un resumen del objeto, es suficiente escribir el nombre del objeto.

Bloque 6.8. Seleccionando variables y creando un objeto de clase `sdcMicroObj` para el proceso SDC en `R`

```
# Seleccionar variables para crear objeto sdcMicro

# Selección de variables categóricas
selectedKeyVars <- c('URBRUR', 'REGION', 'HHSIZE')

# Variables clave continuas
selectedNumVar <- c('TANHHEXP', 'INCTOTGROSSHH')

# PRAM variables
selectedPramVars <- c('ROOF', 'TOILET', 'WATER', 'ELECTCON',
                'FUELCOOK', 'OWNMOTORCYCLE', 'CAR', 'TV', 'LIVESTOCK')

# Peso del hogar
selectedWeightVar <- c('WGTPOP')


# Creando el objeto sdcMicro con las variables asignadas
sdcInitial <- createSdcObj(dat          = file,
                           keyVars      = selectedKeyVars,
                           numVar       = selectedNumVar,
                           weightVar    = selectedWeightVar,
                           pramVars     = selectedPramVars)

# Resumen del objeto
sdcInitial

## ----------------------------------------------------------------------------
```

La Tabla **??** presenta los nombres de los elementos y sus respectivos contenidos. Los nombres de los elementos se pueden listar usando la función `slotNames()`, que se ilustra en el Bloque **??**. Algunos espacios se llenan solo después de aplicar ciertos métodos, por ejemplo, evaluar una medida de riesgo específica. Se puede acceder a ciertos elementos de los objetos mediante funciones de acceso (por ejemplo, `extractManipData` para extraer los datos anónimos) o funciones de impresión (por ejemplo, `print()`) con los argumentos apropiados. También se puede acceder directamente al contenido de un espacio con el operador '@' y el nombre del espacio. Esto se ilustra para el elemento o atributo de riesgo en el Bloque **??**. Esta funcionalidad puede ser práctica para guardar resultados intermedios y comparar los resultados de diferentes métodos. Además, para cambios manuales en los datos durante el proceso SDC, como cambiar códigos

de valores faltantes o recodificación manual, es útil el acceso directo de los datos en los elementos o propiedades con los datos manipulados (es decir, nombres de elemento que comienzan con 'manip'). Dentro de cada elemento generalmente hay varios elementos. Sus nombres se pueden mostrar con la función `names()` y se puede acceder a ellos con el operador '$'. Esto se muestra para el elemento con el riesgo individual en el elemento de riesgo.

Bloque 6.9. Visualización de nombres de elementos o propiedades y acceso a elementos o propiedades de un objeto S4

```
# Lista de todos los slots de objeto sdcMicro
slotNames(sdcInitial)

# Accediendo al slot de riesgos
sdcInitial@risk

# Lista de nombres dentro del slot de riesgo
names(sdcInitial@risk)
## [1] "global"  "individual"  "numeric"

# Dos formas de acceder al riesgo individual dentro del slot de riesgo
sdcInitial@risk$individual
get.sdcMicroObj(sdcInitial, "risk")$individual
```

Hay dos opciones para guardar los resultados después de aplicar los métodos SDC:

1. Sobrescribir el objeto `sdcMicro` existente, o
2. creando un nuevo objeto `sdcMicro`. El objeto original no se modificará y se puede utilizar para comparar resultados. Esto es especialmente útil para comparar varios métodos y seleccionar la mejor opción.

En ambos casos, el resultado de cualquier función debe reasignarse a un objeto con el operador '<-'. Ambos métodos se ilustran en el Bloque **??**.

Bloque 6.10. Guardado de resultados de la aplicación de métodos SDC

```
# Aplicar supresión local y reasignar los resultados al mismo objeto sdcMicro
sdcInitial <- localSuppression(sdcInitial)

# Aplicar supresión local y asignar los resultados a un nuevo objeto sdcMicro
sdc1 <- localSuppression(sdcInitial)
```

Si los resultados se reasignan al mismo objeto sdcMicro, es posible deshacer el último paso del proceso SDC. Esto es útil al cambiar los parámetros. Sin

Tabla 6.2: Nombres de elementos o propiedades y descripción de los elementos o propiedades del objeto 'sdcMicro'

Nombre de elemento	Contenido
'origData'	datos originales como se especifica en el argumento dat de la función 'createSdcObj()'.
'keyVars'	índices de columnas en 'origData' con variables clave categóricas especificadas.
'pramVars'	índices de columnas en 'origData' con variables PRAM especificadas.
'numVars'	índices de columnas en 'origData' con variables clave numéricas especificadas.
'ghostVars'	índices de columnas en 'origData' con 'ghostVars' especificados.
'weightVar'	índices de columnas en 'origData' con variable de peso especificada.
'hhId'	índices de columnas en 'origData' con variable de clúster especificada.
'strataVar'	índices de columnas en 'origData' con variable de estratos especificada.
'sensibleVar'	índices de columnas en 'origData' con variables sensibles especificadas para *l-diversity*.
'manipKeyVars'	variables clave categóricas manipuladas después de aplicar métodos SDC (ver elemento 'key
'manipPramVars'	variables PRAM manipuladas después de aplicar PRAM (ver elemento 'pramVars').
'manipNumVar'	variables clave numéricas manipuladas después de aplicar métodos SDC (ver elemento 'num
'manipGhostVars'	variables fantasma manipuladas (ver elemento 'ghostVars').
'manipStrataVar'	variables de estratos manipulados (ver elemento 'strataVar').
'originalRisk'	medidas de riesgo globales e individuales antes de la anonimización.
'risk'	medidas de riesgo global e individual después de la aplicación de métodos SDC.
'utility'	medidas de utilidad (il1 y eigen).
'pram'	detalles sobre PRAM después de aplicar PRAM.
'localSuppression'	número de supresión por variable después de la supresión local.
'options'	opciones especificadas.
'additionalResults'	resultados adicionales.
'set'	lista de elemento actualmente en uso (para uso interno).
'prev'	información para deshacer un paso con la función 'undo()'.
'deletedVars'	variables eliminadas (identificadores directos).

embargo, los resultados del último paso se pierden después de deshacer ese paso.

La función `undolast()` se puede usar para retroceder solo un paso, no varios. El resultado también debe ser reasignado al mismo objeto. Esto se ilustra en el Bloque **??**.

Bloque 6.11. Deshacer último paso en proceso SDC

```
# Deshacer el último paso en el proceso SDC
sdcInitial <- undolast(sdcInitial)
```

6.7 Estructura del hogar

Si los datos tienen una estructura jerárquica y algunas variables se miden en el nivel jerárquico más alto y otras en el nivel más bajo, el proceso SDC debe adaptarse en consecuencia (véanse también la sección Riesgo jerárquico (o del hogar)). Un ejemplo común en los datos de encuestas sociales son los conjuntos de datos con una estructura de hogar. Las variables que se miden a nivel del hogar son, por ejemplo, los ingresos del hogar, el tipo de vivienda y la región. Las variables medidas a nivel individual son, por ejemplo, la edad, el nivel educativo y el estado civil. Algunas variables se miden a nivel individual, no obstante, son las mismas para todos los miembros del hogar en casi todos los hogares. Estas variables deben ser tratadas como medidas a nivel de hogar desde la perspectiva del SDC. Un ejemplo es la variable religión para algunos países.

El proceso SDC debe dividirse en dos etapas en los casos en que los datos tengan una estructura de hogar. Primero, las variables en el nivel superior (hogar) deben anonimizarse; posteriormente, las variables de nivel superior tratadas deben fusionarse con las variables individuales y anonimizarse conjuntamente. En esta sección, explicamos cómo extraer variables del hogar de un archivo y fusionarlas con las variables de niveles individuales después del tratamiento en R. Ilustramos este proceso con un ejemplo de variables a nivel individual y del hogar.

Estos pasos se ilustran en el Bloque **??**. Requerimos una identificación individual y una identificación familiar en el conjunto de datos; si faltan, deben generarse. La identificación individual debe ser única para cada individuo en el conjunto de datos y la identificación del hogar debe ser única para todos los hogares. El primer paso es extraer las variables del hogar y guardarlas en un nuevo marco de datos. Especificamos las variables que se miden a nivel del hogar en el vector de cadena "HHVars" y restamos solo estas variables del conjunto de datos. Este marco de datos tendrá para cada hogar el mismo número de entradas que miembros del hogar (por ejemplo, si un hogar tiene cuatro miembros, este hogar aparecerá cuatro veces en el archivo). A continuación, aplicamos la función `unique()` para seleccionar solo un registro por hogar. Este argumento de la

función `unique()` es la identificación del hogar, que es la misma para todos los miembros del hogar,

Bloque 6.12. Crear un archivo a nivel de hogar con registros únicos (eliminar duplicados) [7]

```
# Crear subconjunto de archivo con solo variables medidas a nivel de hogar
HHVars <- c('IDH', selectedKeyVars, selectedPramVars, selectedNumVar, selectedWeightVar)
fileHH <- file[,HHVars]

# Elimine las filas duplicadas en función de la identificación del hogar /
# solo cada hogar una vez en el fileHH
fileHH <- unique(fileHH, by = c('HID'))

# Dimensiones del fileHH (número de hogares)
dim(fileHH)
```

Después de anonimizar las variables del hogar con base en el marco de datos "fileHH", recombinamos las variables del hogar anonimizadas con las variables originales, que se miden a nivel individual. Podemos extraer las variables de nivel individual del conjunto de datos original usando "INDVars", un vector de cadena con los nombres de las variables de nivel individual. Para extraer los datos anonimizados del objeto sdcMicro, podemos usar la función `extractManipData()` del paquete `sdcMicro`. A continuación, fusionamos los datos usando la función `merge()`. El argumento 'by' en la función `merge()` especifica la variable utilizada para la combinación; en este caso, la identificación del hogar, que tiene el mismo nombre de variable en ambos conjuntos de datos. Todas las demás variables deben tener nombres diferentes en ambos conjuntos de datos. Estos pasos se ilustran en Bloque **??**.

