DANE | SEN 2.0 Sistema Estadístico Nacional-Colombia

Lineamientos para la integración y el reemplazo de fuentes directas por registros administrativos

Sistema Estadístico Nacional (SEN)

Departamento Administrativo Nacional de Estadística (DANE)

B. Piedad Urdinola Contreras
Directora

Andrea Ramírez Pisco
Subdirectora

Álvaro Fernando Guzmán Lucero
Secretario General

Directores técnicos

Javier Sebastián Ruiz Santacruz
Dirección de Censos y Demografía

Diana María Bohórquez Losada
Dirección de Difusión y Cultura Estadística

Elkin Ernesto Ramírez Niño
Dirección de Geoestadística

César Mauricio López Alfonso
Dirección de Metodología y Producción Estadística

Liliana Ibeth Avila Robles (e)
Dirección de Recolección y Acopio

Juan Pablo Cardoso Torres
Dirección de Síntesis y Cuentas Nacionales

Julieth Alejandra Solano Villa
Dirección de Regulación, Planeación, Estandarización y Normalización

Elaboración del documento

Dirección Técnica de Recolección y Acopio (DRA)

Daniel Felipe Ortiz Cuevas
German Leónidas Orjuela Borda
Gina Fernanda Otalora
Jacobo Rozo Álzate
Juan Sebastian Ordoñez

Revisión

Jose Alejandro Rojas
Dirección de Recolección y Acopio

Federico Segui
Consultor BID

Portada

Jazmin Carolina Waltero Arguello

Corrección de estilo

Sonia Marcela Naranjo

Maquetación y diagramación

Jazmin Carolina Waltero Arguello

Contenido

Figuras

Tablas

Siglas y acrónimos

DANE Departamento Administrativo Nacional de Estadística
RR. AA. Registros administrativos
INE Instituto Nacional de Estadística
IPPR Índice de Precios de la Propiedad Residencial
SNR Superintendencia de Notariado y Registro
UAECD Unidad Administrativa Especial de Catastro Distrital
PILA Planilla Integrada de Liquidación de Aportes
RELAB Registro Estadístico de Relaciones Laborales
ECV Encuesta Nacional de Calidad de Vida
SIMAT Sistema Integrado de Matrícula
RLCPD Registro para la Localización y Caracterización de Personas con Discapacidad
CNPV Censo Nacional de Población y Vivienda 2018
DIREDU Directorio Estadístico de Sedes Educativas
UARIV Unidad para la Atención y Reparación Integral a las Víctimas
SICODE Sistema de Identificación y Caracterización de Oferta y Demanda Estadística
DPS Departamento para la Prosperidad Social
REBP Registro Estadístico Base de Población
GIT Grupo Interno de Trabajo

Introducción

Los registros administrativos (RR. AA.) son usualmente aprovechados por los institutos nacionales de estadísticas (INE) en cuatro principales aplicaciones: la construcción de registros estadísticos; la creación de marcos estadísticos para el muestreo de las encuestas; el apoyo en la producción de censos de población basados en registros, y el complementar o el apoyar la producción estadística de encuestas en una etapa distinta al muestreo[1].

El presente documento brinda los lineamientos internos del Departamento Administrativo Nacional de Estadística (DANE) y una guía para las entidades del Sistema Estadístico Nacional (SEN), para la ejecución de los procedimientos relacionados con la cuarta aplicación. En dicho contexto, el objetivo de este documento es proporcionar lineamientos para la integración de fuentes primarias o directas[2] con registros administrativos en la producción estadística; cabe destacar que los lineamientos son aplicables, tanto para la integración de registros administrativos como para el reemplazo de información en encuestas con registros administrativos o estadísticos.

La incorporación de nuevas fuentes de información, como registros administrativos o fuentes alternativas a los sistemas estadísticos tiene antecedentes en una amplia literatura que aborda la estimación de los errores de medición de las variables incluidas en las diferentes encuestas. Entre estos antecedentes se destacan la estimación del ingreso laboral de individuos (Angel et al., 2019; Bound et al., 1994; Moore et al., 200; Kim y Tamborini, 2014). Lo anterior implica que integrar RR. AA. con encuestas busca complementar la información de las fuentes primarias y reemplazar algunas variables con el fin de mejorar la calidad de la información.

El aprovechamiento estadístico de los RR. AA. también involucra tipos de riesgos que se diferencian de los riesgos de las operaciones estadísticas de encuestas. Entre estos riesgos, es posible encontrar que los microdatos necesarios para las integraciones no estén disponibles en el momento requerido, problemas de cobertura, cambios en el modelo de datos sin previo aviso, cambios en la normatividad que sustenta el registro, la falta de cooperación de las entidades administradoras de los RR. AA. para garantizar la calidad, entre otros. Algunos de estos riesgos son mitigados con las acciones que se plantean en los lineamientos desarrollados en este documento.

Asimismo, los procedimientos de reemplazo pueden contribuir al fortalecimiento de las capacidades estadísticas necesarias para obtener resultados más sólidos en el futuro, aprovechando de esta manera los RR. AA. disponibles. En este sentido, EUROSTAT (2017) señala que “en algunos casos, las nuevas fuentes de datos no podrán replicar los resultados de los procesos tradicionales, pero pueden

1 Wallgren y Wallgren, 2021.
2 Hace referencia a datos que recoge directamente la organización por medio de encuestas.

producir estadísticas de mayor relevancia para los usuarios".

Al completar la implementación de estos lineamientos, el equipo temático responsable y las instancias de validación de la entidad dispondrán de la información necesaria para tomar decisiones relacionadas con la publicación de los resultados. Esta toma de decisiones depende de varios factores, incluida la madurez de los registros administrativos involucrados en términos de su consolidación, su estabilidad y la confianza en el proveedor.

Los lineamientos hacen parte de la creación de nuevas capacidades de producción de estadísticas oficiales, fomentada por los principios del SEN en relación con la innovación. Estos lineamientos abordan cinco etapas secuenciales interrelacionadas que es recomendable ejecutar en cualquier procedimiento de integración y reemplazo: definición de objetivos y justificación; acopio de registros administrativos; validación de unidades de análisis homogéneas; integración de los microdatos de forma determinística, probabilística e integración de datos geográficos; e validación de indicadores de cobertura de dos vías. Estas etapas tienen dos procesos transversales que son descritos en la sección 4.1: instancias de control y aprobación e instrumentos del procedimiento.

Cabe anotar que el alcance de este documento es establecer un esquema general a seguir cuando se integran fuentes primarias con registros administrativos en la producción estadística de encuestas, derivado de la revisión de buenas prácticas internacionales y de la experiencia que ha venido desarrollando el DANE en la materia. En consecuencia, los lineamientos presentan de manera general los métodos comúnmente utilizados en las diferentes etapas de la integración y reemplazo, sin que se busque contar con un compendio exhaustivo de los mismos.

1. Fuentes de información y antecedentes

Definición de fuentes de información

El Instituto Nacional de Estadística de Uruguay (INE) trabaja usualmente con tres tipos de información: censos, encuestas muestrales (conocidas en inglés como sample survey) y RR. AA. (Wallgren y Wallgren, 2021). Los censos se centran en recolectar información de todos los miembros de una población. Debido al propósito de los censos, las oficinas estadísticas se centran en identificar las características de la población como sexo, edad, pertenencia étnica, nivel cultural, situación económica, entre otros aspectos que representan las condiciones de vida, pero no logran capturar variables más sensibles como los ingresos del hogar. Por su parte, las encuestas muestrales recolectan información de manera aleatoria de una muestra de los miembros de una población; por la naturaleza de estas encuestas, las oficinas estadísticas deben identificar y cubrir una población objetivo, la cual busca representar a la población total o en las desagregaciones geográficas definidas por el diseño muestral (Chun et al., 2021; Wallgren y Wallgren, 2021). Es importante agregar a las fuentes de información primaria aquellas encuestas que no tiene un diseño muestral, sino que se focalizan por medio de estimar una población objetivo, para el caso de Colombia, tenemos la Encuesta Anual de Servicios y la Encuesta Anual Manufacturera. Las recomendaciones de este texto también sirven para este tipo de fuentes.

Por otro lado, los RR. AA. están formados por información que se recolecta con un propósito no estadístico, es decir, como un insumo fundamental para cumplir objetivos asociados a su misionalidad o razón social, así como para la operación y la ejecución de los diferentes programas del gobierno u otras entidades (Zhang, 2021; Wallgren y Wallgren, 2021). Estos tienen información de un subconjunto de la población con características particulares con base en la función del registro mismo. Por ejemplo, los RR. AA. de transferencias monetarias o subsidios contienen información de beneficiarios de ayudas sociales, es decir, un grueso de la población con menores recursos. Por su parte, el RR. AA. de pensionados tiene información de un grupo de personas, en principio mayores, que cuentan con una pensión, ya sea por vejez, invalidez o sobrevivencia. Así es que los RR. AA. adquieren mayor valor para el aprovechamiento estadístico cuando han recolectado información durante muchos años, con lo que las oficinas estadísticas pueden asegurar la continuidad de la producción estadística (Künn, 2015; Chun et al., 2021).

Antecedentes internacionales de institutos oficiales de estadísticas

Noruega

En el Instituto Nacional de Estadísticas de Noruega (Statistics Norway, 2012) se producen el 85% de las estadísticas oficiales de Noruega. Este instituto cuenta con el apoyo de otros productores que forman parte del Consejo de Estadísticas, formado por 25 productores de estadística y algunos

propietarios de registros. Desde 1950 y 1960 las estadísticas oficiales se basan en registros administrativos, para reducir la carga en la recopilación de datos, por lo que la información del instituto noruego procede de dos fuentes principales: los RR. AA. y los cuestionarios de encuestas. Asimismo, el instituto ha mantenido una estrategia con tres componentes principales: contar con un sistema de registros administrativos adecuado y establecer un marco legislativo que regule el intercambio de información; conocer la calidad de los registros administrativos que se recopilan para mantener una combinación adecuada al realizar los principales análisis estadísticos, y realizar la transformación y la combinación de la información que procede de diversas fuentes para lograr una producción estadística de calidad. Además, el instituto aplica algunos métodos estadísticos que deben utilizarse al vincular los datos de los registros y las encuestas a nivel individual como, por ejemplo, estratificación posterior, rastrillaje y calibración y estimación de áreas pequeñas[3].

Por otra parte, el acceso a los RR. AA. está reglamentado por la Ley de Estadísticas, que le permite al INE de Noruega elaborar estadísticas a partir de las cifras existentes. Sin embargo, si los datos no están disponibles en un registro administrativo, la información puede recogerse a través de cuestionarios electrónicos que se realizan a empresas o particulares; en este proceso el instituto noruego se encarga de minimizar las cargas a las empresas y los individuos para facilitar el uso de los datos administrativos noruegos en la producción de estadísticas.

Suecia

El Instituto de Estadísticas de Suecia (Statistics Sweden[4]) es responsable del 40% de los productos o las encuestas en las estadísticas oficiales y produce la mitad de las encuestas de las que son responsables otras agencias. Hay 24 agencias gubernamentales adicionales que son responsables de áreas temáticas de las estadísticas oficiales. Además, tiene acceso a todos los datos administrativos necesarios (lo anterior también aplica para las demás agencias gubernamentales). Esta INE cuenta con alrededor de 50 registros estadísticos[5] y 70 registros para diagnóstico médico; los muestreos se calculan basados en estos registros, que representan a la población, a las empresas y demás transacciones mercantiles. Entrando en detalle el instituto sueco utiliza algunas variables de los RR. AA. para la estratificación y como variables auxiliares durante el proceso de estimación y calibración para ajustar la falta de respuesta o para definir con mayor precisión los muestreos.

Para vincular los RR. AA., el instituto usa una única llave de integración, con la cual es posible combinar 125 registros de población, constituyendo el registro longitudinal más grande del Instituto de Estadísticas de Suecia. Este registro cuenta con una alta calidad en las llaves

3 Estratificación posterior: en este proceso se utilizan los totales poblacionales de una o varias variables para ajustar los ponderadores (INE Uruguay, Encuesta Longitudinal de Protección Social).
Rastrillaje y calibración: es un proceso de calidad posterior a la toma de los datos, donde se evalúan los resultados del muestreo (INE Uruguay, Encuesta Longitudinal de Protección Social).