Bloque 6.13. Fusión de variables anonimizadas a nivel de hogar con variables a nivel individual

```
# Crea objeto sdcMicro inicial para variables de nivel de hogar
sdcHH <- createSdcObj(dat = fileHH, keyVars = selectedKeyVars,
                      pramVars = selectedPramVars, weightVar = selectedWeightVar,
                      numVars = selectedNumVar)
numHH <- length(fileHH[,1]) # número de hogares

# Extrae variables de nivel de hogar manipuladas del objeto SDC
HHmanip <- extractManipData(sdcHH)

# Selecciona variables (nivel individual)
```

[7] Se recomienda verificar que el objeto fileHH tenga después de la aplicación de la función `unique()` la cantidad de filas esperadas (ej.: N° de viviendas encuestadas) y que no haya valores perdidos no esperados.

```
selectedKeyVarsIND = c('GENDER', 'REL', 'MARITAL','AGEYRS',
                       'EDUCY', 'INDUSTRY1') # lista de variables clave seleccionadas

# Peso de la muestra (WGTHH, pesos individuales)
selectedWeightVarIND = c('WGTHH')

# ID hogar
selectedHouseholdID = c('IDH')

# Todas las variables individuales
INDVars <- c(selectedKeyVarsIND)

# Recombinando los datos HH anonimizados y las variables a nivel individual
indVars <- c("IDH", "IDP", selectedKeyVarsIND, "WGTHH") # IDH y todas las variantes no
fileInd <- file[indVars] # subset de file sin HHVars
fileCombined <- merge(HHmanip, fileInd, by.x = c('IDH'))
fileCombined <- fileCombined[order(fileCombined[,'IDH'],  fileCombined[,'IDP']),]

dim(fileCombined)

# Objeto SDC con solo las variables a nivel IND
sdcCombined <- createSdcObj(dat = fileCombined, keyVars = c(selectedKeyVarsIND),
                            weightVar = selectedWeightVarIND, hhId = selectedHouseholdI

# Objeto SDC con ambos niveles de variables, a HH y IND
sdcCombinedAll <- createSdcObj(dat = fileCombined,
                               keyVars = c(selectedKeyVarsIND, selectedKeyVars ),
                               weightVar = selectedWeightVarIND,
                               hhId = selectedHouseholdID)
sdcCombinedAll
```

El archivo fileCombined se utiliza para el proceso SDC con todo el conjunto de datos. En el estudio de casos de la sección Caso de estudio se ilustra cómo tratar los datos con la estructura del hogar.

El tamaño de un hogar también puede ser un identificador indirecto, incluso si el tamaño del hogar no está incluido en el conjunto de datos como variable. Con el fin de evaluar el riesgo de divulgación, podría ser necesario crear dicha variable mediante un recuento de los miembros de cada hogar. El Bloque **??** muestra cómo generar la variable de tamaño de hogar, con valores para cada individuo en función de la identificación del hogar (IDH). Se muestran dos casos: 1) el archivo ordenado por IDH y 2) el archivo no ordenado.

Bloque 6.14. Generando la variable tamaño del hogar

```
# Ordenado por IDH
file$hhsize <- rep(unname(table(file$IDH)), unname(table(file$IDH)))

# Desordenado
file$hhsize <- rep(diff(c(1, 1 + which(diff(file$IDH) != 0), length(file$IDH) + 1)),
                   diff(c(1, 1 + which(diff(file$IDH) != 0), length(file$IDH) + 1)))
```

En algunos casos, el orden de las personas dentro de los hogares puede proporcionar información que podría conducir a la reidentificación.

Un ejemplo es la información sobre la relación con el jefe de hogar. En muchos países, el primer miembro del hogar es el cabeza de familia, el segundo es la pareja del cabeza de familia y los siguientes son los hijos. Por lo tanto, el número de línea dentro del hogar podría correlacionarse bien con una variable que contiene información sobre la relación con el jefe de hogar. Una forma de evitar esta divulgación involuntaria de información es cambiar el orden de los individuos dentro de cada hogar al azar. El Bloque **??** ilustra una manera de hacer esto en R.

Bloque 6.15. Cambiando el orden de los individuos dentro de los hogares

```
# Cargando datos anonimizados
dataAnon<-readRDS("dataAnon.RDS")

# Lista de tamaños de hogar por hogar
hhsize <- diff(c(1, 1 + which(diff(dataAnon$IDH) != 0), length(dataAnon$IDH) + 1))

# Números de línea asignados al azar dentro de cada hogar
set.seed(123)
dataAnon$INDID <- unlist(lapply(hhsize,
                                function(n){sample(1:n, n, replace = FALSE,
                                                   prob = rep(1/n, n))}))

# Ordene el archivo por IDH y INDID aleatorio (número de línea)
dataAnon <- dataAnon[order(dataAnon$IDH, dataAnon$INDID),]
```

6.8 Tiempo de cómputo

Algunos métodos SDC pueden tardar mucho tiempo en evaluarse en términos de cómputo. Por ejemplo, la supresión local con la función `localSuppression()` del paquete `sdcMicro` en `R` puede tardar días en ejecutarse en grandes conjuntos de datos de más de 30.000 personas que tienen muchos identificadores indirectos categóricos. El uso de la función `groupVars()`, por ejemplo, no es computacionalmente intensivo, pero aún puede llevar mucho tiempo si el conjunto de datos es grande y las medidas de riesgo deben volver a calcularse.

Nuestra experiencia revela que el tiempo de cómputo es una función de los siguientes factores: el método SDC aplicado; tamaño de los datos, es decir, número de observaciones, número de variables y número de categorías o niveles de factores de cada variable categórica; complejidad de los datos (por ejemplo, el número de diferentes combinaciones de valores de identificadores indirectos en los datos); así como las especificaciones de la computadora (procesador, la memoria RAM y los medios de almacenamiento).

El uso de la paralelización puede mejorar el rendimiento incluso en una sola computadora con un procesador con múltiples núcleos. `R` no utiliza múltiples núcleos a menos que se le indique que lo haga. La paralelización permite que los trabajos/escenarios [8] en los conjuntos de datos puedan procesarse simultáneamente a través de la asignación eficiente de tareas a diferentes procesadores. Sin paralelización, dependiendo del servidor/computadora, solo se usa un núcleo cuando se ejecutan los trabajos secuencialmente. Ejecutar el programa de anonimización sin paralelización conduce a un tiempo de ejecución significativamente mayor. Sin embargo, tenga en cuenta que la paralelización en sí misma también provoca una sobrecarga. Por lo tanto, una suma de los tiempos que lleva ejecutar cada tarea en paralelo no equivale necesariamente al tiempo que puede llevar ejecutarlas secuencialmente. Sin embargo, el hecho de que la RAM se comparta podría reducir ligeramente las ganancias de la paralelización [9].

6.9 Errores comunes

En esta sección, presentamos algunos errores comunes y sus causas, que pueden encontrarse al usar el paquete `sdcMicro` en `R` para la anonimización de microdatos:

1. La clase de una determinada variable no es aceptada por la función, por ejemplo, una variable categórica de clase numérica debe recodificarse primero a la clase requerida (por ejemplo, factor o data.frame). En la sección Clases en `R` se muestra cómo hacerlo.

2. Después de realizar cambios manualmente en las variables, el riesgo no cambió, ya que no se actualiza automáticamente y debe volver a calcularse manualmente mediante la función `calcRisks()`.

[8] Aquí, un escenario se refiere a una combinación de métodos SDC y sus parámetros.

[9] El siguiente sitio web proporciona una descripción general de los paquetes y soluciones de paralelización en `R` : http://cran.r-project.org/web/views/HighPerformanceComputing.html.

Capítulo 7

Medición de riesgos

7.1 Tipos de divulgación

Medir el riesgo de divulgación es una parte importante del proceso SDC: las medidas de riesgo se utilizan para juzgar si un archivo de datos es lo suficientemente seguro para su liberación. Antes de medir el riesgo de divulgación, debemos definir qué tipo de divulgación es relevante para los datos disponibles, a saber: divulgación de identidad, divulgación de atributos y divulgación inferencial (ver **?** y **?**).

- **Divulgación de identidad**, que ocurre si el intruso asocia a un individuo conocido con un registro de datos publicado. Por ejemplo, el intruso vincula un registro de datos publicado con información externa o identifica a un informante con valores de datos extremos. En este caso, un intruso puede explotar un pequeño subconjunto de variables para realizar la vinculación, y una vez que la vinculación es exitosa, el intruso tiene acceso a toda la demás información en los datos publicados relacionados con el informante específico.

- **Divulgación de atributos**, que ocurre si el intruso puede determinar algunas características nuevas de un individuo en función de la información disponible en los datos publicados. La divulgación de atributos ocurre si se vuelve a identificar correctamente a un informante y el conjunto de datos incluye variables que contienen información que el intruso desconocía previamente. La divulgación de atributos también puede ocurrir sin divulgación de identidad. Por ejemplo, si un hospital publica datos que muestran que todas las pacientes de 56 a 60 años que tienen cáncer, un intruso conoce la condición médica de cualquier paciente de 56 a 60 años en el conjunto de datos sin tener que identificar a la persona específica.

- **Divulgación inferencial**, que ocurre si el intruso es capaz de determinar el valor de alguna característica de un individuo con mayor precisión con los datos liberados de lo que hubiera sido posible de otro modo. Por ejemplo, con un modelo de regresión altamente predictivo, un intruso puede inferir la información confidencial de ingresos de un informante utilizando atributos registrados en los datos, lo que lleva a una divulgación inferencial.

Los métodos SDC para microdatos están destinados a evitar la divulgación de identidades y atributos. La divulgación inferencial generalmente no se aborda en SDC en el entorno de microdatos, ya que los microdatos se liberan precisamente para que los investigadores puedan hacer inferencias estadísticas y comprender las relaciones entre las variables. En ese sentido, la inferencia no puede compararse con la divulgación. Además, las inferencias están diseñadas para predecir el comportamiento agregado, no individual y, por lo tanto, suelen ser malos predictores de valores de datos individuales.

7.2 Clasificación de variables

A los efectos del proceso SDC, utilizamos las clasificaciones de variables descritas en los siguientes párrafos (consulte la Figura **??** para obtener una descripción general). La clasificación inicial de variables en variables de identificación y no identificación depende de la forma en que los intrusos pueden utilizar las variables para la reidentificación:

- **Variables de identificación**: contienen información que puede conducir a la identificación de los informantes y se pueden clasificar en:
  - **Los identificadores directos**, que revelan de manera directa e inequívoca la identidad del informante. Algunos ejemplos son nombres, números de pasaporte, números de identificación social y direcciones. Los identificadores directos deben eliminarse del conjunto de datos antes de su publicación. La eliminación de identificadores directos es un proceso sencillo y siempre es el primer paso para producir un conjunto de microdatos seguro para su publicación. Sin embargo, la eliminación de identificadores directos a menudo no es suficiente.
  - **Los identificadores indirectos (cuasi-identificadores o variables clave)** contienen información que, cuando se combina con otros identificadores indirectos en el conjunto de datos, puede conducir a la reidentificación de los informantes. Este es especialmente el caso cuando se pueden usar para hacer coincidir la información con otra información o datos externos. Ejemplos de identificadores indirectos son la raza, la fecha de nacimiento, el sexo y el código postal,

que pueden combinarse o vincularse fácilmente con información externa disponible públicamente y hacer posible la identificación. Las combinaciones de valores de varios identificadores indirectos se denominan claves. Los valores de los identificadores indirectos por sí mismos a menudo no conducen a la identificación (por ejemplo, hombre/mujer), pero una combinación de varios valores de identificador indirecto puede hacer que un registro sea único (por ejemplo, hombre, 18 años, casado) y, por lo tanto, identificable. En general, no es aconsejable eliminar simplemente los identificadores indirectos de los datos para resolver el problema. En muchos casos, serán variables importantes para cualquier análisis sensato. En la práctica, cualquier variable en el conjunto de datos podría potencialmente usarse como un identificador indirecto. SDC aborda esto mediante la identificación de variables como identificadores indirectos y anonimizándolas mientras mantiene la información en el conjunto de datos para su publicación.