4 Disponible en https://unstats.un.org/unsd/dnss/docViewer.aspx?docID=666#start

5 Disponible en http://www.asasrms.org/Proceedings/y2011/Files/300347_64422.pdf

de integración. No obstante, puede tener algunos errores o inconsistencias que son solventados por los diferentes equipos del instituto de estadísticas.

Chile

Chile tiene un contexto institucional que se rige por un marco legal amparado por la Ley 17.374 de 1970, soportado en diferentes leyes de la constitución. Esta ley ordena tres aspectos generales dentro del INE de Chile: las funciones y las organizaciones del INE; regula el acceso a datos de personas naturales y jurídicas (incluidas las instituciones públicas), y establece el secreto estadístico como norma de protección de datos de los informantes. Este marco legal garantiza para la actividad estadística unilateralidad en el acceso a los RR. AA., por lo que se puede acceder a cierta información de registros para transformarla en información estadística. La materialización del cuerpo legal (acceso a registros) se hace por medio de convenios de colaboración con instituciones públicas, formalizando el qué, cómo y cuándo se entregan los datos de las partes.

Uruguay

Para 2021 el INE de Uruguay (2021) estaba terminando la implementación del Data Warehouse Estadístico, que es la arquitectura de base que soporta el sistema integrado de registros estadísticos y que está basado en el modelo de los países escandinavos. Cuenta con un sistema de tres registros base (población, inmuebles y empresas/entidades), los cuales se integran a nivel de microdatos por medio de llaves de identificación con otros registros dentro del mismo sistema (el Registro Único Tributario para el caso de las empresas, la Cédula de Identidad para la población y un ID de domicilio para el caso de inmuebles).

Para la ronda de censos 2030, el INE planea llevar a cabo un piloto de censo a partir de RR.AA. en paralelo al centro tradicional y aunque reconocen que se está lejos de realizar censos basados en registros administrativos, argumentan que deben comenzar por esto. Para 2023, se pretende comparar los datos obtenidos de forma tradicional y los de los registros administrativos, para comparar la cobertura y calidad de los datos.

Antecedentes del DANE

El 2020 trajo la coyuntura de la pandemia del COVID-19 y consigo grandes retos en términos de producción estadística. Entre marzo y julio de 2020, el DANE tuvo que implementar un proceso de adaptación de la Gran Encuesta Integrada de Hogares (GEIH) pasando de una metodología de recolección presencial a un operativo telefónico, lo que implicó una reducción importante del formulario tradicional (DANE, 2021). En aras de preservar la continuidad de la producción estadística de las medidas de pobreza monetaria y desigualdad, así como mantener una coherencia en los datos, el DANE decidió emplear información proveniente de RR. AA. de ayudas institucionales (transferencias monetarias) y pensiones como fuente de contraste de la información recolectada en la GEIH 2020 (DANE, 2021).

El aprovechamiento de RR. AA. permitió identificar “una disminución en la cobertura del operativo de recuperación de ingresos con relación al universo que la GEIH tradicionalmente identifica” (DANE, 2021). El anterior escenario motivó a que el DANE evaluara la posibilidad de integrar

de manera regular los RR. AA. de ayudas institucionales y pensiones con la GEIH, con el fin de mejorar la medición de dichas fuentes de ingresos.

Si bien, uno de los principales usos de los RR. AA. por parte del DANE ha sido la integración con la GEIH para producir las estimaciones de pobreza monetaria y desigualdad, este no ha sido el único. El DANE ha utilizado diferentes RR. AA. en distintas áreas temáticas. Por ejemplo, en el Índice de Precios de la Propiedad Residencial (IPPR), una operación estadística que genera índices de precios de vivienda para Bogotá y utiliza los RR. AA. de la Superintendencia de Notariado y Registro (SNR) y de la Unidad Administrativa Especial de Catastro Distrital (UAECD) para la producción estadística (DANE, 2021); también existen otros ejemplos que utilizan RR. AA., como la PILA o el SISBEN. Asimismo, los RR. AA. de la Planilla Integrada de Liquidación de Aportes (PILA) del Ministerio de Salud han sido utilizados en la generación de estadísticas interseccionales y granulares presentadas en el Registro Estadístico de Relaciones Laborales (RELAB).

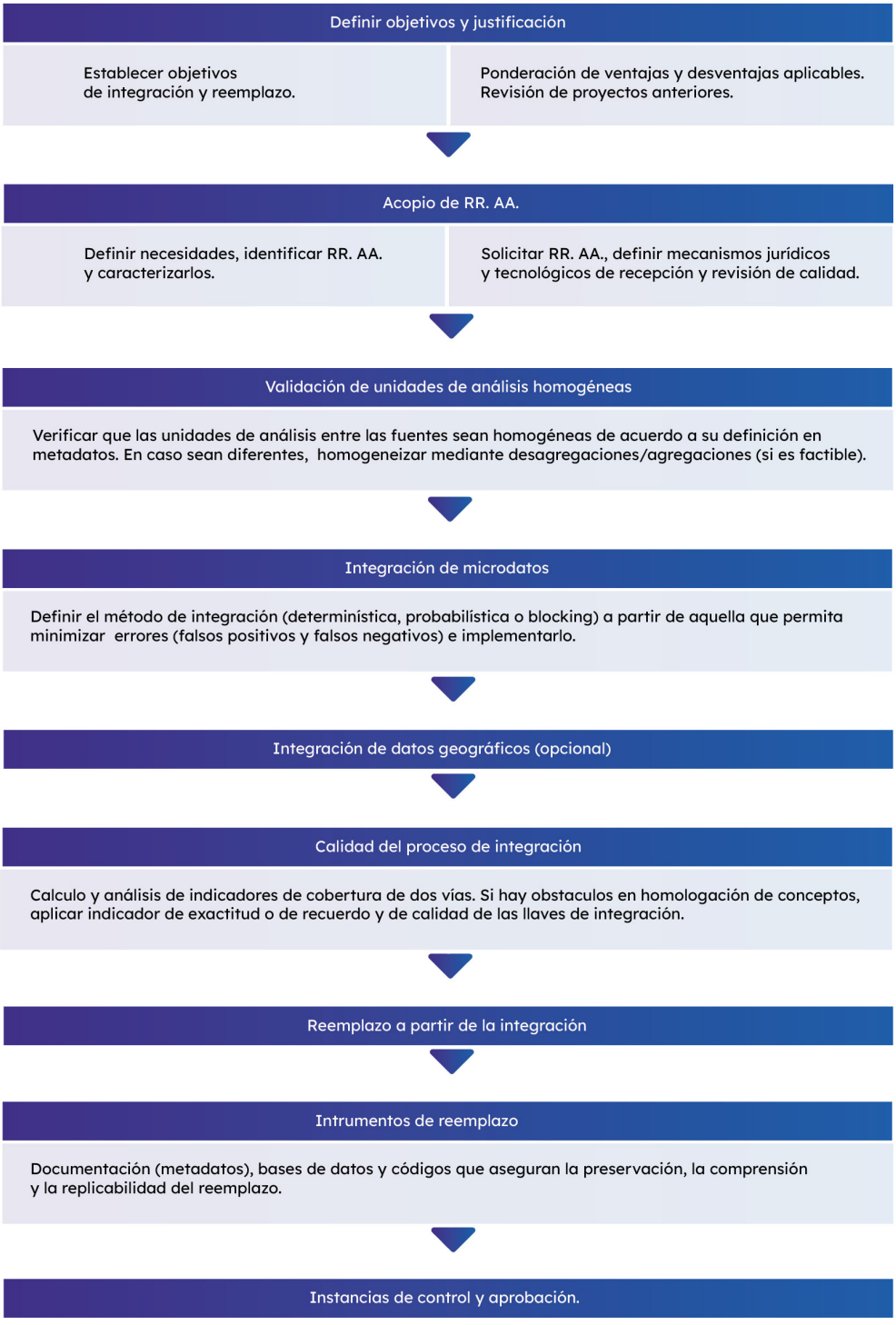
Respecto a la pobreza multidimensional, en el contexto de la coyuntura COVID-19, el indicador de inasistencia escolar sufrió cambios metodológicos en 2020. En la medida que la Encuesta Nacional de Calidad de Vida 2020 (ECV) no lograba capturar los efectos de la pandemia en su totalidad, se propuso emplear los RR. AA. de instituciones educativas y matriculados (Formulario C-600) y Sistema Integrado de Matrícula (SIMAT) del Ministerio de Educación (DANE, 2021). Adicionalmente, con el propósito de informar la discusión sobre los indicadores para la población de personas con discapacidad, el DANE realizó un proceso de integración del Registro para la Localización y Caracterización de Personas con Discapacidad (RLCPD) con la ECV. En la temática de educación, se realizó un proceso de triangulación y armonización del Censo Nacional de Población y Vivienda 2018 (CNPV) con el Directorio Estadístico de Sedes Educativas (DIREDU) y el SIMAT. Este ejercicio condujo a un análisis de accesibilidad a centros educativos utilizando la distancia entre hogares y centros educativos.

Desde el 2020, el DANE utilizó los RR. AA. de la PILA y ayudas institucionales (subsidios monetarios) como fuente de contraste e imputación. Lo anterior analizar la cobertura, que consistió en contrastar el valor observado en la encuesta con el valor de los RR. AA., y así realizar un análisis de sesgos de reporte (DANE, 2021). En la misma temática de pobreza monetaria y desigualdad, el DANE realizó un proceso de integración del Registro único de Víctimas (RUV) de la Unidad para la Atención y Reparación Integral a las Víctimas (UARIV) y la GEIH, esto con el fin de caracterizar a las personas que han sido víctimas del conflicto, describiendo esta población con la información presente en la GEIH tradicional (DANE, 2022).

2. Integración y reemplazo

En esta sección se presentan los pasos a seguir para realizar un proceso de integración de RR.AA. o integración y reemplazo de fuentes directas o primarias con registros administrativos, los cuales se resumen en el siguiente esquema.

Figura 1. Pasos para la integración



2.1. Definición de los objetivos y la justificación

Como primer paso para la integración de RR. AA. o la integración y el reemplazo de información de encuestas con RR. AA. se requiere la definición de los objetivos y la justificación correspondiente, que desarrolle los retos, las ventajas y las desventajas de estos. Esta sección presenta los elementos que son usuales encontrar en los objetivos y la justificación de la integración o la integración y el reemplazo. El reemplazo de información de la encuesta con información proveniente de registros administrativos o registros estadísticos puede generar diversas aplicaciones y beneficios para los usuarios; la UNECE (2011), clasifica estas aplicaciones según el grado en que se complementa las unidades de análisis de la población y de las variables. Existen lo siguientes enfoques que motivan realizar los reemplazos directos con registros administrativos:

- *Enfoque de partición de la población:* es un enfoque donde se utilizan todas las variables de una parte de las unidades de análisis, por lo que reduce el número de cuestionarios o entrevistas realizadas, lo que a su vez reduce el tiempo y el costo de la operación estadística.

- *Enfoque de partición de las variables:* este enfoque busca complementar algunas variables para la población completa con RR. AA. o fuentes secundarias. Reduce la longitud o la complejidad de los cuestionarios suministrados en las encuestas. En algunos casos, el enfoque de partición de las variables provee información precargada verificada o corregida durante la entrevista, lo que reduce el tiempo de respuesta y permite verificar la información proveniente de los registros para quienes verifican o corrijan la información precargada.

- *Enfoque mixto:* este enfoque corresponde a una mezcla de los anteriores, donde se suministra información en algunas variables para algunas unidades de la población.

El aprovechamiento de los RR. AA. debe definir un objetivo según los términos planteados. Los objetivos del reemplazo pueden ser: apoyar la imputación de elementos sin respuesta o con valores atípicos de un cuestionario o en cuestionarios sin ninguna respuesta; enriquecer la encuesta con nuevas variables u adicionar opciones de respuestas de las preguntas; apoyar la fase de análisis de los sesgos de respuesta y errores de medición, o construir modelos de áreas pequeñas (Meyer & Mittag, 2021; UNECE, 2011).