- **Las variables de no identificación** son variables que no se pueden utilizar para volver a identificar a los informantes. Esto podría deberse a que estas variables no están contenidas en ningún otro archivo de datos u otras fuentes externas y no son observables por un intruso. No obstante, las variables de no identificación son importantes en el proceso SDC, ya que pueden contener información confidencial/sensible, que puede resultar perjudicial si se produce una divulgación como resultado de la divulgación de la identidad basada en variables de identificación.

Estas clasificaciones de variables dependen parcialmente de la disponibilidad de conjuntos de datos externos que pueden contener información que, cuando se combina con los datos actuales, podría conducir a la divulgación. La identificación y clasificación de variables como identificadores indirectos depende, entre otros, de la disponibilidad de información en conjuntos de datos externos. Un paso importante en el proceso SDC es definir una lista de posibles escenarios de divulgación en función de cómo los identificadores indirectos podrían combinarse entre sí y la información en conjuntos de datos externos, y luego tratar los datos para evitar la divulgación. Analizamos los escenarios de divulgación con más detalle en la sección Escenarios de divulgación.

Para el proceso SDC, también es útil clasificar aún más los identificadores indirectos en variables categóricas, continuas y semicontinuas o discretas. Esta clasificación es importante para determinar los métodos SDC apropiados para esa variable, así como la validez de las medidas de riesgo.

- **Las variables categóricas** toman valores sobre un conjunto finito, y cualquier operación aritmética que las utilice generalmente no tiene sentido o no está permitida. Ejemplos de variables categóricas son género, región y nivel educativo.

- **Las variables continuas** pueden tomar un número infinito de valores en un conjunto denso. Algunos ejemplos son los ingresos, la altura del cuerpo y el tamaño del terreno. Las variables continuas se pueden transformar en variables categóricas mediante la construcción de intervalos (como bandas de ingresos)[1].

- **Las variables semicontinuas o discretas** son variables continuas que toman valores limitados a un conjunto finito. Un ejemplo es la edad medida en años, que podría tomar valores en el conjunto {0, 1,..., 100}. La naturaleza finita de los valores de estas variables significa que pueden tratarse como variables categóricas a los efectos de SDC [2].

Además de estas clasificaciones de variables, el proceso SDC clasifica aún más las variables según su sensibilidad o confidencialidad. Tanto las variables identificadoras indirectas como las de no identificación pueden clasificarse como sensibles (o confidenciales) o no sensibles (o no confidenciales). Esta distinción no es importante para los identificadores directos, ya que los identificadores directos se eliminan de los datos publicados.

- **Las variables sensibles** contienen información confidencial que no debe liberarse sin un tratamiento adecuado, utilizando los métodos de SDC para reducir el riesgo de divulgación. Algunos ejemplos son los ingresos, la religión, la afiliación política y las variables relativas a la salud. Que una variable sea sensible depende del contexto y del país: una determinada variable puede considerarse sensible en un país y no sensible en otro.

- **Las variables no sensibles** contienen información no confidencial sobre el informante, como el lugar de residencia o el área de residencia rural/urbana. Sin embargo, la clasificación de una variable como no sensible no significa que no deba ser considerada en el proceso de SDC. Las variables no sensibles aún pueden servir como identificadores indirectos cuando se combinan con otras variables u otros datos externos.

[1] Recodificar una variable continua a veces es útil en los casos en que los datos contienen solo unas pocas variables continuas. Veremos en la sección Riesgo individual que muchos métodos utilizados para el cálculo del riesgo dependen de si las variables son categóricas. También veremos que es más fácil para la medición del riesgo si los datos contienen solo variables categóricas o solo continuas.

[2] Esto se discute con mayor detalle en las siguientes secciones. En los casos en que el número de valores posibles sea grande, se recomienda recodificar la variable, o partes del conjunto en el que toma valores, para obtener menos valores distintos.

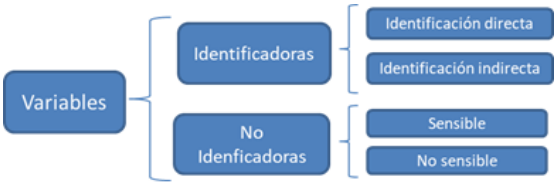

Figura 7.1: Clasificación de las variables.

7.3 Escenarios de divulgación

La evaluación del riesgo de divulgación se lleva a cabo con referencia a las fuentes de datos disponibles en el entorno donde se liberará el conjunto de datos. En este contexto, el riesgo de divulgación es la posibilidad de volver a identificar correctamente a una unidad en el archivo de microdatos publicado [3] comparando sus datos con un archivo externo basado en un conjunto de identificadores indirectos. La evaluación de riesgos se realiza mediante la identificación de los llamados escenarios de divulgación o intrusión. Un escenario de divulgación describe la información potencialmente disponible para el intruso (por ejemplo, datos del censo, padrones electorales, registros de población o datos recopilados por empresas privadas) para identificar a los informantes y las formas en que dicha información puede combinarse con el conjunto de microdatos que se liberará y utilizará para reidentificación de registros en el conjunto de datos. Normalmente, estos conjuntos de datos externos incluyen identificadores directos. En ese caso, la reidentificación de los registros en el conjunto de datos publicado conduce a la divulgación de la identidad y, posiblemente, de los atributos. El principal resultado de la evaluación de los escenarios de divulgación es la identificación de un conjunto de identificadores indirectos (es decir, variables clave) que deben tratarse durante el proceso SDC (ver **?**).

Un ejemplo de un escenario de divulgación podría ser el reconocimiento espontáneo de un informante por parte de un investigador. Por ejemplo, mientras revisa los datos, el investigador reconoce a una persona con una combinación inusual de las variables edad y estado civil. Por supuesto, esto solo puede suceder si la persona es bien conocida o es conocida por el investigador. Otro ejemplo de un escenario de divulgación para un archivo disponible públicamente sería si las variables en los datos pudieran vincularse a un registro electoral disponible públicamente. Un intruso podría hacer coincidir todo el conjunto de datos con las personas del registro. Sin embargo, esto puede ser difícil y requerir experiencia especializada, o *software*, y se deben cumplir otras condiciones. Los ejemplos son que el momento en el que se recopilaron los conjuntos de datos

[3] No todos los datos externos son necesariamente de dominio público. También se deben tener en cuenta los conjuntos de datos de propiedad privada o los conjuntos de datos que no se divulgan para determinar el escenario de divulgación adecuado.

debe coincidir aproximadamente y el contenido de las variables debe ser (casi) idéntico. Si no se cumplen estas condiciones, la coincidencia exacta es mucho menos probable.

La evaluación del riesgo de divulgación se basa en los identificadores indirectos, que se identifican en el análisis de escenarios de riesgo de divulgación. El riesgo de divulgación depende directamente de la inclusión o exclusión de variables en el conjunto de identificadores indirectos elegidos. Por lo tanto, este paso en el proceso SDC (hacer la elección de los identificadores indirectos) debe abordarse con gran reflexión y cuidado. Veremos más adelante, cuando discutamos los pasos en el proceso de SDC con más detalle, que el primer paso para cualquier oficina de estadística es realizar un ejercicio en el que se compila un inventario de todos los conjuntos de datos disponibles en el país. Se consideran tanto los conjuntos de datos publicados por la oficina nacional de estadística (como el INE) como por otras fuentes y se analiza su disponibilidad para los intrusos, así como las variables incluidas en estos conjuntos de datos.

7.4 Niveles de riesgo

Con microdatos de encuestas y censos, a menudo tenemos que preocuparnos por la divulgación a nivel individual o de unidad, es decir, identificar a los informantes individuales. Los informantes individuales suelen ser personas físicas, pero también pueden ser unidades, como empresas, escuelas, centros de salud, etc. Los archivos de microdatos suelen tener una estructura jerárquica en la que las unidades individuales pertenecen a grupos, por ejemplo, las personas pertenecen a hogares. La estructura jerárquica más común en los microdatos es la estructura del hogar en los datos de las encuestas de hogares. Por lo tanto, en esta guía, a veces llamamos al riesgo de divulgación de datos con una estructura jerárquica "riesgo hogar". Sin embargo, los conceptos se aplican por igual a los datos del establecimiento y otros datos con estructuras jerárquicas, como los datos de la escuela con los alumnos y profesores o los datos de la empresa con los empleados.

Veremos que es importante tener en cuenta esta estructura jerárquica al medir el riesgo de divulgación. Para los datos jerárquicos, la información recopilada en el nivel jerárquico superior (por ejemplo, nivel del hogar) sería la misma para todos los individuos del grupo que pertenecen a ese nivel jerárquico superior (por ejemplo, el hogar) [Los supuestos para esta medida de riesgo son estrictos y el riesgo se estima en muchos casos mayor que el riesgo real. Entre otras suposiciones, se supone que todos los individuos de la muestra también están incluidos en el archivo externo utilizado por el intruso para compararlos. Si no es así, el riesgo es mucho menor; si el individuo en el archivo liberado no está incluido en el archivo externo, la probabilidad de una coincidencia correcta es cero. Otras suposiciones son que los archivos no contienen errores y que ambos conjuntos de datos se recopilaron simultáneamente, es decir, contienen

la misma información. Estos supuestos a menudo no se cumplen en general, pero son necesarios para el cálculo de una medida. Un ejemplo de una violación de las últimas suposiciones podría ocurrir si los conjuntos de datos se recopilan en diferentes puntos en el tiempo y los registros han cambiado. Esto podría suceder cuando las personas se mudan o cambian de trabajo y hace imposible la coincidencia correcta. **Los supuestos son conservadores y asumen el mayor riesgo de divulgación**.]. Algunos ejemplos típicos de variables que tendrían los mismos valores para todos los miembros de una misma unidad jerárquica superior son, en el caso de los hogares, las relativas a la vivienda y los ingresos del hogar. Estas variables difieren de una encuesta a otra y de un país a otro [4]. Esta estructura jerárquica crea un mayor nivel de riesgo de divulgación por dos razones:

1. si se reidentifica a una persona del hogar, la estructura del hogar permite la reidentificación de los demás miembros del hogar en el mismo hogar,
2. los valores de las variables para otros miembros del hogar que son comunes para todos los miembros del hogar pueden usarse para volver a identificar a otro individuo del mismo hogar. Esto se analiza con más detalle en la sección Riesgo jerárquico (o del hogar).

A continuación, primero analizamos las medidas de riesgo utilizadas para evaluar el riesgo de divulgación en ausencia de una estructura jerárquica. Esto incluye medidas de riesgo que buscan agregar el riesgo individual para todos los individuos en el archivo de microdatos; el objetivo es cuantificar una medida de riesgo de divulgación global para el archivo. Luego discutimos cómo cambian las medidas de riesgo cuando se tiene en cuenta la estructura jerárquica de los datos.