Los RR. AA. son proveídos principalmente por el sector público, la función principal de estos no es su uso estadístico, sino que cumple un rol administrativo en la operación y la ejecución de algún programa o política pública particular. Sin embargo, uno de los primeros usos de estos registros, por parte de los diferentes institutos de estadística, fue como un marco para el muestreo de sus encuestas (Zhang, 2021). Los RR. AA. contienen mucha información, por lo que la integración de estos con diferentes encuestas de hogares ha sido atractiva para las oficinas estadísticas del mundo. La decisión de integrar los RR. AA. o enriquecer encuestas con los RR. AA. debe contemplar las razones de su adopción, el objetivo principal y justificar de manera detallada las ventajas y las desventajas de su uso.

Tabla 1. Ventajas y desventajas de usar RR. AA.

Principales ventajas de utilizar RR. AA.	Principales desventajas de utilizar RR. AA.
1. Bajos costos relativos en la producción de información estadística.	**1.** No contar con RR. AA. de buena calidad y con cobertura suficiente.
2. Mayores niveles de desagregación geográfica y mayor frecuencia en el reporte de los datos.	**2.** Baja frecuencia de actualización del registro.
3. Reducción de la carga de realizar una encuesta, tanto para los encuestados como para los encuestadores.	**3.** Poca capacidad técnica y tecnológica para procesar e integrar los datos.
4. Según la aplicación, permiten mejorar la cobertura y mejorar los niveles de desegregación.	**4.** Restricciones legales para el uso de la información.

Fuente: Elaborada a partir de Estadísticas basadas en registros: Aprovechamiento estadístico de datos administrativos, Wallgren & Wallgren, 2016, p. 14.

La recolección tradicional de los datos para las encuestas de hogares, por medio de una llamada telefónica o visitas presenciales, puede generar sesgos en la información recogida (DANE, 2021), en la medida que las personas encuestadas podrían experimentar una "pérdida" de privacidad de la información a la hora de responder preguntas confidenciales o sensibles como lo son los ingresos, pertenencia a programas sociales, nivel educativo, entre otros (Wallgren y Wallgren, 2016).

El reemplazo de las encuestas con los RR. AA. o con integraciones de RR. AA. permiten incluir en la generación de indicadores de la encuesta información de variables que no fueron capturadas en la encuesta, valores faltantes de los cuestionarios o el tratamiento de valores extremos. En la recolección tradicional de encuestas, algunos encuestados suelen negarse a responder algunas preguntas o incluso secciones completas, particularmente sobre preguntas sensibles. Para solucionar este problema, algunas oficinas estadísticas usan métodos de imputación, como el "método hot-deck" (Little & Rubin, 2002) que utiliza solo información de la misma fuente, mientras que algunos investigadores aplicados suelen abordar estos problemas usando solo cuestionarios completos. No obstante, este tipo de estrategias asumen que los valores ausentes por no respuesta se distribuyen de manera aleatoria y no están asociados con las características propias de las personas. Es por esto por lo que, la integración de los RR. AA. provee una medida mucho más acertada de la población objetivo (Meyer y Mittag, 2021).

En este orden de ideas, el uso de los RR. AA. generará un impacto positivo en la calidad de las estimaciones en la medida que logran rescatar información verídica e integrarla con la información que se tienen en las encuestas (Wallgren y Wallgren,

2021). A su vez, esta integración ayudará a desarrollar evaluaciones de estos sesgos de medición mediante la comparación de los microdatos observados en las encuestas y los reportados en los RR. AA. (UNECE, 2011; CEPAL, 2020).

De acuerdo con Meyer y Mittag (2021), la integración de RR. AA. puede mejorar, particularmente, las mediciones de pobreza y desigualdad. Los autores explican que los programas de transferencias monetarias son una de las herramientas principales que tienen los gobiernos para combatir la pobreza y la desigualdad. Bajo esta lógica, si estos programas están bien focalizados, la información proveniente de los registros puede ayudar a identificar a las personas con mayor probabilidad de estar en condición de pobreza o vulnerabilidad. Asimismo, contar con la información oficial de los diferentes programas sociales, permite reducir los errores de medición propios de las variables más sensibles como lo son el ingreso y la pertenencia a programas sociales (Wallgren y Wallgren, 2016).

Por otro lado, para aprovechar plenamente la integración de RR. AA. es fundamental que estos sean de alta calidad y cuenten con cobertura poblacional y geográfica necesaria. Por lo tanto, es necesario implementar un proceso de validación de la calidad del registro para determinar su idoneidad para la integración con el propósito de reemplazo, lo que implica evaluar la cobertura, la completitud de variables, las inconsistencias, la calidad de las claves de integración y otras dimensiones relevantes. Además, si se tiene la intención de adoptar esta nueva estrategia de producción estadística de manera constante, es fundamental establecer mecanismos que aseguren la puntualidad, el acceso y la calidad por parte del administrador de los RR. AA. necesarios para la producción estadística (este tema se trata en profundidad en la **Sección 2.2** – Acopio de registros administrativos).

Pueden existir diversos factores que limitan el uso de los RR. AA. (Künn, 2015; Wallgren y Wallgren, 2021):

- Los RR. AA. pueden presentar una baja cobertura poblacional. Wallgren y Wallgren (2021) argumentan que muchos países no logran tener una cobertura lo suficientemente robusta, principalmente en los países en desarrollo que cuentan con una gran parte de su población en el sector informal o viviendo en zonas rurales. Asimismo, los autores argumentan que algunos registros pueden tener una frecuencia de actualización relativamente baja, por lo tanto, muchos de los reportes pueden ser incorrectos (por ejemplo: la dirección de residencia).

- Los RR. AA. pueden no estar estandarizados debido a que no tienen un fin estadístico como tal, sino administrativo.

- Los institutos de estadística oficiales pueden tener poca capacidad técnica y tecnológica para procesar e integrar los datos. Las bases de datos de los RR. AA. suelen contener la información bruta, por lo que se requiere una gran capacidad técnica y tecnológica para su procesamiento.

- Pueden existir restricciones legales para el uso de la información, en la medida que los datos contenidos suelen ser de carácter confidencial.

Antes de realizar la integración de los RR. AA. con las encuestas, los equipos de trabajo deben identificar si las ventajas y las desventajas expuestas anteriormente se ajustan al análisis que quieren realizar, y ponderar si el proceso de integración es costo-efectivo en su respectivo proyecto. El aprovechamiento de los RR. AA. dependerá del objetivo que se tenga. Por ejemplo, una posible razón para utilizar un registro administrativo es identificar una población con alguna característica particular que la encuesta no logre capturar. Otra razón podría ser incluir una nueva variable que provenga del registro y que la encuesta no haya capturado, o que la encuesta haya capturado de manera deficiente por la materialización de contingencias que afectaron el proceso de recolección.

Los anteriores ejemplos clasifican en un mismo método de imputación llamado "imputación cold-deck". Este método de imputación consta de un emparejamiento de las encuestas con los diferentes RR. AA., donde se actualiza la información de la pertenencia de una persona a un programa en particular o con una característica en particular en función del cruce directo. Usualmente este método se usa cuando no hay respuesta en una o más variables de la encuesta, falsos valores en cero, valores atípicos, adicionar variables nuevas que son difíciles de medir en las encuestas de hogares, o cuando se busca una mayor desagregación de una pregunta en particular (Meyer & Mittag, 2021; UNECE, 2011).

Igualmente, en la justificación del reemplazo de las variables de la encuesta por variables del RR. AA., se deben identificar e incluir los antecedentes de ejercicios o procesos de integración similares dentro de la entidad como a nivel nacional, del área temática pertinente. Esto con el objetivo de asegurar que se reutilicen los resultados y las conclusiones de proyectos anteriores y evitar la duplicación de esfuerzos. La justificación del reemplazo puede venir de la etapa de evaluación de la operación estadística. Es decir, después de ejecutar un ciclo de la operación, en la fase de evaluación se diagnostica un problema que lleva a la necesidad de hacer un reemplazo.

Bajo esta lógica, es pertinente llevar a cabo una integración con RR. AA. siempre y cuando se detecten posibles mejoras de la información de la fuente primaria a partir de los mismos. No obstante, esta conclusión depende de los objetivos específicos de cada grupo de trabajo, así como las condiciones particulares de los RR. AA. y las encuestas. Es importante que el grupo de trabajo cuente con los elementos necesarios para justificar la integración respectiva siguiendo los principios de oportunidad, comparabilidad, coherencia, eficiencia y seguridad y privacidad de la información.

2.2. Acopio de registros administrativos

Para acopiar las fuentes secundarias requeridas en la producción de estadísticas, es necesario diseñar e implementar esquemas de relacionamiento que faciliten los tramites jurídicos y contengan incentivos temáticos y técnicos (como la visualización de datos o las revisiones de calidad respectivamente), para que el productor del registro administrativo remita la información solicitada en los tiempos acordados y con niveles de calidad y cobertura óptimos. Adicionalmente, el acopio debe estar sustentado en el cumplimento de la normatividad vigente,

y cuando se considere necesario, en acuerdos vinculantes entre las partes[6].

Los proveedores de estas fuentes de información comúnmente se benefician del aprovechamiento en la medida que son usuarios de la información publicada por el DANE y reciben retroalimentación sobre la calidad de los datos y los metadatos utilizados para propósitos misionales. Con la maduración en el aprovechamiento de los RR. AA., se configuran los registros estadísticos que se convertirán en una fuente común para la producción estadística, complementado así el rol de productor del DANE.

Cabe reiterar, que el aprovechamiento estadístico de los RR. AA. requiere un sistema de registros integrados que permita un uso eficiente de los datos, la cooperación entre las entidades para difundir información sobre el funcionamiento del RR. AA., la coherencia entre los registros estadísticos; y el envío regular de los datos (Rothbard, 2015; Wallgren y Wallgren, 2021).

En el DANE las tareas de relacionamiento y análisis de la calidad de los RR. AA. acopiados (bases de datos y metadatos) son llevadas a cabo en dentro del proceso de gestión de proveedores, mientras que la recepción, el almacenamiento y la distribución al interior del DANE de los RR. AA. requeridos para el aprovechamiento estadístico está a cargo del GIT de Gestión de Datos, todo lo anterior se enmarca en la Ley SEN (Ley 2335 de 2023).

Como primer paso, el equipo temático debe identificar sus necesidades de información en términos de variables, unidades de observación, identificadores, periodicidad, grado de anonimización y demás aspectos relevantes para cumplir con el objetivo. Posteriormente, debe identificar las fuentes secundarias que atienden su necesidad, para lo cual puede utilizar inventarios de registros administrativos del sistema estadístico nacional (en DANE se cuenta con el Sistema de Identificación y Caracterización de Oferta y Demanda Estadística (SICODE)), consultar expertos temáticos y/o rastrear los antecedentes del aprovechamiento de los RR. AA. de interés al interior de la entidad (buscando reutilizar los resultados de proyectos anteriores). El rastreo de antecedentes puede también contener información sobre relacionamientos vigentes, las frecuencias actuales y deseables de recepción del RR.AA., las versiones disponibles, los metadatos y las reglas de actualización del RR. AA.

Cabe recordar que los metadatos corresponden a la Información necesaria para el uso y la interpretación de los datos, entre los cuales encontramos los diccionarios de datos, la normatividad asociada con la fuente secundaria, los manuales de diligenciamiento, los manuales de proceso, las modificaciones a la producción de la fuente secundaria, los cambios en las reglas de la fuente secundaria que puedan impactar la comparabilidad del aprovechamiento estadístico, etc.

A partir del listado de potenciales fuentes secundarias a aprovechar, se procede a caracterizarlas mediante la identificación de sus principales atributos como, objetivo, sustento normativo, periodicidad, variables, mecanismos de recolección y procesamiento, entre otros. En el DANE, el GIT de Planificación de DIRPEN, cuenta con un formulario para la caracterización

6 Parte del marco jurídico se encuentra en la Ley SEN o Ley 2335 de 2023

de los registros administrativos que se incluyen en SICODE, el cual es posible aprovechar para esta actividad. De forma paralela, se procede a identificar las principales características del productor de la fuente secundaria, con énfasis en posibles datos de contacto para el relacionamiento. Es importante mencionar que el GIT de Gestión de Proveedores de Datos cuenta con un directorio de fuentes que puede facilitar este proceso.