También discutiremos cómo las medidas de riesgo difieren para los identificadores indirectos categóricos y continuos. Para las variables categóricas, utilizaremos el concepto de unicidad de combinaciones de valores de identificadores indirectos (las llamadas "claves") que se utilizan para identificar a las personas en riesgo. El concepto de unicidad, sin embargo, no es útil para variables continuas, ya que es probable que todos o muchos individuos tengan valores únicos para esa variable, por definición de una variable continua. Las medidas de riesgo para variables categóricas son generalmente medidas a priori, es decir, pueden evaluarse antes de aplicar métodos de anonimización ya que se basan en el principio de unicidad. Las medidas de riesgo para variables continuas son medidas a posteriori; se basan en la comparación de los microdatos antes y después de la anonimización y son, por ejemplo, basadas en la proximidad de observaciones entre conjuntos de datos originales y tratados (anonimizados).

Los archivos que se limitan solo a identificadores indirectos categóricos o continuos son los más fáciles para medir el riesgo. Veremos en secciones posteriores

[4] Consulte la sección Objetos de la clase `sdcMicroObj` para obtener más información sobre los *slots* y la estructura del objeto `sdcMicro`.

que, en los casos en que ambos tipos de variables están presentes, la recodificación de variables continuas en categorías es un enfoque para simplificar el proceso SDC, pero también veremos que desde una perspectiva de utilidad esto puede no ser deseable. Un ejemplo podría ser el uso de quintiles de ingresos en lugar de las variables de ingresos reales. Veremos que medir el riesgo de divulgación con base en las variables categóricas y continuas por separado generalmente no es un enfoque válido.

Las medidas de riesgo discutidas en la siguiente sección se basan en varios supuestos. En general, estas medidas se basan en suposiciones bastante restrictivas y, a menudo, conducirán a estimaciones de riesgo conservadoras. Estas medidas de riesgo conservadoras pueden exagerar el riesgo ya que suponen el peor de los casos. Sin embargo, se deben cumplir dos suposiciones para que las medidas de riesgo sean válidas y significativas; los microdatos deben ser una muestra de una población más grande (no censo) y las ponderaciones de la muestra deben estar disponibles.

7.5 Riesgo individual

7.5.1 Identificadores indirectos categóricos y recuentos de frecuencia

El enfoque principal de la medición del riesgo para los identificadores indirectos categóricos es la divulgación de la identidad. La medición del riesgo de divulgación se basa en la evaluación de la probabilidad de reidentificación correcta de las personas en los datos publicados. Utilizamos medidas basadas en los microdatos reales que se publicarán. En general, cuanto más rara sea la combinación de valores de los identificadores indirectos (es decir, clave) de una observación en la muestra, mayor será el riesgo de revelación de identidad. Un intruso que intente hacer coincidir a una persona que tiene una clave relativamente rara dentro de los datos de muestra con un conjunto de datos externo en el que existe la misma clave tendrá una mayor probabilidad de encontrar una coincidencia correcta que cuando un número mayor de personas comparten la misma clave. Esto se puede ilustrar con el siguiente ejemplo que se ilustra en la Tabla **??**.

La Tabla **??** muestra los valores de 10 informantes para los identificadores indirectos “área”, “género”, “nivel educacional” y “situación laboral”. En los datos, encontramos siete combinaciones únicas de valores de identificadores indirectos (es decir, patrones o claves) de los cuatro identificadores indirectos. Ejemplos de claves son {‘urbano’, ‘femenino’, ‘secundaria incompleta’, ‘ocupado’} y {‘urbano’, ‘femenino’, ‘primaria incompleta’, ‘no FL’}. Sea f_k la frecuencia de muestreo de la *k-ésima* clave, es decir, el número de individuos de la muestra con valores de los identificadores indirectos que coinciden con la clave k. Este sería 2 para la clave {urbano, femenino, secundaria incompleta, ocupado}, ya que esta

Tabla 7.1: Conjunto de datos de ejemplo que muestra frecuencias de muestra, frecuencias de población y riesgo de divulgación individual

Id	Área	Género	Nivel educacional	Situación laboral	Peso (w_{i})	f_{k}	F_{k}
1	Urbano	Femenino	Secundaria incompleta	Ocupado	180	2	360
2	Urbano	Femenino	Secundaria incompleta	Ocupado	180	2	360
3	Urbano	Femenino	Primaria incompleta	No FL	215	1	215
4	Urbano	Masculino	Secundaria completa	Ocupado	76	2	152
5	Rural	Femenino	Secundaria completa	Desocupado	186	1	186
6	Urbano	Masculino	Secundaria completa	Ocupado	76	2	152
7	Urbano	Femenino	Primaria completa	No FL	180	1	180
8	Urbano	Masculino	Post secundaria	Desocupado	215	1	215
9	Urbano	Femenino	Secundaria incompleta	No FL	186	2	262
10	Urbano	Femenino	Secundaria incompleta	No FL	76	2	262

clave es compartida por los individuos 1 y 2 y 1 para la clave {'urbano', 'femenino', 'primaria incompleta', 'no FL'}, que es exclusivo del individuo 3. Por definición, f_k es el mismo para cada registro que comparte una clave particular.

Fuente: Adaptación de (**?**, p.28)

Cuanto menos personas con las que una persona comparte su combinación de identificadores indirectos, más probable es que la persona coincida correctamente en otro conjunto de datos que contenga estos identificadores indirectos. Incluso cuando los identificadores directos se eliminan del conjunto de datos, esa persona tiene un mayor riesgo de divulgación que otras, suponiendo que sus pesos de muestra sean los mismos. La Tabla **??** reporta las frecuencias de muestreo f_k de las llaves para todos los individuos. Las personas con las mismas claves tienen la misma frecuencia de muestreo. Si $f_k = 1$, este individuo tiene una combinación única de valores de identificadores indirectos y se denomina "muestra única". El conjunto de datos de la Tabla **??** contiene cuatro muestras únicas. Las medidas de riesgo se basan en esta frecuencia de muestreo.

En el Bloque **??**, mostramos cómo usar el paquete `sdcMicro` para crear una lista de frecuencias de muestra f_k para cada registro en un conjunto de datos. Esto se hace usando la función `sdcMicro freq()`. Un valor de 2 para una observación significa que en la muestra hay un individuo más con exactamente la misma combinación de valores para los identificadores indirectos seleccionados. En el Bloque **??**, la función `freq()` se aplica a "sdcInitial", que es un objeto `sdcMicro`. Los objetos se usan cuando se hace SDC con `sdcMicro`. La función `freq()` muestra la frecuencia de muestreo de las claves construidas sobre un conjunto definido de identificadores indirectos. El Bloque **??** corresponde a los datos de la Tabla **??**.

Bloque 7.1. Cálculo f_k usando `sdcMicro`

```
setwd("..\Capacitación\GitHub") # directorio de trabajo

library(sdcMicro) # carga paquete sdcMicro
# Set up conjunto de datos
data <- as.data.frame(cbind(as.factor(c('Urbano', 'Urbano', 'Urbano', 'Urbano',
                                        'Rural', 'Urbano', 'Urbano', 'Urbano',
                                        'Urbano', 'Urbano')),
                            as.factor(c('Femenino', 'Femenino', 'Femenino',
                                        'Masculino','Femenino', 'Masculino',
                                        'Femenino', 'Masculino', 'Femenino',
                                        'Femenino')),
                            as.factor(c('Sec in', 'Sec in', 'Prim in', 'Sec com',
                                        'Sec com', 'Sec com', 'Prim com', 'Post-sec',
                                        'Sec in', 'Sec in')),
                            as.factor(c('Ocu', 'Ocu', 'No-FL', 'Ocu', 'Desocu', 'Ocu',
                                        'No-FL', 'Desocu', 'No-FL','No-FL')),
                            as.factor(c('Sí', 'Sí', 'Sí', 'Sí', 'Sí', 'No', 'No',
                                        'Sí', 'No', 'Sí')),
                            c(180, 180, 215, 76, 186, 76, 180, 215, 186, 76)
))

# Especificar nombres de variables
names(data) <- c('Área', 'Género', 'Educ', 'SitLab', 'Salud', 'Pesos')

# Set up objeto sdcMicro con especificación de identificadores indirectos y pesos
sdcInitial <- createSdcObj(dat = data, keyVars = c('Área', 'Género', 'Educ', 'SitLab')
                           weightVar = 'Pesos')
data$fk<-freq(sdcInitial, type = 'fk')
```

Para datos de muestra, es más interesante mirar F_k, la frecuencia de población de una combinación de identificadores indirectos (clave) k, que es el número de individuos de la población con la clave que corresponde a la clave k. Se desconoce la frecuencia poblacional si los microdatos son una muestra y no un censo. Bajo ciertas suposiciones, el valor esperado de las frecuencias de la población se puede calcular utilizando el peso del diseño de la muestra w_i(en una muestra simple, esta es la inversa de la probabilidad de inclusión) para cada individuo i.

$$F_k = \sum_{i|individuo\, i\, correspondiente\, a\, la\, clave\, k} w_i$$

F_k es la suma de los pesos muestrales de todos los registros con la misma clave k. Por lo tanto, como f_k, F_k es el mismo para cada registro con clave k. El riesgo de una reidentificación correcta es la probabilidad de que la clave coincida con el individuo correcto de la población. Dado que cada individuo en la muestra con

clave k corresponde a F_k individuos en la población, la probabilidad de reidentificación correcta es $1/F_k$. Esta es la probabilidad de reidentificación en el peor de los casos y puede interpretarse como riesgo de divulgación. Los individuos con la misma clave tienen las mismas frecuencias, es decir, la frecuencia de la clave.

Si $F_k = 1$, la clave k es tanto una muestra como una población única y el riesgo de divulgación sería 1. Las características únicas de la población son un factor importante a considerar al evaluar el riesgo y merecen especial atención.

Además, f_k, la frecuencia de muestreo de la clave k (es decir, el número de individuos en la muestra con la combinación de identificadores indirectos correspondientes a la combinación especificada en la clave k) y F_k, la frecuencia de población estimada de k, se puede visualizar en `sdcMicro`. El Bloque **??** ilustra cómo devolver listas de longitud n de frecuencias para todos los individuos. Las frecuencias se muestran para cada individuo y no para cada clave.

Bloque 7.2. Cálculo de frecuencias muestrales y poblacionales usando `sdcMicro`

```
# frecuencia muestral de individuos
data$fk<-freq(sdcInitial, type = 'fk')
# frecuencia poblacional de individuos
data$FK<-freq(sdcInitial, type = 'Fk')
```

En la práctica, este enfoque conduce a estimaciones de riesgo conservadoras, ya que no tiene en cuenta adecuadamente los métodos de muestreo. En este caso, las estimaciones del riesgo de reidentificación pueden ser demasiado altas. Si se utiliza este riesgo sobreestimado, los datos pueden estar sobreprotegidos (es decir, la pérdida de información será mayor que la necesaria) al aplicar las medidas de SDC.