Como siguiente paso, se contrasta la caracterización de las fuentes secundarias con el objetivo de la integración y el reemplazo, con el propósito de seleccionar las fuentes que serán aprovechadas estadísticamente. Una vez identificadas las fuentes secundarias requeridas, confirmando previamente que esta información no se encuentra ya en la entidad como registro estadístico o fuente aprovechada por otra dependencia, se inicia el proceso de relacionamiento.

Uno de los retos más comunes en el aprovechamiento de fuentes secundarias está asociado al establecimiento de mecanismos de cooperación efectivos con las entidades o empresas que producen dichos registros, que son fuente de la información necesaria para la mejora de la producción estadística (Rothbard, 2015; Wallgren y Wallgren, 2021).

Si alguna entidad del SEN (Ley 2335 de 2023) quiere aprovechar un RR.AA. que produce otra entidad, el primer paso del relacionamiento consiste en solicitar de manera oficial la información requerida, dando a conocer como mínimo: la normatividad que sustenta la solicitud, el uso que se dará a la información, las necesidades en términos de periodos, variables, desagregaciones y grado de anonimización, así como las condiciones de reserva y seguridad aplicables.

Continúa con la definición entre las dos partes de un mecanismo jurídico para el relacionamiento, mediante el cual se hacen explícitos los compromisos respecto a las condiciones de uso y protección de la información por parte del receptor y de la periodicidad, calidad y notificación de actualizaciones por parte del dueño de la fuente secundaria. En este sentido, en DANE comúnmente se utilizan acuerdos o convenios para establecer él envió de información, cuya aplicabilidad depende del concepto de la oficina jurídica (acuerdos) u oficina de compras públicas (convenios). Cabe anotar que, teniendo en cuenta las facultades con que cuenta el DANE respecto a la solicitud de información para la producción de estadísticas oficiales, también es viable obtener las fuentes secundarias que se requieren sin la firma de ningún acuerdo o convenio; esto ocurre comúnmente con envíos puntuales de información.

De forma paralela, se acuerda entre las oficinas de Tecnologías de la Información (TI) de las dos entidades, los mecanismos para el envío y la recepción de la información, especificando la infraestructura tecnológica y el software a emplear, así como los puntos de contacto de cada entidad responsables del procedimiento.

Una vez acordados los términos del relacionamiento, se realiza la solicitud de recepción de la fuente secundaria a la Oficina de Sistemas (GIT Gestión de Datos), con la finalidad de cumplir con las medidas de seguridad y privacidad de la información. La solicitud de recepción puede contener periodos de referencia,

periodos de actualización o versión, la especificación de un separador de columnas que conduzca a la legibilidad de los datos, la trazabilidad de modificaciones o alteraciones de los datos crudos o cualquier metadato disponible.

Previo al aprovechamiento estadístico, es recomendable realizar un diagnóstico de la calidad de la base de datos y verificar que se cuenta con metadatos suficientes y actualizados para el adecuado aprovechamiento de la información. Para este análisis, el GIT de Gestión de Proveedores aplica y pone a disposición del DANE la herramienta "Revisión de calidad de datos"[7], la cual permite determinar la calidad de una fuente secundaria a partir de la revisión de 15 dimensiones o aspectos de calidad, agrupadas a su vez en tres hiperdimensiones. Esta herramienta está compuesta por indicadores/lista de chequeo de tipo cualitativo y cuantitativo, la cual se resume en la siguiente tabla.

Tabla 2. Hiperdimensiones de calidad

Hiperdimensiones de calidad	Dimensión	Descripción de la dimensión
Fuente (cualitativo): aspectos de calidad relacionados con la entidad Responsable de la fuente secundaria (proveedor) y de los procesos de acceso/ transferencia	**Proveedor**	Análisis de la confianza y la fiabilidad en el titular de los datos, la efectividad del contacto y la comunicación con el proveedor (periódica / esporádica).
	Relevancia	Análisis del propósito estadístico de los datos: utilidad, uso previsto, demanda de información y carga a la fuente.
	Privacidad y seguridad	Análisis de provisiones legales, confidencialidad y seguridad en la trasmisión de datos.
	Entrega	Análisis de costos asociados a la transferencia, los acuerdos existentes, la puntualidad, el formato de envió y el cumplimiento con selección de datos solicitada.
	Procedimientos	Análisis del modo de transmisión de datos, nivel de descifrado de datos y otros elementos de seguridad dispuestos por el proveedor al momento del acceso/transferencia de datos.

7 Esta herramienta se desarrolla por el GIT GDP de la DRA.

Hiperdimensiones de calidad	Dimensión	Descripción de la dimensión
Metadatos (cualitativo): aspectos de los documentos explicativos de la fuente secundaria. Los indicadores ayudan a comprender la configuración de la fuente secundaria y a determinar si existen verificaciones y / o modificaciones en los datos.	**Claridad**	Análisis de la confianza y la fiabilidad en el titular de los datos: identificación de variables por tipología e identificación de cambios.
	Comparabilidad	Análisis del marco conceptual presente en la fuente secundaria frente a las necesidades del aprovechamiento.
	Llaves únicas	Análisis de la descripción de las llaves únicas presentes en la fuente secundaria.
	Tratamiento de datos	Análisis para determinar si el proveedor de la fuente secundaria realiza comprobaciones y/o modifica los datos de la fuente, agrupados en chequeos del proveedor y modificaciones.

Hiperdimensiones de calidad	Dimensión	Descripción de la dimensión
Datos (cuantitativo): aspectos de calidad de los datos (hechos) en la fuente secundaria. Los indicadores están predominantemente relacionados con la precisión y verifican la legibilidad del archivo de datos y la conformidad del conjunto de datos con la definición de metadatos.	**Chequeos técnicos**	Indicadores de extracción, legibilidad, convertibilidad, cumplimiento de los metadatos, la declaración de cumplimiento y la carga.
	Exactitud	Indicadores de autenticidad, inconsistencias y valores atípicos.
	Completitud	Indicadores de subcobertura poblacional, sobrecobertura poblacional y selectividad.
	Temporalidad	Indicadores de oportunidad de uso y retraso del dato
	Integralidad	Indicadores asociados a la calidad de las variables de integración y comparabilidad de objetos / unidades.
	Declaración de cumplimiento	Análisis de calidad por variable: tipo de variable, cumplimiento de respuesta obligatoria, cumplimiento de dominio y rango de respuesta, cumplimiento de longitud y de las reglas de validación del productor.

Fuente: Elaborada a partir de Dimensions, quality indicators, and methods for Source, Daas, Ossen, Vis-Visschers, & Arends-Toth, 2009, p.7.

Cabe mencionar que los RR. AA. de uso frecuente en el DANE cuentan con diagnósticos de calidad más completos, los cuales se pueden realizar a partir del aprovechamiento o del cumplimiento de las características técnicas. Tomando como referencia los resultados obtenidos en el diagnóstico, el área temática decide si es posible aprovechar o no la fuente.

A su vez, la fuente tiene la posibilidad de definir un plan de fortalecimiento a partir de los resultados de la revisión de calidad, con el cual pueda mejorar la calidad de la información, aprovechando las metodologías y el acompañamiento que ofrece el DANE. El plan de fortalecimiento traduce las oportunidades de mejora identificadas en acciones concretas que permitirán mejorar la calidad de la fuente de información, a ser ejecutadas en un periodo de un año como máximo (plazo sugerido). El seguimiento al cumplimiento de este plan se realiza mediante el envío al DANE de las evidencias de la implementación de las acciones de mejora.

Por otra parte, en cuanto a la recepción de datos de encuestas producidas por el DANE, es necesario verificar que se cuenta con acceso a las variables y el diccionario de la encuesta, de acuerdo con las necesidades del aprovechamiento. Esto para procurar la mayor cantidad de información posible que optimice la identificación de las unidades de análisis. El intercambio de información de encuestas es un proceso regular dentro de la entidad y cuenta con lineamientos que procuran la seguridad y la privacidad de la información.

Cabe anotar que los procedimientos de relacionamiento y diagnóstico de la calidad los lidera el GIT de Gestión de Proveedores al interior del DANE, no obstante, las áreas productoras de información también pueden implementarlos en coordinación con este GIT.

2.3. Validación de unidades de análisis homogéneas

Los RR. AA. suministran información que complementa las encuestas y los indicadores basados en encuestas al aportar variables adicionales. Los RR. AA. pueden arrojar información sobre las mismas unidades presentes en las encuestas y contener campos similares o variables homologables que sirvan para comparar las fuentes. Esta información en común entre las fuentes constituye el puente que habilita su integración y fortalece así los indicadores generando economías de escala con los recursos iniciales. En ese sentido, la coherencia estadística entre las unidades de análisis de las fuentes de información necesita definiciones y unidades de medida comunes.

Los RR. AA. pueden contener información de muchos tipos de individuos, ya sean personas, hogares o empresas. Las unidades de análisis son personas o entidades caracterizadas con los valores que toman las variables de la encuesta. Esta no es una relación rígida sino flexible porque las unidades de análisis también se definen gracias a los valores que toman en algunas variables esenciales, por ejemplo, las siguientes pueden ser unidades de análisis: trabajadores, trabajadores formales, trabajadores con cotización a seguridad social, trabajadores asalariados en una ciudad determinada, relaciones laborales del sector de la construcción, etc.

Para realizar el proceso de integración y de homogenización, se debe implementar herramientas enfocadas a la calidad de

base de datos, como la revisión de chequeos técnicos, comparabilidad, integrabilidad y exactitud, conceptos que se encuentran y desarrollan en la guía "Metodológica de Diagnóstico de Registros Administrativos" elaborada por el DANE. La validación de las unidades de análisis es necesaria para la construcción de indicadores a partir de integrar información de RR. AA. y fuentes primarias, lo que implica que este proceso es independiente de si posteriormente se realizan remplazos en la información de cada una de las fuentes.

Tabla 3. Revisiones de calidad

Dimensión	Definición
Revisiones técnicas	Usabilidad técnica del archivo y los datos en el archivo, según diccionario de base de datos. Permite la verificación de la accesibilidad y la coherencia del registro.
Comparabilidad	Es la característica que permite que los datos contenidos en los registros administrativos puedan relacionarse, agregarse e interpretarse entre sí o con respecto a algún parámetro común.
Integrabilidad	Grado en el que la fuente de datos puede integrarse.
Exactitud	Grado en que los datos contenidos en los registros administrativos se aproximan y describen correctamente las cantidades o características de los elementos registrados.
Completitud	Grado en el cual la fuente de datos incluye datos que describen un grupo de objetos y variables del mundo real.
Temporalidad	Indicadores relacionados con el tiempo o estabilidad. Para el ejercicio de diagnóstico esta dimensión no se evalúa debido al objetivo de diagnosticar un registro administrativo.

Fuente: Elaboración a partir Buenas practicas para la configuración de registros administrativos para su aprovechamiento estadístico en el SEN, DANE, 2024, p 142 - 143.

Las mallas de validación del registro administrativo muestran las combinaciones de valores para un conjunto de variables que están marcadas como inconsistentes. Estos procesos son llevados a cabo por el proveedor de RR. AA., sin embargo, los datos crudos del RR. AA. no reflejan todas las validaciones y las transformaciones de calidad aplicadas por el proveedor. En estos casos, el equipo de trabajo debe desarrollar las reglas o las transformaciones adecuadas para el objetivo planteado. En el caso de persistir ambigüedad en el desarrollo y la implementación de las transformaciones, el equipo puede entrar en contacto con el proveedor o con otros grupos de trabajo que tengan experiencia con el registro administrativo. Otra opción consiste en generar escenarios para evaluar la robustez de los resultados ante los cambios en las transformaciones.

La información entre las fuentes puede tener unidades de análisis distintas, y, por tanto, se debe realizar un proceso de homogenización de las unidades de análisis. Por ejemplo, en Colombia, el registro administrativo del programa de transferencias monetarias Más Familias en Acción, custodiado por el Departamento para la Prosperidad Social (DPS), contiene información de las personas titulares y beneficiarias del subsidio entregado por el gobierno nacional. A pesar de que el registro administrativo reporta a todas las personas beneficiarias, los valores monetarios asignados a cada persona no corresponden a los montos efectivos girados a cada persona, sino que es el monto total asignado al hogar durante un ciclo de pagos. Lo anterior es un caso de una base de datos por persona-transacción, pero que en la práctica se debe trabajar como una base de datos hogar-transacción.