La medida del riesgo r_k es como f_k y F_k, el mismo para todos los individuos que comparten el mismo patrón de valores de identificadores indirectos y se denomina riesgo individual. Los valores r_k también puede interpretarse como la probabilidad de divulgación de 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 los identificadores indirectos. Esta medida de riesgo se basa en ciertos supuestos [5], que son estrictos y pueden conducir a una medida de riesgo relativamente conservadora. En `sdcMicro`, la medida de riesgo r_k se

[5]Los supuestos para esta medida de riesgo son estrictos y el riesgo se estima en muchos casos mayor que el riesgo real. Entre otras suposiciones, se supone que todos los individuos de la muestra también están incluidos en el archivo externo utilizado por el intruso para compararlos. Si no es así, el riesgo es mucho menor; si el individuo en el archivo liberado no está incluido en el archivo externo, la probabilidad de una coincidencia correcta es cero. Otras suposiciones son que los archivos no contienen errores y que ambos conjuntos de datos se recopilaron simultáneamente, es decir, contienen la misma información. Estos supuestos a menudo no se cumplen en general, pero son necesarios para el cálculo de una medida. Un ejemplo de una violación de las últimas suposiciones podría ocurrir si los conjuntos de datos se

calcula automáticamente al crear un objeto `sdcMicro` y se guarda en el *slot* de "riesgo" [6]. El Bloque **??** muestra cómo recuperar las medidas de riesgo usando `sdcMicro` para nuestro ejemplo. Las medidas de riesgo también se presentan en la Tabla **??**.

Bloque 7.3. Slot de riesgo individual en el objeto `sdcMicro`

```
sdcInitial@risk$individual
```

Los principales factores que influyen en el riesgo individual son las frecuencias de muestreo f_k y los pesos de diseño de muestreo w_i. Si un individuo tiene un riesgo relativamente alto de divulgación, en nuestro ejemplo serían los individuos 3, 5, 7 y 8 en la Tabla **??** y el Bloque **??**, la probabilidad de que un posible intruso relacione correctamente a estos individuos con un archivo de datos externo es relativamente alta. En nuestro ejemplo, la razón del alto riesgo es el hecho de que estos individuos son muestras únicas (es decir, $f_k = 1$). Este riesgo es el riesgo del peor de los casos y no implica que la persona sea reidentificada con certeza con esta probabilidad. Por ejemplo, si un individuo incluido en los microdatos no está incluido en el archivo de datos externo, la probabilidad de una coincidencia correcta es cero. No obstante, la medida del riesgo calculada a partir de las frecuencias será positiva como medida de evaluación.

7.5.2 k-anonimato

La medida del riesgo k- anonimato se basa en el principio de que, en un conjunto de datos seguro, el número de personas que comparten la misma combinación de valores (claves) de identificadores indirectos categóricos debe ser superior a un umbral especificado k. El k-anonimato es una medida de riesgo basada en los microdatos a publicar, ya que solo tiene en cuenta la muestra. Un individuo viola el k-anonimato si la frecuencia de muestreo f_k para la llave k es menor que el umbral especificado k. Por ejemplo, si un individuo tiene la misma combinación de identificadores indirectos que otros dos individuos en la muestra, estos individuos satisfacen el 3-anonimato pero violan el 4-anonimato. En el conjunto de datos de la Tabla **??**, seis personas satisfacen el 2-anonimato y cuatro violan el 2-anonimato. Los individuos que violan el 2-anonimato son muestras únicas. La medida de riesgo es el número de observaciones que violan el k-anonimato para un cierto valor de k, que es

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

recopilan en diferentes puntos en el tiempo y los registros han cambiado. Esto podría suceder cuando las personas se mudan o cambian de trabajo y hace imposible la coincidencia correcta. **Los supuestos son conservadores y asumen el mayor riesgo de divulgación**.

[6] Consulte la sección Objetos de la clase `sdcMicroObj` para obtener más información sobre los *slots* y la estructura del objeto `sdcMicro`.

donde I es la función indicadora e i se refiere al *i-ésimo* registro. Esto es simplemente un recuento del número de personas con una frecuencia de muestreo de su clave inferior a k. El recuento es mayor para los k más grandes, ya que si un registro satisface k-anonimato, también satisface (k+1)- anonimato. La medida del 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. Si los pesos de la muestra 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 más bajo. Los pesos de muestra grandes van de la mano con conjuntos de datos más pequeños. 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.

En `sdcMicro` podemos mostrar el número de observaciones que violan un determinado umbral de k-anonimato. En el Bloque **??**, usamos `sdcMicro` para calcular la cantidad de infractores para los umbrales $k = 2$ y $k = 3$. Se da tanto el número absoluto de infractores como el número relativo como porcentaje del número de individuos en la muestra. En el ejemplo, cuatro observaciones violan el 2-anonimato y las 10 observaciones violan el 3-anonimato.

Bloque 7.4. Uso de la función `print()` para mostrar observaciones que violan k-anonimato

```
print(sdcInitial, 'kAnon')
```

Para otros niveles de k-anonimato, es posible calcular el número de personas infractoras utilizando los recuentos de frecuencia de muestreo en el objeto `sdcMicro`. El número de infractores es el número de personas con recuentos de frecuencia de muestreo inferiores al umbral especificado k. En el Bloque **??**, mostramos un ejemplo de cómo calcular cualquier umbral para k usando las medidas de riesgo ya almacenadas disponibles después de configurar un objeto `sdcMicro` en `R`. k se puede reemplazar con cualquier umbral requerido. La elección del umbral requerido que deben cumplir todas las personas en el archivo de microdatos depende de muchos factores y se analiza más adelante en la sección Supresión local sobre la supresión local. En muchas instituciones, los umbrales típicamente requeridos para k-anonimato son 3 y 5.

Bloque 7.5. Violaciones de k-anonimato para distintos valores de k

```
k=10
sum(sdcInitial@risk$individual[,2] < k)
```

Es importante tener en cuenta que los valores faltantes (`NA`s en `R` [7]) se tratan como si fueran cualquier otro valor. Dos personas con claves {'Masculino', `NA`,

[7] En `sdcMicro`, es importante utilizar el código de valor faltante estándar `NA` en lugar de otros

Tabla 7.2: Conjunto de datos de ejemplo para ilustrar el efecto de los valores faltantes en el k-anonimato

Id	Género	Nivel educacional	Situación laboral	f_{k}
1	Masculino	Secundaria completa	Ocupado	2
2	Masculino	Secundaria incompleta	Ocupado	2
3	Masculino	NA	Ocupado	3

'Ocupado'} y {'Masculino', 'Secundaria completa', 'Ocupado'} comparten la misma clave y, de manera similar, {'Masculino', `NA`, 'Ocupado'} y {'Masculino', 'Secundaria incompleta', 'Ocupado'} también comparten la misma clave. Por lo tanto, el valor que falta en la primera clave se interpreta primero como 'Secundaria completa' y luego como 'Secundaria incompleta'. Esto se ilustra en la Tabla **??**.

Fuente: Adaptación de (**?**, p.32)

Si un conjunto de datos satisface k-anonimato, un intruso siempre encontrará al menos k individuos con la misma combinación de identificadores indirectos. El k-anonimato suele ser un requisito necesario para la anonimización de un conjunto de datos antes de su publicación, pero no es necesariamente un requisito suficiente. La medida de k-anonimato solo se basa en recuentos de frecuencia y no tiene en cuenta (las diferencias en) los pesos de las muestras. Con frecuencia el k-anonimato se logra aplicando primero la recodificación y luego la supresión local y, en algunos casos, mediante la microagregación, antes de utilizar otras medidas de riesgo y métodos de divulgación para reducir aún más el riesgo de divulgación. Estos métodos se analizan en la sección Métodos SDC.

7.5.3 l-diversity

El k-anonimato ha sido criticado por no ser lo suficientemente restrictivo. La información confidencial puede divulgarse incluso si los datos satisfacen el k-anonimato. Esto puede ocurrir en los casos en que los datos contienen variables categóricas confidenciales (de no identificación) que tienen el mismo valor para todas las personas que comparten la misma clave. Ejemplos de tales variables sensibles son aquellas que contienen información sobre el estado de salud de un individuo. La Tabla **??** ilustra este problema utilizando los mismos datos que se utilizaron anteriormente, pero agregando una variable sensible, "salud". Los dos primeros individuos cumplen 2-anonimato para los identificadores indirectos "área", "género", "nivel educacional" y "situación laboral". Esto significa que

códigos, como 9999 o cadenas. En la sección Valores faltantes, se analizó más detalladamente cómo establecer otros códigos de valores faltantes de `NA` en `R`. Esto es necesario para garantizar que los métodos de `sdcMicro` funcionen correctamente. Cuando los valores faltantes tienen códigos distintos de `NA`, los códigos de valores faltantes se interpretan como un nivel de factor distinto en el caso de las variables categóricas.

un intruso encontrará al menos dos personas al hacer coincidir el conjunto de microdatos publicado en función de esos cuatro identificadores indirectos. Sin embargo, si el intruso sabe que alguien pertenece a la muestra y tiene la clave {'Urbano', 'Femenino', 'Secundaria incompleta' y 'Ocupado'}, con certeza se revela el estado de salud ('sí'), porque para ambos las observaciones con esta clave tienen el mismo valor. Esta información se revela así sin la necesidad de coincidir exactamente con el individuo. Este no es el caso de los individuos con clave {'Urbano', 'Masculino', 'Secundaria completa', 'Ocupado'}.

El concepto de l-diversity (distinto) aborda esta deficiencia del k-anonimato. Un conjunto de datos satisface l-diversity si para cada clave k hay por lo menos l diferentes valores para cada una de las variables sensibles. En el ejemplo, los primeros dos individuos satisfacen solo 1-diversity, los individuos 4 y 6 satisfacen 2-diversity. El nivel requerido de l-diversity depende del número de valores posibles que puede tomar la variable sensible. Si la variable sensible es una variable binaria, el nivel más alto de l-diversity que se puede conseguir es 2. Una muestra única siempre solo satisfará 1-diversity.

Para computar l-diversity para variables sensibles en `sdcMicro`, se puede usar la función `ldiversity()`. Esto se ilustra en el Bloque **??**. Como argumentos, especificamos los nombres de las variables sensibles [8] en el archivo, así como una constante para l-diversity [9] y el código de valores faltantes en los datos. La salida se guarda en el *slot* de "riesgo" del objeto `sdcMicro`. El resultado muestra el mínimo, máximo, media y cuantiles de las l-puntuaciones de diversidad para todos los individuos de la muestra. El resultado del Bloque **??** reproduce los resultados según los datos de la Tabla **??**.

Fuente: Adaptación de (**?**, p.33)

Bloque 7.6. Función para l-diversity en `sdcMicro`

```
# Calculando l-diversity

sdcInitial <- ldiversity(obj = sdcInitial, ldiv_index = c("Salud"),
                         l_recurs_c = 2, missing = NA)
# Resultado para l-diversity
sdcInitial@risk$ldiversity

# l-diversity score para cada registro
sdcInitial@risk$ldiversity[,'Salud_Distinct_Ldiversity']
```

[8] Alternativamente, las variables sensibles se pueden especificar al crear el objeto `sdcMicro` usando la función `createSdcObj()` en el argumento sensibleVar. Esto se explica con más detalle en la sección Objetos de la clase `sdcMicroObj`. En ese caso, no es necesario especificar el argumento ldiv_index en la función `ldiversity()`, y las variables en el argumento sensibleVar se usarán automáticamente para calcular l-diversity.

[9] Además de l-diversity distintos, hay otros métodos de l-diversity: entropía y recursivo. l-diversity distinto es el más utilizado.

Tabla 7.3: Ilustración de l-diversity

Id	Área	Género	Nivel educacional	Situación laboral	Salud	f_{k}	$F_{k
1	Urbano	Femenino	Secundaria incompleta	Ocupado	Enfermo	2	360
2	Urbano	Femenino	Secundaria incompleta	Ocupado	Enfermo	2	360
3	Urbano	Femenino	Primaria incompleta	No FL	Enfermo	1	215
4	Urbano	Masculino	Secundaria completa	Ocupado	Enfermo	2	152
5	Rural	Femenino	Secundaria completa	Desocupado	Enfermo	1	186
6	Urbano	Masculino	Secundaria completa	Ocupado	Sano	2	152
7	Urbano	Femenino	Primaria completa	No FL	Sano	1	180
8	Urbano	Masculino	Post secundaria	Desocupado	Enfermo	1	215
9	Urbano	Femenino	Secundaria incompleta	No FL	Sano	2	262
10	Urbano	Femenino	Secundaria incompleta	No FL	Enfermo	2	262

l-diversity es útil si los datos contienen variables sensibles categóricas que no son identificadores indirectos en sí mismos. No es posible seleccionar identificadores indirectos para calcular la l-diversity. La l-diversity debe calcularse para cada variable sensible por separado.

7.6 Medidas de riesgo para variables continuas

El principio de rareza o unicidad de combinaciones de identificadores indirectos (claves) no es útil para variables continuas, porque es probable que todos o muchos individuos tengan claves únicas. Por lo tanto, se explotan otros enfoques para medir el riesgo de divulgación de las variables continuas. Estos métodos se basan en la unicidad de los valores en la vecindad de los valores originales. La unicidad se define de diferentes formas: en términos absolutos (medida de intervalo) o en términos relativos (vinculación de registros). La mayoría de las medidas son medidas a posteriori: se evalúan después de anonimizar los datos sin procesar, comparar los datos tratados con los datos sin procesar y evaluar para cada individuo la distancia entre los valores en los datos sin procesar y los tratados. Esto significa que estos métodos no son útiles para identificar personas en riesgo dentro de los datos sin procesar, sino que muestra la distancia/diferencia entre el conjunto de datos antes y después de la anonimización y, por lo tanto, puede interpretarse como una evaluación del método de anonimización. Por esa razón, se asemejan a las medidas de pérdida de información discutidas en la sección Medición de la utilidad y la pérdida de información. Finalmente, las medidas de riesgo para identificadores indirectos continuos también se basan en la detección de valores atípicos. Los valores atípicos juegan un papel importante en la reidentificación de estos registros.

7.6.1 Vinculación de registros (o coincidencia de registros)

Vinculación de registros (o Record linkage, en inglés) es un método a posteriori que evalúa el número de vínculos correctos al vincular los valores perturbados con los valores originales. El algoritmo de vinculación se basa en la distancia entre el original y los valores perturbados (es decir, vinculación de registros basada en la distancia). Los valores perturbados se emparejan con el individuo más cercano. Es importante señalar que este método no brinda información sobre el riesgo inicial, sino que es una medida para evaluar el algoritmo de perturbación (es decir, está diseñado para indicar el nivel de incertidumbre introducido en la variable al contar la cantidad de registros que podría coincidir correctamente).

Los algoritmos de vinculación de registros difieren con respecto a qué medida de distancia se utiliza. Cuando una variable tiene una escala muy diferente a la de otras variables continuas en el conjunto de datos, se recomienda volver a escalar las variables antes de usar la vinculación de registros. Escalas muy diferentes pueden dar lugar a resultados no deseados al medir la distancia multivariada entre registros en función de varias variables continuas. Dado que estos métodos se basan tanto en datos sin procesar como en datos tratados, los ejemplos de sus aplicaciones requieren la introducción de métodos SDC y, por lo tanto, se posponen a los estudios de casos en la sección Caso de estudio.

Además de la vinculación de registros basada en la distancia, otro método de vinculación es la vinculación de registros probabilísticos. La literatura muestra, sin embargo, que los resultados de la vinculación de registros basados en la distancia son mejores que los resultados de la vinculación de registros probabilísticos. Las personas en los datos tratados que están vinculadas a las personas correctas en los datos sin procesar se consideran en riesgo de divulgación.

7.6.2 Medida de intervalo

La aplicación exitosa de un método SDC debería dar como resultado valores perturbados que no se consideran demasiado cercanos a sus valores iniciales; si el valor es relativamente cercano, la reidentificación puede ser relativamente fácil. En la aplicación de medidas de intervalo, se crean intervalos alrededor de cada valor perturbado y luego se determina si el valor original de esa observación perturbada está contenido en este intervalo. 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. 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. Este método está implementado en la función `dRisk()` en `sdcMicro`. El Bloque **??** muestra cómo imprimir o mostrar el valor de riesgo calculado por `sdcMicro` comparando las variables de ingresos antes y después de la anonimización. “sdcObj” es un objeto `sdcMicro` y

"compExp" es un vector que contiene los nombres de las variables de ingresos. El tamaño de los intervalos es k veces la desviación estándar, donde k es un parámetro en la función `dRisk()`. El más largo k, cuanto más grandes son los intervalos y, por lo tanto, mayor es el número de observaciones dentro del intervalo construidas alrededor de sus valores originales y mayor es la medida de riesgo. El resultado 1 indica que todas (100 por ciento) las observaciones están fuera del intervalo de 0,1 veces la desviación estándar alrededor de los valores originales.

Bloque 7.7. Ilustración de medida de intervalo

```
dRisk(obj = sdcObj@origData[,compExp], xm = sdcObj@manipNumVars[,compExp],
      k = 0.1)
[1] 1
```

Para la mayoría de los valores, este es un enfoque satisfactorio. Sin embargo, no es una medida suficiente para valores atípicos. Después de la perturbación, los valores atípicos seguirán siendo valores atípicos y se pueden volver a identificar fácilmente, incluso si están lo suficientemente lejos de sus valores iniciales. Por lo tanto, los valores atípicos deben tratarse con precaución.

7.6.3 Detección de valores atípicos

Los valores atípicos son importantes para medir el riesgo de reidentificación en microdatos continuos. Los datos continuos suelen estar sesgados, especialmente a la derecha. Esto significa que hay algunos valores atípicos con valores muy altos en relación con las otras observaciones de la misma variable. Algunos ejemplos son los ingresos en los datos de los hogares, donde solo unas pocas personas/hogares pueden tener ingresos muy altos, o los datos de facturación de empresas que son mucho más grandes que otras empresas de la muestra. En casos como estos, incluso si estos valores se perturban, aún puede ser fácil identificar estos valores atípicos, ya que seguirán siendo los valores más grandes incluso después de la perturbación (la perturbación habrá creado incertidumbre en cuanto al valor exacto, pero debido a que el valor comenzó mucho más lejos de otras observaciones, aún puede ser fácil vincularlo con el individuo de altos ingresos o la empresa muy grande). Los ejemplos serían el único médico en un área geográfica con altos ingresos o una sola gran empresa en un tipo de industria. Por lo tanto, la identificación de valores atípicos en datos continuos es un paso importante cuando se identifican personas con alto riesgo. En la práctica, identificar los valores de una variable continua que son mayores que un valor predeterminado p%-percentil 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.

Podemos calcular el p%-percentil de una variable continua en `R` y mostrar los

individuos que tienen ingresos superiores a este percentil. El Bloque **??** proporciona una ilustración del percentil 90.

Bloque 7.8. Cómputo del percentil 90 % de la variable INCWAGE

```
setwd("../Capacitación/GitHub") # directorio de trabajo

fname = "data.dta" # nombre del archivo
library(haven) # carga el paquete requerido para la función de lectura/escritura
               # para archivos STATA
file <- read_dta(fname)

# Cómputo de 90 % percentil para variable INCWAGE
 perc90 <- quantile(file[,'INCWAGE'], 0.90, na.rm = TRUE)

 # Muestra ID de observaciones con valores para INCWAGE mayores al 90 % percentil
 file[(file[, 'INCWAGE'] >= perc90), 'IDP']
```

Un segundo enfoque para la detección de valores atípicos es una medida a posteriori que compara los datos tratados y sin procesar. Se construye un intervalo alrededor de los valores perturbados como se describe en la sección anterior. Si los valores originales caen dentro del intervalo alrededor de los valores perturbados, los valores perturbados se consideran inseguros ya que están demasiado cerca de los valores originales. Existen diferentes formas de construir dichos intervalos, como intervalos basados en rangos e intervalos basados en desviación estándar. **?** proponen una alternativa robusta para estos intervalos. Construyen los intervalos en función de la distancia robusta de Mahalanobis (RMD, por su sigla en inglés) al cuadrado de los valores individuales. El RMD escala los intervalos de manera que los valores atípicos obtienen intervalos más grandes y, por lo tanto, deben tener una perturbación mayor para que se consideren seguros que los valores que no son atípicos. Este método se implementa en `sdcMicro` en la función `dRiskRMD()`, que es una extensión de la función `dRisk()`.

7.7 Riesgo global

Para construir una medida de riesgo agregado a nivel global para el conjunto de datos completo, podemos agregar las medidas de riesgo a nivel individual de varias maneras. Las medidas de riesgo global deben usarse con precaución: detrás de un riesgo global aceptable pueden esconderse algunos registros de muy alto riesgo que se compensan con muchos registros de bajo riesgo.

7.7.1 Media de las medidas de riesgo individuales

Una forma sencilla de agregar las medidas de riesgo individuales es tomar la media de todos los individuos de la muestra, que es igual a sumar todas las claves de la muestra si se multiplica por las frecuencias de muestra de estas claves y se divide por el tamaño de la muestra, n:

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

donde r_k es el riesgo individual de clave k que el *i-ésimo* registro comparte (ver la sección Identificadores indirectos categóricos y recuentos de frecuencia). Esta medida se informa como riesgo global en `sdcMicro`, se almacena en el *slot* de "riesgo" y se puede imprimir como se muestra en el Bloque **??**. Indica que la probabilidad de reidentificación promedio es 0,01582 o 0,1582%.

Bloque 7.9. Cómputo de la medida de riesgo global

```
# Riesgo global (probabilidad promedio de re-identificación)
 sdcInitial@risk$global$risk
```

El riesgo global en los datos de ejemplo de la Tabla **??** es 0,01582, que es la proporción esperada de todos los individuos de la muestra que un intruso podría volver a identificar. Otra forma de expresar el riesgo global es el número de reidentificaciones esperadas, $n * R_1$, que es en el ejemplo 10 * 0,01582. El número esperado de reidentificaciones también se guarda en el objeto `sdcMicro`. El Bloque **??** muestra cómo imprimir esto.