Esto también sucede en los RR. AA. de programas como Devolución del IVA, Ingreso Solidario (custodiados por DPS) y transferencias monetarias de entidades territoriales de las alcaldías de las principales ciudades. Por otro lado, hay RR. AA. de programas de transferencias monetarias como Jóvenes en Acción y Colombia Mayor los cuales son registros que contienen información de cada uno de los titulares y los montos asignados corresponden únicamente a los titulares del programa. La misma lógica se aplica en el RR. AA. pensionados (PILA – Ministerio de Salud), el cual reporta únicamente las personas que tienen asignada una pensión.

Lo anterior, presenta un reto importante a la hora de decidir la unidad de análisis más precisa. No obstante, la elección de dicha unidad de análisis no depende solamente del registro administrativo empleado, sino también del propósito de la integración y de las fuentes de datos a utilizar. Para la estimación de la tasa de incidencia de pobreza monetaria y pobreza extrema se calculan los ingresos totales de una unidad de gasto, para después calcular un ingreso per cápita por unidad de gasto. Lo anterior supone una agregación de todos los ingresos de la unidad de gasto sin importar la procedencia ni el aportante. En este caso se muestra como el tratamiento de la unidad de gasto se deriva de la demanda de información estadística y de los conceptos del área temática respectiva y, por consiguiente, la utilización de los RR. AA. necesita ceñirse a esta definición.

Desde la guía Lineamientos para documentar la metodología de operaciones estadísticas, censos y encuestas por muestreo, publicada por el DANE, se define la unidad de análisis para las operaciones estadísticas, concepto que se puede extrapolar a los RR. AA. (p 20):

"La unidad de análisis corresponde a la entidad objeto de estudio en una medición y sobre la que se presentan las conclusiones de la investigación. La operación estadística puede tener varias unidades de análisis. Por ejemplo, en las encuestas agropecuarias pueden ser el predio, la finca y la unidad productora agropecuaria; en las encuestas sociales las personas, los hogares y las viviendas".

Para el caso del RR. AA. de Más Familias en Acción la unidad de análisis es el hogar, pero la base que se cruza con la encuesta es a nivel de persona. Por lo que los indicadores generados deben considerar que la información de transferencia del registro administrativo y de la encuesta hagan referencia a una unidad de análisis similar.

La homogeneidad de las unidades de análisis de las fuentes de información es un proceso que se retroalimenta con la integración de los microdatos. Los conceptos del área temática correspondiente pueden señalar una cercanía teórica entre las fuentes, sin embargo, esto necesita ser contrastado con estadísticas descriptivas generadas a partir de la integración de los microdatos. Estos resultados pueden llevar a su vez a ajustar los conceptos utilizados para definir la unidad de análisis. La Sección 2.4 – Integración de los microdatos de las distintas fuentes muestra los indicadores calculados para las estadísticas descriptivas requeridas para el anterior análisis.

Desde el punto de vista de programación, las unidades de análisis sirven para organizar mejor las bases de datos al garantizar que una unidad de análisis ocupa a lo sumo una fila dentro de los datos. Un tratamiento usual en el alistamiento del registro es ordenar la base de datos de tal forma que cada observación represente un único individuo. Por ejemplo, los RR. AA. de ayudas institucionales suelen presentarse por titular-transacción, lo que hace que haya más de una única observación por persona. Asimismo, se deben identificar aquellas observaciones que son duplicados exactos y eliminarlos, de tal forma que la base de datos cuente con una entrada única de persona-transacción. En el caso de los RR. AA. de ayudas institucionales, las entradas únicas persona-transacción se entienden como una base de datos con un identificador de persona repetido por una variable tiempo que corresponde a la fecha de la transacción. Para llevar la variable de tiempo a las columnas solo es necesario trasponer los datos utilizando variables que identifiquen la unidad de análisis y la fecha de la transacción.

2.4. Integración de los microdatos de las distintas fuentes

La integración de múltiples fuentes de datos satisface necesidades de información que no se podrían atender con las fuentes tradicionales e individuales. Optimizar la integración de las fuentes aprovechando la información disponible de las llaves de identificación permite tener estimaciones más certeras y robustas y aumenta el valor público de los datos con economías a escala.

La integración entre RR. AA. o el cruce entre los RR. AA. y las encuestas se puede realizar con llaves de integración directas como números y tipos de documento o indirectas como nombres, apellidos, fecha de nacimiento y lugar de nacimiento, el primer tipo de llaves (número y tipo de documento) se denominan llaves primarias,

donde también podemos encontrar el NIT de las organizaciones, sean de naturaleza pública o privada.

Como se mencionó anteriormente, los países con un fuerte componente de RR. AA. tienen una muy buena calidad de números y tipos de identificación en todas sus bases: censos, encuestas muestrales y RR. AA. No obstante, en los sistemas tradicionales, basados en encuestas muestrales representativas, este tipo de variables de identificación no suelen ser indispensables, y, por lo tanto, la información que se tiene es deficiente al principio. En particular, uno de los retos iniciales, para el caso colombiano, es el cambio de los distintos tipos de documentos de identificación que puede tener una misma persona o unidad económica a lo largo del tiempo. Recordemos que en el caso de tipos de documentos para personas (antes del NUIP) una persona pasaba por registro civil, tarjeta de identidad y cédula de ciudadanía.

En este contexto se han desarrollado diferentes metodologías de cruce para aprovechar cada una de estas llaves de integración. La integración determinística (*deterministic record linkage*) se usa con llaves directas, mientras la integración probabilística (*probabalistic record linkage*) se usa con llaves indirectas.

Integración determinística

Se define como aquella integración en la que se aprueba o rechaza un cruce si sus llaves de integración son iguales o no, es decir, manejan una lógica booleana que me descarta o aprueba un cruce si ambos valores del registro son diferentes o iguales, respectivamente. En esta lógica si los registros tienen la misma llave de integración entonces es una integración positiva, y si son diferentes es una integración negativa. Sus ventajas se ven representadas en la rapidez del procedimiento y el bajo costo computacional, además de que, si sus llaves de integración son de buena calidad, la tasa de falsos positivos y negativos[8] suele ser baja. Por ejemplo, una llave de identificación directa es la combinación de tipo y número de documento, luego al hacer la integración determinística únicamente tendremos en cuenta aquellos registros que, en ambas bases, tienen mismo tipo y número de documento.

Aquí se puede diferenciar una metodología como es la integración determinística 'stepwise' (*stepwise record linkage*), que consiste en definir n variables, usualmente llaves indirectas, como nuestras llaves de integración, pero no preocuparnos porque las n crucen, sino que sean m (m<=n) cualesquiera de estas las que sí lo hagan (donde m son el número de condiciones necesarias para cruzar una observación). Es decir, si por ejemplo se definen las llaves como dirección, municipio, sexo, edad y ocupación, se observa que n=5 y se puede tomar m=3, entonces mientras que algún par de registros en ambas bases tengan 3 de estas 5 variables idénticas (sexo, edad y ocupación, entre otras) entonces se validará como un cruce exitoso.

8 Un falso positivo se genera cuando se logran integrar dos observaciones que no corresponden a la misma persona o unidad administrativa. Un falso negativo se genera cuando una misma persona o unidad administrativa se encuentra en ambas bases de datos, pero esta no cruza, por lo general por problemas en la llave de integración.

Tabla 4. Ejemplo base 1 para cruce

ID_1	Dirección	Municipio	Sexo	Edad	Ocupación
1	Calle 3A 4 32	Chicoral	M	46	Administrador
2	Carrera 12 18 01	Pacho	F	31	-
3	Diagonal 157 sur 35 78	Bogotá	F	22	Panadero
4	Al frente de la tienda de Doña Rita	Vereda Dulceyes, Jenesano	M	63	Campesino
5	Calle 32 26 13	Santa Marta	F	50	Ingeniera

Fuente: elaboración propia.

Tabla 5. Ejemplo base 2 para cruce

ID_1	Dirección	Municipio	Sexo	Edad	Ocupación
1	Calle 3A 4 32	Chicoral	M	44	-
2	Tienda Doña Rita	Vereda Dulceyes, Jenesano	-	63	Campesino
3	Calle 12 18 01	Pasto	F	31	Músico
4	Carrera 26 32 13	Cartagena	M	19	Ingeniera
5	Diagonal 157 sur 35 78	Bogotá D.C.	F	22	Pastelero

Fuente: elaboración propia.

Se evidencia que al hacer la integración determinística 'stepwise' con n=5 y m=3 de las anteriores tablas se obtienen los siguientes resultados:

- ID_1 = 1 e ID_2 = 1 tienen las variables Dirección, Municipio y Sexo idénticas, luego es una **integración positiva**

- ID_1 = 2 e ID_2 = 3 tienen las variables Sexo y Edad idénticas, luego es una **integración negativa**

- ID_1 = 3 e ID_2 = 5 tienen las variables Dirección, Municipio, Sexo y Edad idénticas, luego es una **integración positiva**

- ID_1 = 4 e ID_2 = 2 tienen las variables Municipio, Edad y Ocupación idénticas, luego es una **integración positiva**

- ID_1 = 5 e ID_2 = 4 tienen la variable Ocupación idéntica, luego es una **integración negativa**

Considerando las limitaciones y la información sobre la calidad de las llaves de integración, la integración determinística es la metodología más común utilizada en el DANE. Por ejemplo, con la metodología utilizada en DANE para cruce de personas, se sigue un protocolo de integración de dos etapas, útil para grupos de edad con cambios en el tipo de documento, pero no en el número de identificación (DANE, 2021). Esta consta de una primera vuelta con un cruce simple de las llaves de integración tipo y número de identificación. Luego las observaciones que no cruzan se usan en la segunda vuelta, donde se realiza un cruce solo por el número de documento, pero se valida su pertinencia utilizando el código fonético del primer nombre y del primer apellido o un criterio auxiliar que aproveche información de identificación distinta a los tipos y números de documento.

Solo las observaciones que coincidan son tenidas en cuenta, el resto de las observaciones se descartan en el proceso de integración con el registro. Es importante señalar que, si bien las encuestas en Colombia no preguntan de forma obligatoria los campos del tipo y número de documento de identidad, la completitud y la consistencia evaluadas con el Registro Estadístico Base de Población (REBP) de estos campos suele permitir vincular las encuestas con diferentes conjuntos de datos. El REBP funge como base de referencia, es decir, es un registro maestro de las variables de identificación de las personas (tipo y número de documento), así como el Directorio Estadístico de Empresas lo es para las unidades económicas.

Integración probabilística

Esta se diferencia de la determinista gracias al uso de la lógica difusa y no de la booleana, es decir, no define una integración como positiva y negativa de entrada, sino que da un valor entre 0 y 1 que define la calidad de la integración. Luego, como parte de sus retos, le queda definir al usuario ese umbral de aceptación y rechazo del cruce. No suele haber un estándar universal sobre la definición de este umbral, pues depende de la naturaleza de las variables y del método usado.

Esta metodología tiene sus ventajas al permitir trabajar con variables que son pseudo-identificadores, las cuales suelen ser varias y pueden llegar a ser formadas por diferentes variables. Pero también implica un costo computacional grande.

Las herramientas más usadas para este tipo de cruces suelen ser las distintas métricas de comparación de cadenas de caracteres, entre ellas están Levenshtein, Jaro, Jaro-Winkler, Damerau-Levenshtein, Hamming, etc. Un procedimiento necesario antes de usar estas herramientas es la de estandarizar las cadenas de caracteres a un formato que permita evitar errores ocasionados por falsas diferencias. Con esto nos referimos es a estandarizar, por ejemplo, "A", "a", "á" y "ä", pues estos 4 caracteres son todos diferentes entre sí para estas métricas, pero para nosotros son similares. Para evitar este inconveniente se recomienda quitar todos los acentos y estandarizar a mayúsculas todos los caracteres. A continuación, se explica de forma sintética cada uno de los métodos probabilísticos mencionados.

- *Distancia de Levenshtein:* es una métrica basada en el conteo de inserciones, eliminaciones y sustituciones de caracteres para pasar de una cadena a otra. Cuando las cadenas son iguales su distancia es cero (0), en cambio, si hay cuatro pasos para llegar de una cadena a la otra, entonces decimos que su distancia es de cuatro (4). Esta distancia busca encontrar el camino más rápido para pasar de una cadena a otra, limitándose en las reglas que establece. A continuación, se muestran algunos ejemplos al momento de usar está distancia:

  - *Levenshtein* ("MONTAR", "MATAR"): 2. Ya que el proceso que se usa es eliminar la "N" de la palabra "MONTAR", y sustituir la letra "O" por la "A" en esta misma palabra.

  - *Levenshtein* ("MILENA", "MENILA"): 4. Se reemplaza la "I" por "E" en "MILENA", se reemplaza la "L" por "N" en "MILENA", se reemplaza la "E" por "I" en "MILENA" y se reemplaza la "N" por "L" en "MILENA".

  - *Levenshtein* ("REMERA", "CAMISETA"): 5. Se reemplaza la "R" por "C" en "REMERA", se reemplaza la primera "E" por "A" en "REMERA", se inserta la "I" entre la "M" y "E" en "REMERA", se inserta la "S" entre la "M" y "E" en "REMERA" y se reemplaza la "R" por "T" en "REMERA".

Se evidencia que una de las limitantes de esta métrica es su no-normalización en el valor, es decir, es diferente que el resultado sea 2 para un par de cadenas de longitud 3 (66% diferente) a que sea 2 para un par de cadenas de longitud 10 (20% diferente).

- *Distancia de Levenshtein utilizando el algoritmo Fussy Wuzzy:* este es un algoritmo que se basa en la distancia de Levenshtein descrita anteriormente, a partir de esta medida inicial de distancia se realiza una medida donde importa el orden alfabético de los caracteres que se suma al cálculo inicial de la distancia calculada inicialmente, a partir de las diferencias se calcula una tasa que varía de 0 a 1; esta tasa se construye comparando los verdaderos positivos, o caracteres iguales con los falsos negativos, que seria los caracteres distintos con la siguiente formula:

$$Sensitivity(s) = \frac{n(Tp)}{n(Tp) + n(Fn)}$$

Como lo muestran Rao et al. (2022) este algoritmo muestra mejores resultados que utilizar únicamente la distancia de Levenshtein.

- *Distancia de Damerau-Levenshtein:* esta métrica es una variación de la distancia de Levenshtein, pues añade transposiciones a sus reglas de conteo para pasar de una cadena a la otra. Por lo demás mantiene las mismas características de que cuando las cadenas son iguales su distancia es cero (0), y cuenta la cantidad de pasos que le toma pasar de una cadena a otra, donde esa es la distancia asociada. Esta distancia busca encontrar el camino más rápido para pasar de una cadena a otra, limitándose en las reglas que establece. Este tipo de distancia se puede ejemplificar de la siguiente manera:

  - *Damerau-Levenshtein* ("MONTAR", "MATAR"): 2. Ya que el proceso que se usa es eliminar la "N" de la palabra "MONTAR", y sustituir la letra "O" por la "A" en esta misma palabra.

  - *Damerau-Levenshtein* ("MILENA", "MENILA"): 2. Se transpone la "I" por "E" en "MILENA" y se transpone la "L" por "N" en "MILENA".

  - *Damerau-Levenshtein* ("REMERA", "CAMISETA"): 5. Se reemplaza la "R" por "C" en "REMERA", se reemplaza la primera "E" por "A" en "REMERA", se inserta la "I" entre la "M" y "E" en "REMERA", se inserta la "S" entre la "M" y "E" en "REMERA" y se reemplaza la "R" por "T" en "REMERA".

Se identifica que una de las limitantes, al igual que la distancia de Levenshtein, es su no-normalización en el valor, es decir, es diferente que el resultado sea 2 para un par de cadenas de longitud 3 (66% diferente) a que sea 2 para un par de cadenas de longitud 10 (20% diferente). Además, se evidencia que el añadir esta regla de transposición efectivamente mejora algunas diferencias, así como en el ejemplo de "MILENA" – "MENILA".

- *Similaridad de Jaro:* es una métrica basada en el conteo de caracteres que se interceptan entre las cadenas de caracteres, y la cantidad de transposiciones entre los caracteres para acercarse lo más posible de una cadena a otra. Esta, a diferencia de Levenshtein, no busca llegar de una cadena a otra, sino la similaridad entre estas. Por esto, si no hay caracteres compartidos entre las cadenas se dice que la similaridad es de cero (0), en cambio si todos los caracteres coinciden y en el orden adecuado la similaridad es de uno (1). Algunos ejemplos al momento de usar está similaridad serían

  - *Jaro* ("MONTAR", "MATAR"): 0,822. Este valor se obtiene ya que tienen 4 caracteres en común y están en orden, pero no se puede lograr, vía transposición, igualar ambas cadenas.

  - *Jaro* ("MILENA", "MENILA"): 0,889. Este valor se obtiene ya que tienen los 6 caracteres en común pero no están en total orden.

  - *Jaro* ("REMERA", "CAMISETA"): 0,625. Este valor se obtiene ya que apenas tienen 3 caracteres en común, no están en el orden adecuado y no es posible llegar, vía transposición, de una cadena a otra.

Se evidencia que esta similaridad ya está normalizada.

- *Similiaridad de Jaro-Winkler:* este es un reescalamiento de la similaridad de Jaro, pues toma el número de

caracteres compartidos entre las cadenas (en los primeros 4 caracteres) y lo multiplica por un factor (arbitrario) que está entre cero (0) y un cuarto (¼). Luego, mediante una combinación lineal similaridades de Jaro, que se operan con el factor elegido, se obtiene un valor nuevo. Este reescalamiento se hace para favorecer las cadenas que tienen un inicio muy parecido, pues se trata de tener en cuenta las raíces de las palabras, pero este no afecta el rango de la función, pues sigue siendo cero (0) si no hay ninguna similaridad de las cadenas y uno (1) si los registros son iguales. Esto son ejemplos de la similaridad Jaro-Winkler:

- Jaro-Winkler ("MONTAR", "MATAR"): 0,84. Este valor se obtiene ya que tienen 4 caracteres en común y la raíz de las cadenas casi no comparten caracteres.
- Jaro-Winkler ("MILENA", "MENILA"): 0,9. Este valor se obtiene ya que tienen los 6 caracteres en común pero no están en total orden.
- Jaro-Winkler ("REMERA", "CAMISETA"): 0,625. Este valor se obtiene ya que apenas tienen 3 caracteres en común y las raíces de las cadenas son muy diferentes.

Se confirma que esta similaridad, comparado con la de Jaro, aumenta el valor de similaridad entre cadenas que compartan algunos caracteres al inicio, por lo tanto, esta suele ser más útil cuando queremos revisar verbos que están conjugados.

- *Distancia de Hamming:* esta es una de las distancias más limitantes pues únicamente puede comparar cadenas con el mismo número de caracteres y bajo la regla de sustitución. Si dos cadenas son iguales se afirma que la distancia es cero (0), pero si se encuentran dos cadenas diferentes primero se debe verificar que tengan el mismo número de caracteres y luego contar la cantidad de sustituciones que se deben hacer para pasar de una cadena a otra. Esta distancia busca encontrar el camino más rápido para pasar de una cadena a otra, limitándose en las reglas que establece. Esta distancia se calcula de la siguiente manera:

- *Hamming* ("MONTAR", "MATAR"): No se puede calcular.
- *Hamming* ("MILENA", "MENILA"): 4. Se reemplaza la "I" por "E" en "MILENA", se reemplaza la "L" por "N" en "MILENA", se reemplaza la "E" por "I" en "MILENA" y se reemplaza la "N" por "L" en "MILENA".
- *Hamming* ("REMERA", "CAMISETA"): No se puede calcular.

Se confirma que una de las limitantes, al igual que la distancia de Levenshtein, es su no-normalización en el valor, es decir, es diferente que el resultado sea 2 para un par de cadenas de longitud 3 (66% diferente) a que sea 2 para un par de cadenas de longitud 10 (20% diferente). Además, al no poder comparar cadenas de longitud diferente su uso es muy reducido.

También, se puede diferenciar una metodología como lo es el Blocking, el cual pretende hacer uso de la integración determinista y la probabilística. Esto lo logra haciendo una primera integración determinística con sus llaves directas, y luego una probabilística sobre los cruces positivos para verificar la calidad del

cruce, y otra sobre los cruces negativos para obtener más cruces positivos. Nace como forma de optimizar las herramientas disponibles y hacer consistente la integración.

En la práctica, cuando no se cuenta con una buena calidad de las llaves de integración o por características propias de la información a cruzar, se utiliza la metodología Blocking. Estos métodos probabilísticos se suelen usar con nombres, fechas de nacimiento y lugares de nacimiento, entre otras variables propias, pero no identificadores únicos de las unidades. En la medida que estos métodos pueden resultar en cruces erróneos (mismatch), la integración probabilística no suele ser tan común y óptima si la calidad de los datos es la adecuada (Rothbard, 2015; Künn, 2015; Reiter, 2021; Wallgren y Wallgren, 2021). El motivo se explica por determinar el umbral desde el que el puntaje de similitud entre cadenas de texto se considera adecuado puede resultar ambigua, pero es el único método de integración disponible si no hay tipos y números de documentos de identificación.

En general, si el método de integración depende de parámetros seleccionados, deben documentarse y chequearse con escenarios para evaluar la robustez del método a los cambios en los parámetros. Por lo anterior, el desempeño del método de integración puede ser planteado en términos de dos errores: las unidades que siendo distintas quedan emparejadas (falsos positivos) y las unidades que siendo iguales no quedan emparejadas (falsos negativos). Un método de integración será mejor en la medida que se pueda reducir los dos tipos de errores, aunque conocer si uno de los dos tipos de errores es menos probable que otro puede ser útil para esperar de la integración sobrecobertura o subcobertura de la población objetivo. No es común que se pueda medir los dos errores de integración, de hecho, si fuese posible medirlos el método de integración debería cubrirlos para llevarlos a cero.

Para llegar a saber la calidad de la integración se debe poder detectar estos falsos positivos y falsos negativos, pues esto permite tener confianza en la base de datos obtenida. En este sentido se han desarrollado diferentes metodologías para hallar estos registros, una que aprovecha las metodologías de integración, es una variación del Blocking. En esta se hace la integración vía integración determinista y luego se usan los métodos probabilísticos sobre los pesudo-identificadores para validar que la integración sea correcta. Otras metodologías que se pueden usar para estimar dichos registros provienen de modelos probabilísticos que usan muestras para hacer esas estimaciones, también hay modelos de regresión que se pueden plantear basados en variables relevantes y, finalmente, algunos modelos de Machine Learning se han usado para predecir observaciones que sean verdaderos positivos.

Para evaluar la coherencia de los datos y detectar posibles sesgos y errores en la imputación, se aplican pruebas de consistencia a las variables. Estas validaciones incluyen comparaciones de las distribuciones y frecuencias de las variables de interés que comparten los RR. AA. con la información de la fuente primaria, teniendo en cuenta segmentos por regiones, grupos de edad y género (ver información sobre otras pruebas en la Sección 2.5 – Calidad del proceso de integración). Lo anterior, permite asegurar que los valores que se imputen reflejen

con precisión la información contenida en los RR. AA. Este es un paso clave en la medida que brinda información sobre la pertinencia de la integración realizada. Tras realizar estas validaciones y siguiendo reglas establecidas que consideran el contexto y el propósito del reemplazo, se procede con el proceso de imputación o sustitución de la información de la encuesta por las variables derivadas de los registros administrativos.

Por último, la integración debe contemplar tratamientos para variables sustitutas entre las fuentes, por lo que puede darse el caso que una variable en común entre las fuentes tenga diferentes valores para la misma unidad. Esto puede ocurrir por errores de medición que son inherentes a las operaciones de este tipo. Sin embargo, también puede ser por la naturaleza del RR. AA., lo que implica que pueden ser escalas distintas o tener una estacionalidad distinta; un ejemplo de integración de RR. AA. y encuestas se muestra en la Sección 2.10., en este caso se contrasta la información recogida por la GEIH con el formulario 201 de la DIAN, que da cuenta del pago por declaración de renta. En este caso es necesario homologar la información de las fuentes, para compararla y realizar análisis de calidad sobre la variable objetivo. Esto con el fin de generar el indicador más preciso posible.