Bloque 7.10. Cómputo del número esperado de reidentificaciones

```
#  # Riesgo global(Número esperado de re-identificaciones)
 sdcInitial@risk$global$risk_ER
```

7.7.2 Recuento de personas con riesgos superiores a un cierto umbral

Todos los individuos pertenecientes a la misma clave tienen el mismo riesgo individual, r_k. Otra forma de expresar el riesgo total en la muestra es el número total de observaciones que superan un determinado umbral de riesgo individual. La fijación del umbral puede ser absoluta (por ejemplo, todas aquellas personas que tengan un riesgo de divulgación superior a 0,05 o 5%) o relativa (por ejemplo, todas aquellas personas con riesgos superiores al cuartil superior del riesgo individual). El Bloque **??** muestra cómo utilizando `R`, se contaría el número de observaciones con un riesgo de reidentificación individual superior al 5%. En el ejemplo, ninguna persona tiene un riesgo de divulgación superior a 0,05.

Bloque 7.11. Número de personas con riesgo individual superior al umbral 0,05

```
sum(sdcInitial@risk$individual[,1] > 0.05)
```

Estos cálculos se pueden usar para tratar los datos de las personas cuyos valores de riesgo están por encima de un umbral predeterminado. Más adelante veremos que hay métodos en `sdcMicro`, como `localSupp()`, que se pueden usar para suprimir valores de ciertas variables clave para aquellas personas con riesgo por encima de un umbral específico. Esto se explica con más detalle en la sección Supresión local.

7.8 Riesgo jerárquico (o del hogar)

En muchas encuestas sociales, los datos tienen una estructura jerárquica donde un individuo pertenece a una entidad de nivel superior (ver la sección Niveles de riesgo). Ejemplos típicos son los hogares en las encuestas sociales o los alumnos en las escuelas. La reidentificación de un miembro del hogar también puede conducir a la reidentificación de los otros miembros del hogar. Por tanto, es fácil ver que, si tenemos en cuenta la estructura del hogar, el riesgo de reidentificación es el riesgo de que al menos uno de los miembros del hogar sea reidentificado.

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

donde A_j es el evento que el *j-ésimo* miembro del hogar sea 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 en función de sus respectivas claves 0,02, 0,03 y 0,03, respectivamente, el riesgo del hogar es

$$1 - ((1 - 0,02)(1 - 0,03)(1 - 0,03)) = 0,078$$

El riesgo jerárquico o del hogar no puede ser menor que el riesgo individual, y el riesgo del hogar es siempre el mismo para todos los miembros del hogar. El riesgo del hogar debe utilizarse en los casos en que los datos contengan una estructura jerárquica, es decir, cuando la estructura del hogar esté presente en los datos. Usando `sdcMicro`, si se especifica un identificador de hogar (en el argumento hhId en la función `createSdcObj()`) al crear un objeto `sdcMicro`, el riesgo del hogar se calculará automáticamente. El Bloque **??** muestra cómo imprimir estas medidas de riesgo.

Bloque 7.12. Cómputo del riesgo del hogar y número esperado de reidentificaciones

```
# Riesgo del hogar
 sdcInitial@risk$global$hier_risk

 # Riesgo del hogar (Número esperado de reidentificaciones)
 sdcInitial@risk$global$hier_risk_ER
```

El tamaño de un hogar es un identificador importante en sí mismo, especialmente para hogares grandes. Sin embargo, la supresión de la variable del tamaño real (por ejemplo, el número de miembros del hogar) no es suficiente para eliminar esta información del conjunto de datos, ya que un simple recuento de los miembros del hogar para un hogar en particular permitirá reconstruir esta variable siempre que un ID del hogar esté en los datos, lo que permite asignar individuos a los hogares. Señalamos esto para la atención del lector ya que es importante.

7.9 Referencias

Capítulo 8

Métodos SDC

Esta sección describe los métodos SDC más utilizados. Todos los métodos se pueden implementar en `R` utilizando el paquete `sdcMicro`. Discutimos qué método es más adecuado para cada tipo de datos, tanto en términos de características como del tipo de dato. Además, se discuten opciones como los parámetros específicos de cada método, así como sus impactos. Las conclusiones pretenden ser orientativas, pero deben utilizarse con precaución, ya que cada operación estadística genera datos con características diferentes y las recomendaciones del documento no siempre serán las más adecuadas para sus datos en particular.

Para determinar qué métodos de anonimización son adecuados para variables y/o conjuntos de datos específicos, comenzamos presentando algunas clasificaciones de los métodos SDC.

8.1 Clasificación de los métodos SDC

Los métodos SDC pueden clasificarse en no perturbativos y perturbativos (**?**).

- **Los métodos no perturbativos** reducen el detalle de los datos mediante la generalización o la supresión de ciertos valores (enmascaramiento) sin distorsionar la estructura de los datos.
- **Los métodos perturbativos** no suprimen los valores del conjunto de datos, sino que alteran los valores para limitar el riesgo de divulgación creando incertidumbre en torno a los valores reales. Tanto los métodos no perturbativos como los perturbativos pueden utilizarse para variables categóricas y continuas.

También distinguimos entre métodos probabilísticos y deterministas SDC.

Tabla 8.1: Métodos SDC y funciones correspondientes en 'sdcMicro'

Método	Clasificación del método SDC	Tipo de datos	Función en
Recodificación Global	no perturbativo, determinista	continuo y categórico	'globalReco
Codificación Superior e Inferior	no perturbativo, determinista	continuo y categórico	'topBotCod
Supresión Local	no perturbativo, determinista	categórico	'localSuppr
PRAM	perturbativo, probabilístico	categórico	'pram'
Micro agregación	perturbativo, probabilístico	continuo	'microaggre
Adición de Ruido	perturbativo, probabilístico	continuo	'addNoise'
Shuffling	perturbativo, probabilístico	continuo	'shuffle'
Rank swapping	perturbativo, probabilístico	continuo	'rankSwap'

Los **métodos probabilísticos** dependen de un mecanismo de probabilidad o de un mecanismo de generación de números aleatorios. Cada vez que se utiliza un método probabilístico, se genera un resultado diferente. Para estos métodos se suele recomendar que se establezca una semilla (con la función `set.seed()`) para el generador de números aleatorios si se quiere producir resultados replicables.

Los **métodos deterministas** siguen un algoritmo determinado y producen los mismos resultados si se aplican repetidamente a los mismos datos con el mismo conjunto de parámetros.

Los métodos SDC para microdatos pretenden evitar la revelación de identidad y de atributos. Para cada tipo de control de la divulgación se utilizan diferentes métodos SDC. Métodos como la recodificación y la supresión local se aplican a los identificadores indirectos para evitar la divulgación de identidad, mientras que la codificación superior de un identificador indirecto (por ejemplo, los ingresos) o la perturbación de una variable sensible evitan la divulgación de atributos.

Discutiremos los métodos SDC que se implementan en el paquete `sdcMicro` o que pueden implementarse fácilmente en R. Estos son los métodos más comúnmente aplicados en la literatura y utilizados en la mayoría de las agencias con experiencia en el uso de estos métodos. La Tabla **??** ofrece una visión general de los métodos de SDC discutidos en esta guía, su clasificación, los tipos de datos a los que son aplicables y los nombres de sus funciones en el paquete `sdcMicro`.

8.2 Métodos no perturbativos

8.2.1 Recodificación

La recodificación es un método determinista utilizado para disminuir el número de categorías o valores distintos de una variable. Se realiza combinando o agru-

pando categorías para las variables categóricas o construyendo intervalos para las variables continuas. La recodificación se aplica a todas las observaciones de una determinada variable y no solo a las que corren el riesgo de ser reveladas. Existen dos tipos generales de recodificación: la recodificación global y la codificación superior e inferior.

8.2.1.1 Recodificación global

La recodificación global combina varias categorías de una variable categórica o construye intervalos para variables continuas. Esto reduce el número de categorías disponibles en los datos y, potencialmente, el riesgo de divulgación, especialmente para las categorías con pocas observaciones, pero también, y esto es importante, reduce el nivel de detalle de la información disponible para el analista. Para ilustrar la recodificación, utilizamos el siguiente ejemplo. Supongamos que tenemos cinco regiones en nuestro conjunto de datos. Algunas regiones son muy pequeñas y, cuando se combinan con otras variables clave del conjunto de datos, producen un alto riesgo de reidentificación para algunos individuos de esas regiones. Una forma de reducir el riesgo sería combinar algunas de las regiones al recodificarlas. Podríamos, por ejemplo, hacer tres grupos de los cinco, llamarlos "Norte", "Centro" y "Sur" y en consecuencia reetiquetar los valores. De este modo, el número de categorías de la región variable se reduce de cinco a tres.

> **Nota**: Cualquier agrupación debe ser una agrupación pertinente para los objetivos analíticos de la operación estadística y no una unión aleatoria de categorías.

Algunos ejemplos serían agrupar las comunas en provincias, las regiones en macrozonas o las categorías detalladas de agua limpia. Agrupar todas las regiones pequeñas sin proximidad geográfica no es necesariamente la mejor opción desde el punto de vista de los servicios públicos. La Tabla **??** lo ilustra con un conjunto de datos de ejemplo muy simplificado. Antes de la recodificación, tres individuos tienen claves distintas, mientras que después de la recodificación (agrupando la "Región 1" y la "Región 2" en el "Norte", la "Región 3" en el "Centro" y la "Región 4" y la "Región 5" en el "Sur"), el número de claves distintas se reduce a cuatro y la frecuencia de cada clave es de al menos dos, basándose en los tres identificadores indirectos seleccionados. Los recuentos de la frecuencia de las claves fk se muestran en la última columna de la Tabla **??**. Un intruso encontraría al menos dos individuos para cada clave y no podría distinguir más entre los individuos 1 - 3, los individuos 4 y 6, los individuos 5 y 7 y los individuos 8 - 10, basándose en las variables clave seleccionadas.

La recodificación suele ser el primer paso de un proceso de anonimización. Puede utilizarse para reducir el número de combinaciones únicas de valores de las variables clave. Por lo general, esto aumenta los recuentos de frecuencia de la mayoría de las claves y reduce el riesgo de divulgación. La reducción del número de

Tabla 8.2: Ilustración del efecto de la recodificación en los recuentos de frecuencia de las variables clave

Antes de recodificar					Después de recodificar			
Individuo	Región	Sexo	Religión	f_k	Región	Sexo	Religión	f_k
1	Región 1	Mujer	Católica	1	Norte	Mujer	Católica	3
2	Región 2	Mujer	Católica	2	Norte	Mujer	Católica	3
3	Región 2	Mujer	Católica	2	Norte	Mujer	Católica	3
4	Región 3	Mujer	Protestante	2	Centro	Mujer	Protestante	2
5	Región 3	Hombre	Protestante	1	Centro	Hombre	Protestante	2
6	Región 3	Mujer	Protestante	2	Centro	Mujer	Protestante	2
7	Región 3	Hombre	Protestante	2	Centro	Hombre	Protestante	2
8	Región 4	Hombre	Musulmán	2	Sur	Hombre	Musulmán	3
9	Región 4	Hombre	Musulmán	2	Sur	Hombre	Musulmán	3
10	Región 5	Hombre	Musulmán	1	Sur	Hombre	Musulmán	3

Tabla 8.3: Ilustración del efecto de la recodificación en el número de combinaciones teóricamente posibles de un conjunto de datos

Número de categorías	Región	Estado civil	Edad	Posibles combinaciones
antes de la recodificación	20	8	100	16.000
después de la recodificación	6	6	15	540

combinaciones posibles se ilustra en la Tabla **??** con los identificadores indirectos "región", "estado civil" y "edad". La Tabla **??** muestra el número de categorías de cada variable y el número de combinaciones teóricamente posibles, que es el producto del número de categorías de cada identificador indirecto, antes y después de la recodificación. La "edad" se interpreta como una variable semicontinua y se trata como una variable categórica. El número de combinaciones posibles y, por tanto, el riesgo de reidentificación se reducen en gran medida con la recodificación. Hay que tener en cuenta que el número de combinaciones posibles es un número teórico; en la práctica, pueden incluirse combinaciones muy improbables, como edad = 3 y estado civil = viudo, y el número real de combinaciones en un conjunto de datos puede ser inferior.