En caso de presentarse conflictos en los valores de variables que son esenciales para el reemplazo, deben ser documentados y calculados indicadores de calidad que midan la extensión de la discrepancia entre las fuentes, con el objetivo de explicar estas diferencias en términos de la metodología propia de cada fuente de información. Lo anterior, con el propósito de contar con una justificación para los criterios de elección de la fuente más adecuada, según el objetivo del reemplazo.

2.5. Calidad del proceso de integración

La unificación de un marco conceptual para integrar a RR. AA. con encuestas bajo una misma operación requiere la producción de indicadores que informen el grado de homologación o armonización alcanzado entre estas. En un primer momento, los conceptos involucrados en la definición de la unidad de análisis son requeridos para construir los indicadores y, a su vez, los indicadores pueden apoyar la armonización de los conceptos. En un segundo momento, al consolidarse o converger a una situación estable del proceso de retroalimentación entre la definición de las unidades de análisis y el cálculo de los indicadores de cobertura, se obtiene una visión de conjunto sobre la calidad del procedimiento de reemplazo en relación con el objetivo final del reemplazo.

Las unidades de análisis en la integración no suelen alcanzar una cobertura perfecta mutua, desde el registro administrativo hacia la encuesta, como desde la encuesta hacia el RR.AA., porque las fuentes de información estadística contienen errores de medición inherentes al proceso de recolección de datos, por ejemplo, los sesgos de memoria o recordación. Estas dos perspectivas, desde el RR. AA. o desde la encuesta, son la razón del nombre de los indicadores de cobertura de dos vías. Además, los métodos de integración generan falsos cruces entre entidades disímiles como también falsos no-cruces

entre entidades que deberían haberse emparejado.

En este sentido, es importante verificar si el método de integración induce a sesgos de selección, es decir, si aquellas unidades emparejadas a nivel de los microdatos son sistemáticamente distintas en los valores de sus variables a aquellas que deben representar en la encuesta y en el registro administrativo, según la definición de la unidad de análisis. Una tercera fuente de diferencias son las definiciones utilizadas en la metodología de las operaciones estadísticas como los errores de muestreo, referencias de tiempo distintas, o las limitaciones administrativas que impone el RR.AA. en su uso estadístico, por consiguiente, en estos casos no es posible generar una homologación perfecta para las distintas fuentes.

Los indicadores de cobertura de dos vías no solo informan la decisión sobre la incorporación del ejercicio definitivo de reemplazo a la operación de la encuesta a las instancias pertinentes sino, además, constituye una operación de monitoreo de calidad para las futuras iteraciones de la producción del reemplazo en la encuesta. La medición de la calidad se debe evaluar en conjunto con la medición del costo del reemplazo versus otras alternativas que puedan estar siendo evaluadas por parte del equipo temático de la encuesta, como pueden ser operaciones de recuperación para cuestionarios incompletos o reducción de tasas de no-respuesta.

El cálculo de los indicadores de cobertura de dos vías tiene como insumos principales las operaciones de derivación de las variables en el RR.AA., o en registros estadísticos, y la integración de los microdatos de las fuentes. En este punto, se deben utilizar solo las variables que han sido seleccionadas y homologadas entre las fuentes para asegurar la comparabilidad priorizada según el objetivo del reemplazo.

La medición inicial de la cobertura comprende la medición del conjunto "Cruce + Complemento" que corresponde a aquellas que pertenecen al conjunto de cruce con el registro administrativo o pertenecen a las unidades de análisis según la definición de la encuesta. Aquellas unidades que hacen parte del análisis según la información de la encuesta, pero no están incluidas en el conjunto de referencia del cruce con RR. AA. recibe el nombre de "Complemento encuesta". Estos indicadores se comparan con el número de las unidades de análisis según el RR.AA., según la encuesta u otras fuentes antes de la integración a nivel de microdato. La utilización de los factores de expansión en las encuestas es un paso clave por lo que estos indicadores deberían estar disponibles aplicando los factores y también antes de aplicarlos.

Por ejemplo, para una variable de interés como valores de ingresos, se revisan las distribuciones de las variables en cada una de las fuentes de forma separada y también dentro del cruce a nivel de microdatos. Asimismo, para los microdatos en el cruce están disponibles las variables en común de las distintas fuentes por lo que se pueden calcular ratios o diferencias en el reporte de las distintas fuentes. Es importante destacar que las conclusiones sobre las diferencias entre las fuentes pueden cambiar si se utiliza una fórmula u otra para medirlas. Además, pueden ser sensibles a valores atípicos en la base de datos. Algunas fórmulas incluyen las diferencias de los logaritmos naturales de las dos fuentes, la razón entre las dos fuentes y la siguiente tasa que está

acotada entre -2 y 2 y puede ser calculada incluso si una de las dos fuentes tiene un valor de cero:

$$\frac{x_{i1} - x_{i2}}{\frac{1}{2}(x_{i1} + x_{i2})}$$

Donde x_i1 es el valor de la variable a comparar según la fuente 1 para el individuo i y x_i2 es lo mismo para la fuente 2. Estos indicadores pueden ser descritos utilizando deciles de la variable a comparar en algunas de las dos fuentes o variables demográficas o económicas que se consideren relevantes para el análisis. Esta descripción de las diferencias de reporte para las variables en común entre las fuentes que no son categóricas puede verse enriquecido por la inclusión de la evolución en el tiempo, cuando están disponibles la integración a nivel de microdatos para varios años, meses o días. En el caso de requerir un análisis a mayor profundidad es aconsejable utilizar los coeficientes de variación o los intervalos de confianza por errores de muestreo que tienen disponibles las encuestas, esto con la finalidad de verificar si los valores obtenidos en el cruce o según el registro administrativo, caen dentro de los intervalos de confianza.

En el caso de las variables categóricas o dicotómicas que son comunes entre las fuentes, pueden generarse tablas de tabulación cruzada con los valores en niveles o con la frecuencia relativa para cada una de las celdas. Para este caso también son ideales los gráficos *Sankey,* de cuerdas o similares que muestren las transiciones entre las categorías para las mismas unidades al variar la fuente de información, así como gráficos de series de tiempo con los indicadores de consistencia entre las fuentes.

Para algunas aplicaciones de reemplazo, será necesario un análisis a mayor profundidad sobre la correcta clasificación de las unidades de análisis en las fuentes. Esto ocurre especialmente cuando surgen obstáculos en la homologación de los conceptos de las fuentes a integrar. Para los siguientes indicadores es útil considerar la encuesta o el registro como punto de referencia y la otra fuente como sujeta a evaluación, o calcular los indicadores en ambos escenarios cuando sea posible. En el caso que sea el registro que funja como marco de referencia es importante que incluya información no solo sobre la población objetivo sino sobre otros tipos de poblaciones con el fin de evaluar los problemas de clasificación errónea entre los tipos de población o unidades de análisis.

A continuación, se presentan las definiciones del indicador de exactitud y del indicador de recuerdo (Christen, 2012), cuyos nombres provienen de la predicción de variables con enfoque algorítmico:

Indicador de exactitud: de la población de referencia de la encuesta, cuánto queda cubierta por microdatos en el cruce y a nivel agregado en el registro. Se construye a partir de los verdaderos positivos, que son aquellos registros que cruzan entre el RR. AA. y la encuesta, divididos por el total de registros objetivo del análisis, que son de la suma de los verdaderos positivos con los falsos positivos, los cuales son el total de registros de la encuesta. El indicador de exactitud se interpreta como una medida de qué tanto la definición de la unidad de análisis del registro contiene las unidades de análisis verdaderas según la fuente de referencia. El complemento del indicador de exactitud se interpreta como una sobrecobertura del registro al incluir unidades extras que no hacen parte de la

unidad de análisis como si lo fueran. En la sección 2.10 se muestra un análisis sobre el resultado de la integración de la GEIH con el formulario 201 de la DIAN.

Indicador de recuerdo: el indicador de recuerdo es la proporción de verdaderos positivos que hacen parte de la unidad de análisis según el registro administrativo y según la encuesta, divididos por la suma de los verdaderos positivos y los falsos negativos (la suma de estos dos elementos constituye las unidades totales de la encuesta); los falsos negativos son aquellos registros u observaciones que no pertenecen a la unidad de análisis según el registro, porque no cruzaron o porque en el registro tienen otra clasificación, pero en la encuesta hacen parte de las unidades de análisis. El indicador de recuerdo representa el porcentaje de unidades de análisis en la encuesta o fuente de referencia que es posible recuperar o recordar en el registro como porcentaje de las unidades de la encuesta, así que su complemento representa una subcobertura o faltante del registro con respecto a la encuesta.

Para tener una interpretación adecuada de las variaciones en el tiempo del indicador de exactitud o de recuerdo, es necesario también calcular o estimar el grado de unidades de la encuesta que no reportaron sus identificaciones o no las reportaron correctamente, con la finalidad de comprobar que este porcentaje se mantiene en niveles similares en el tiempo. De lo contrario, cualquier análisis de los indicadores en el tiempo debe incluir la variación en el reporte de calidad de las llaves utilizadas en el método de integración.

Por último, es necesario destacar que los resultados de los reemplazos suelen depender del supuesto esencial según el cual el conjunto de unidades de análisis que no fue posible incluir en el cruce de los microdatos no presentan características sistemáticamente distintas de aquellas donde fue posible. En caso de encontrarse diferencias, están deben quedar documentadas y justificadas a partir de los conceptos teóricos y a partir de los datos.

Como resultado del cálculo de los indicadores de cobertura de dos vías, se puede informar la decisión sobre la calidad evaluada a través de estos indicadores. Además, se pueden generar procesos que mejoren estos indicadores como la recalibración de los factores de expansión de las encuestas o el replanteamiento de las transformaciones aplicadas al registro administrativo. En algunos casos no será posible calcular alguno de los indicadores señalados porque las operaciones de reemplazo pueden ser muy distintas entre sí, para estos casos se recomienda señalar el motivo por el cuál no es posible calcular el indicador o realizar los análisis recomendados.

En las evaluaciones de cobertura de dos vías no necesariamente se busca coherencia perfecta entre las fuentes sino también estudiar sus discrepancias, de este modo, las diferencias entre las fuentes pueden constituirse en justificaciones para el reemplazo cuando existe una motivación apropiada para ello. Por último, estos indicadores pueden llevar a concluir que el reemplazo de las variables de la encuesta por información de los RR. AA. no es satisfactorio y es necesario obtener más periodos de referencia para madurar el procedimiento.

2.6. Integración de datos geográficos

Ubicar geográficamente las unidades administrativas o las observaciones de las bases de datos es muy útil a la hora de diseñar y focalizar programas o políticas públicas, incluso permite planear geográficamente como se va a implementar la estrategia pensada; por otra parte, posibilita esquemas diferentes para integrar información. Realizar cruces utilizando la información geográfica es posible y puede ser una opción dependiendo el caso específico; además de lo anterior, al integrar la información permite las mismas verificaciones de calidad y se tienen las mismas condiciones de remplazo explicadas en la siguiente sección de este documento.

Las encuestas y los RR. AA. también se pueden integrar geográficamente, esto depende principalmente del tipo de variables o metodología de geolocalización que tenga cada base de datos, sin importar si es un registro administrativo o una encuesta. La forma más sencilla de integrar estos dos tipos de fuentes de información es por medio de cruzar los datos utilizando una llave o código de identificación. Dado esto, es posible obtener un nivel de agregación geográfica donde los datos sean representativos; por ejemplo, agregar los datos de la GEIH a nivel de departamento e integrarlos con el registro administrativo utilizando el código DIVIPOLA. Dado esto, para poder integrar encuestas con RR. AA. de forma geográfica, lo primero que debemos hacer es identificar qué tipo de variables geográficas maneja la fuente de información.