Los principales parámetros para la recodificación global son el tamaño de los nuevos grupos, así como la definición de los valores que se agrupan en las nuevas categorías.

> **Nota**: Hay que tener cuidado de elegir las nuevas categorías, deberían generarse en función del uso de los datos por parte de los usuarios finales y de minimizar la pérdida de información como re-

sultado de la recodificación.

Podemos observarlo mediante tres ejemplos:

- *Variable de edad*: Las categorías de edad deben elegirse de forma que sigan permitiendo a los usuarios de los datos realizar cálculos relevantes para el tema que se está estudiando. Por ejemplo, si es necesario calcular indicadores para niños de edades comprendidas entre los 6 y los 11 años y entre los 12 y los 17 años, además, es necesario agrupar la edad para reducir el riesgo, hay que tener cuidado de crear intervalos de edad que sigan permitiendo realizar los cálculos. Una agrupación satisfactoria podría ser, por ejemplo, 0 - 5, 6 - 11, 12 - 17, etc., mientras que una agrupación 0 - 10, 11 - 15, 16 - 18 destruiría la utilidad de los datos para estos usuarios. Aunque es una práctica habitual crear intervalos (grupos) de igual anchura (tamaño), también es posible (si los usuarios de los datos lo requieren) recodificar solo una parte de las variables y dejar algunos valores como estaban originalmente. Esto podría hacerse, por ejemplo, recodificando todas las edades superiores a 20 años, pero dejando las inferiores a 20 años tal y como están. Si los métodos SDC distintos de la recodificación se van a utilizar más tarde o en un paso siguiente, hay que tener cuidado al aplicar la recodificación solo a una parte de la distribución, ya que esto podría aumentar la pérdida de información debida a los otros métodos, ya que la agrupación no protege las variables no agrupadas. La recodificación parcial seguida de métodos de supresión como la supresión local puede, por ejemplo, conducir a un número de supresiones mayor del deseado o necesario en caso de que la recodificación se realice para todo el rango de valores (ver la siguiente sección de **??**). En el ejemplo anterior, el número de supresiones de los valores inferiores a 20 será probablemente mayor que para los valores del rango recodificado. El número desproporcionadamente alto de supresiones en este rango de valores que no se recodifican puede conducir a una mayor pérdida de utilidad para estos grupos.
- *Variables geográficas*: Si los datos originales especifican información de nivel administrativo en detalle, por ejemplo, hasta el nivel de comuna, entonces potencialmente esos niveles inferiores podrían ser recodificados o agregados en niveles administrativos superiores, por ejemplo, la provincia, para reducir el riesgo. Al hacerlo, hay que tener en cuenta lo siguiente: La agrupación de comunas en niveles abstractos que se cruzan con diferentes provincias haría que el análisis de datos a nivel comunal o provincial fuera un reto. Se debe tener cuidado de entender lo que el usuario requiere y la intención del estudio. Si un componente clave de la encuesta es realizar un análisis a nivel comunal, la agregación a nivel provincial podría perjudicar la utilidad de los datos para el usuario. La recodificación debería aplicarse si el nivel de detalle de los datos no es necesario para la mayoría de los usuarios de los datos y para evitar un gran número de supresiones cuando se utilicen posteriormente otros métodos SDC. Si los usuarios necesitan

información a un nivel más detallado, otros métodos, como los Métodos perturbativos, podrían ofrecer una solución mejor que la recodificación.

- *Instalaciones sanitarias*: Un ejemplo de una situación en la que un alto nivel de detalle podría no ser necesario y la recodificación podría hacer muy poco daño a la utilidad es el caso de una variable detallada de instalaciones sanitarias en el hogar que enumera las respuestas para 20 tipos de inodoros. Es posible que los investigadores solo necesiten distinguir entre instalaciones de inodoros mejoradas y no mejoradas y que no necesiten la clasificación exacta de hasta 20 tipos. La información detallada de los tipos de inodoros puede utilizarse para volver a identificar a los hogares, mientras que la recodificación en dos categorías -instalaciones mejoradas y no mejoradas- reduce el riesgo de reidentificación y, en este contexto, apenas reduce la utilidad de los datos. Este enfoque puede aplicarse a cualquier variable con muchas categorías en las que los usuarios de los datos no estén interesados en los detalles, sino en algunas categorías agregadas. La recodificación aborda la agregación para los usuarios de los datos y al mismo tiempo protege los microdatos. Es importante hacer un balance de las agregaciones utilizadas por los usuarios.

La recodificación debe aplicarse solo si la eliminación de la información detallada de los datos no perjudica a la mayoría de las personas usuarias. Si los usuarios necesitan información a un nivel más detallado, entonces la recodificación no es apropiada y otros métodos, como los perturbativos, podrían funcionar mejor.

En `sdcMicro` existen diferentes opciones de recodificación global. En los siguientes párrafos, damos ejemplos de recodificación global con las funciones `groupAndRename()` y `globalRecode()`. La función `groupAndRename()` se utiliza generalmente para las variables categóricas y la función `globalRecode()` para las variables continuas. Por último, discutimos el uso del redondeo para reducir el detalle en las variables continuas.

8.2.1.1.1 Recodificación de una variable categórica mediante la función `sdcMicro groupAndRename()`

Supongamos que se ha creado un objeto de la clase `sdcMicro`, que se llama `sdcInitial` (véase el apartado Objetos de la clase `sdcMicroObj` cómo crear objetos de la clase `sdcMicro`). En el Bloque **??**, la variable "sizeRes" tiene cuatro categorías diferentes: "capital", "ciudad grande", "ciudad pequeña", "pueblo" y "campo"). Las tres primeras se recodifican o reagrupan como "urbano" y la categoría "campo" pasa a llamarse "rural". En los argumentos de la función, especificamos las categorías que se van a agrupar (anterior) y los nombres de las categorías después de la recodificación (posterior). Es importante que los vectores "anterior" y "posterior" tengan la misma longitud. Por lo tanto, tenemos que repetir "urbano" tres veces en el vector (posterior) para que coincida con los tres valores diferentes que se recodifican en "urbano".

Nota: La función `groupAndRename()` solo funciona en variables tipo factor.

Nos referimos a la sección Clases en R sobre cómo cambiar la clase de una variable.

Cargar librerías.

```
require(dplyr)
require(foreign)
require(sdcMicro)
```

Cargaremos la base en formato de dta (stata).

```
#directorio de trabajo
#getwd()
fname <- "data/data.dta"
file <- read.dta(fname, convert.factors = TRUE)
```

Ajustaremos las variables a factores.

```
# Crear variables área y etnia para evaluar recodificación

area_names <- c("capital, large city", "small city", "town", "countryside")
area <- sample(area_names[1:3], nrow(file[file$URBRUR==1,]), replace=TRUE, prob=c(0.60,0.27, 0.13

file <- file %>% mutate(sizeRes = ifelse(URBRUR==1, area, area_names[4])) %>% relocate(sizeRes, .
  mutate(sizeRes = factor(sizeRes, levels = area_names))

etnia_names <- c("mapuche","diaguita","atacameno","otra","No aplica")
etnia <- sample(etnia_names, nrow(file), replace=TRUE, prob=c(0.2,0.1,0.1,0.1, 0.95))


file <- file %>% mutate(etnia = etnia) %>% mutate(etnia = factor(etnia, levels = etnia_names))


selectedKeyVarsHH = c("sizeRes","AGEYRS", "GENDER", "REGION", "etnia","RELIG") #,
#selectedKeyVarsHH = c("URBRUR", "REGION", "HHSIZE", "OWNAGLAND", "RELIG")

file$URBRUR    <- as.factor(file$URBRUR)
file$REGION    <- as.factor(file$REGION)
file$OWNHOUSE  <- as.factor(file$OWNHOUSE)
file$OWNAGLAND <- as.factor(file$OWNAGLAND)
file$RELIG     <- as.factor(file$RELIG)
```

```
numVarsHH      <- c("LANDSIZEHA", "TANHHEXP", "TFOODEXP", "TALCHEXP", "TCLTHEXP", "THOU
                    "TFURNEXP", "THLTHEXP", "TTRANSEXP", "TCOMMEXP", "TRECEXP", "TEDUEX
                    "TRESTHOTEXP", "TMISCEXP", "INCTOTGROSSHH", "INCRMT", "INCWAGE", "I
                    "INCNFARMBSN", "INCRENT", "INCFIN", "INCPENSN", "INCOTHER")

pramVarsHH     <- c("ROOF", "TOILET", "WATER", "ELECTCON", "FUELCOOK", "OWNMOTORCYCLE"
weightVarHH    <- c("WGTPOP")

# Ajuste strata

file$strata_region <- file$REGION
strata_var <- c("strata_region")

# Ajuste para transformar factores
file[,c(pramVarsHH)] <- lapply(file[,c(pramVarsHH)], as.factor)


HHVars <- c("IDH", selectedKeyVarsHH, pramVarsHH, numVarsHH, weightVarHH, strata_var) #
fileHH <- file
fileHH <- fileHH[which(!duplicated(fileHH$IDH)),]
sdcHH <- createSdcObj(dat=fileHH, keyVars=selectedKeyVarsHH, pramVars=pramVarsHH, weigh

sdcInitial <- sdcHH
sdc_respaldo <- sdcHH
```

Bloque 8.1. Uso de la función `sdcMicro groupAndRename()` para recodificar una variable categórica

```
# Frequencias de sizeRes antes de recodificar
table(sdcInitial@manipKeyVars$sizeRes)
```

```
##
## capital, large city          small city                town         countryside
##                 786                 359                 171                 684
```

```
# Recodificar urbano
sdcInitial  <-  groupAndRename(sdcInitial, var = "sizeRes",
                          before = c("capital, large city", "small city", "town"),
                          after = c("urban", "urban", "urban"))

# Recodificar rural
sdcInitial  <-  groupAndRename(obj = sdcInitial, var = c("sizeRes"),
                          before = c("countryside"), after = c("rural"))
```