Existen muchas formas de georreferenciar un registro administrativo o una base de datos, por lo general se utilizan variables que permiten ubicar geográficamente a la unidad administrativa. En este acápite solo se relacionan cuatro formas de ubicar las unidades administrativas espacialmente, pues son las más usadas en Colombia (CONPES 4007 y CONPES 3859), aunque existen más:

- *Código DIVIPOLA:* este código permite identificar la unidad administrativa en la división política territorial del país, donde el mínimo nivel de identificación que permite es el de corregimiento. Por lo anterior, no es una referencia precisa, pero es útil para generar datos a distintos niveles de agregación.

- *La dirección:* este campo puede ser útil a la hora de construir registros administrativos, en especial si estos cubren zonas urbanas; para las áreas rurales suelen ser imprecisas. Además, para ubicar a la unidad administrativa espacialmente puede ser complejo, dado los cambios en nomenclatura o que su digitación puede dificultar el proceso.

- *Latitud y longitud:* este se compone de dos variables; la latitud ubica cualquier punto respecto a la línea del Ecuador y la longitud ubica cualquier punto respecto al meridiano de Greenwich. Estas dos variables permiten georreferenciar cualquier unidad administrativa sin importar su procedencia, lo que implica que puede ser utilizada independientemente si el registro administrativo representa unidades urbanas o rurales.

- *Modelo LADM_COL:* Modelo para el Ámbito de la Administración del Territorio en Colombia (LADM_COL). Es el modelo oficial para los datos catastrales del país como está definido en los documentos CONPES 4007 y CONPES 3859. Este permite una descripción profunda de varios elementos cartográficos y legales. Se compone de varios elementos y el más importante se describe a continuación:

  - *Modelo Núcleo:* consiste en el mínimo de elementos necesarios para representar el territorio en Colombia. Este se compone de la interrelación de la capa catastral de los predios y la base legal de registro de propiedad de inmuebles, sumada a la capa de reservas forestales y parque nacionales. Este elemento permite la interoperabilidad entre las diferentes fuentes catastrales. Para poder aplicarlo es necesario poder identificar el código catastral y la cédula inmobiliaria de la unidad geográfica definida para cada observación.

  - *Otros componentes del modelo LADM:* este se llama en los documentos de soporte del modelo LADM-COL el modelo extendido. Hace referencia a la tipificación de los predios desde otros instrumentos de ordenamiento territorial como los POMCAS, Planes de Ordenamiento territorial, también incluye información como el valor del impuesto predial o los avalúos prediales.

En general existen dos formas de integrar datos que están georreferenciados; cuál utilizar, depende de cómo está construida la base de datos, la información que contiene y si es o no una base de datos geográfica (lo que implica que la información se puede plasmar con alguna figura geométrica en un mapa). A continuación, se describen las dos formas de integrar datos georreferenciados:

I. Si se tiene alguna llave de integración, como la cédula catastral, el folio de matrícula, incluso la dirección de la unidad administrativa es posible realizar una integración determinística o probabilística. Para lograrlo es muy posible que se deban surtir los pasos anteriormente descritos en lo referente a la calidad de las llaves de integración; un ejemplo, es que muy posiblemente para realizar la integración de dos bases de datos utilizando la dirección de la unidad administrativa u observación, sea necesario realizar varios procesos de limpieza y estandarización, y sea necesario recurrir a una integración probabilística, dada la calidad de la variable.

Dependiendo el caso, también es posible agregar la información a un nivel dado al que se puedan integrar ambas fuentes; por ejemplo, si se tiene una encuesta que solo es representativa a nivel de municipio o corregimiento, se podría integrar la información de esta encuesta para complementar la información del registro administrativo. Esto es válido siempre y cuando el objetivo del análisis sea a nivel de corregimiento o a un nivel de agregación mayor. Para este ejercicio se pueden utilizar llaves como el código DIVIPOLA.

Este tipo de integración permite imputar información sobre unidades

administrativas o registros que están georreferenciados y se pueden plasmar espacialmente, lo que implica que si se tiene una llave en las bases de datos que permita integrar la información es posible integrar cualquier fuente de información con una base de datos geográfica.

II. La segunda forma de integración de información tiene unos requerimientos muy detallados. Es necesario que ambas bases de datos, tanto el registro administrativo como la encuesta, sean bases de datos geográficas, lo que implica que la información tiene un reflejo geométrico y se puede graficar en forma de puntos, líneas o polígonos (Sutton. T, 2009).

Dependiendo las condiciones de las bases de datos geográficas se puede realizar una integración espacial (esta función se conoce "spatial joint"; Sutton. T, 2009). Lo que se busca es identificar la mínima distancia entre las diferentes formas ("shapes") y de esta manera poder emparejar las observaciones o las unidades administrativas; el análisis de distancia, o la creación de una matriz de distancia euclidiana son un esfuerzo extra, pero funcionan cuando se tienen dos bases geográficas compuestas de puntos (Sutton. T, 2009). Si ambas bases de datos funcionan bajo el modelo LADM COL es más sencillo integrarlas y realizar verificaciones alfanuméricas y espaciales posteriores, las cuales se pueden complementar.

Además de lo anterior también es posible integrar imágenes, como las coberturas vegetales a las bases de datos geográficas. Este proceso requiere imágenes de gran calidad, pues consiste en asignar los pixeles de la imagen al punto, línea o polígono más cercano; este proceso se conoce como "Raster" (Sutton. T, 2009).

La integración espacial se puede realizar de varias formas dependiendo las condiciones de las bases de datos, este tipo de ejercicios no solo resultan muy útiles para las instituciones del SEN en su ejercicio misional, sino que también permite generar rutinas de verificación de la información para asegurar la calidad de la información y su integración.

2.7. Reemplazo a partir de la integración

Cuando se busca emplear los registros integrados para sustituir valores faltantes o no válidos en una pregunta, o completar una pregunta que en su mayoría no es contestada en la encuesta se puede recurrir al método de imputación conocido como "cold deck". Este método implica reemplazar el valor faltante en la encuesta con el valor obtenido de una fuente externa, en este caso, de registros administrativos. En este contexto, es fundamental considerar la implementación de reglas de remplazo que guíen las decisiones en cada escenario posible. Es decir, se debe definir de manera específica qué valor se imputará cuando se identifique una discrepancia entre un valor en la encuesta y el informe del registro administrativo, o

cuando la encuesta reporte un valor que no tenga una correspondencia en el registro administrativo.

Es importante señalar que este documento no detalla reglas específicas, ya que estas pueden variar según el caso y la temática en la que se esté trabajando. Sin embargo, existen unos pasos generales que se deben seguir:

1. Comprender de manera más profunda los mecanismos subyacentes de la falta de respuesta mediante la recopilación de información sobre aquellos que no respondieron. Para lograrlo, es esencial evaluar y analizar las características de los encuestados que optaron por no responder a la pregunta. Este análisis se realiza con variables demográficas, geográficas y otras obtenidas de fuentes administrativas, para caracterizar e identificar si los no respondientes son aleatorios o si pertenecen a algún grupo específico de la población.

Este análisis facilita la determinación de si el registro administrativo utilizado para abordar la información faltante cuenta con cobertura tanto para la población general como para los grupos específicos identificados. Este enfoque contribuye al éxito en la recuperación de la información faltante de la encuesta mediante el uso del registro administrativo.

2. Establecer las reglas de reemplazo. En este punto, se debe considerar si es necesario realizar alguna modificación o cálculo previo a la sustitución del valor en la pregunta de la encuesta. Esto implica decidir si el valor se toma directamente del registro o si se realizará una operación como una división, promedio o porcentaje del valor encontrado en el registro. Además, es esencial definir cómo actuar en caso de encontrar discrepancias entre los valores en las observaciones de la encuesta y del registro. En este escenario, se puede optar por tomar el mayor valor entre los dos o confiar siempre en el valor del registro, entre otras opciones. Sin embargo, estas decisiones deben tomarse considerando el contexto de la información que se busca sustituir. Los ejemplos mencionados son casos comunes, pero en cada situación específica pueden surgir circunstancias particulares que requieran una evaluación detallada para tomar la mejor opción.

3. Examinar las reglas de reemplazo elegidas. Esto implica realizar un análisis detallado mediante estadísticas descriptivas y evaluación de resultados. Es esencial determinar en qué medida la elección de una regla sobre otra puede impactar los resultados de los indicadores o cálculos que emplean esta variable después de la sustitución de los valores. Esta evaluación posibilita considerar diversos escenarios y tomar una decisión más informada respecto a la mejor opción. Además, se sugiere compartir esta evaluación con las instancias de control y aprobación, aspecto que se aborda con mayor profundidad en la siguiente sección (4.8 Instancias de control y aprobación). En este proceso, la

participación de expertos en las temáticas relevantes resulta crucial para garantizar la coherencia de las decisiones tomadas.

4. Es imperativo documentar las decisiones tomadas y las validaciones realizadas en la selección de estas reglas. Este registro proporcionará referencia para futuros ejercicios similares que puedan desembocar en discusiones similares. Además, si se produce algún cambio en el registro administrativo o en el proceso de homologación de la información con la encuesta, esta documentación facilitará ajustes oportunos a las reglas establecidas.

Igualmente, de lo provechoso del reemplazo, la derivación de nuevas variables producto de esta integración entre RR. AA. y encuestas puede ser un buen producto. Ya que, al tener dos fuentes de información diferente conectadas en una misma base, se puede obtener información nueva que se diferencia y complemente la original. En este sentido, el equipo temático de la operación genera transformaciones a la base resultante según las necesidades de información y el alcance temático de las anteriores bases.

2.8. Instancias de control y aprobación

Una operación de reemplazo debe inscribirse en el marco de gobierno de datos de la entidad, dar cumplimiento a los lineamientos de la Norma de Calidad Estadística y en los casos que aplique, contar con el acompañamiento, la retroalimentación y la aprobación de instancias distintas al equipo de trabajo responsable de diseñar, ejecutar y documentar el reemplazo.

En cuanto a los lineamientos de calidad, establecen que las metodologías de imputación y remplazo deben estar documentadas en la operación estadística, incluyendo aspectos como diseño metodológico, responsables y medidas de calidad/validación. Comúnmente, los procesos de imputación y reemplazo cuentan con la validación del equipo de diseños muestrales de la entidad o de un experto temático / estadístico.

Adicionalmente, de acuerdo con los lineamientos de buenas prácticas de producción estadística, se cuenta en DANE con la realización de pre-comités. En estos espacios las áreas temáticas, logística y de diseños muestrales (si aplica), validan de forma detallada la aplicación del proceso estadístico de acuerdo con los resultados obtenidos y el diseño de la operación estadística. A su vez, a acuerdo con la normatividad vigente, la mayoría de las operaciones estadísticas en DANE deben realizar comités internos de forma previa a la divulgación de los resultados. En los comités internos se valida la aplicación del proceso estadístico, con participación del área temática, logística, diseños muestrales (si aplica), representantes de subdirección/dirección, representantes de DIRPEN y expertos temáticos de otras áreas del DANE.

Por otra parte, existen operaciones estadísticas que, debido a su relevancia y complejidad temática, previo a su divulgación deben ser sometidas a la validación de un comité externo de expertos, quienes analizan el proceso y resultados y buscan asegurar su calidad.

El organigrama del DANE establece las instancias de control interno y de toma de decisiones de los equipos de trabajo. En consecuencia, antes de la publicación de cualquier resultado de reemplazo deben haberse surtido las validaciones y las aprobaciones que resulten aplicables. Las instancias de control y aprobación deben conocer en las fuentes de información integradas para evaluar las repercusiones positivas o negativas de la publicación de los resultados del reemplazo en los resultados de las operaciones relacionadas con el área temática. Esto para garantizar el mejor resultado y el uso de la información estadística por parte del usuario final que la demanda.

Los proveedores de los RR. AA. o las entidades del Gobierno que estén interesados en conocer los resultados de la operación de reemplazo, pueden consultarlos al equipo temático para obtener comentarios o sugerencias orientadas a mejorar la calidad de los resultados obtenidos.

2.9. Instrumentos del reemplazo

Al implementar estos lineamientos, se generan documentos y códigos que facilitan y orientan la toma de decisiones en cada proceso. La generación de estos instrumentos debe seguir la misma estructura para cada una de las etapas, la cual es la siguiente: identificar los insumos disponibles; elaborar e implementar una serie de transformaciones, reglas y estándares que habiliten el aprovechamiento estadístico del registro administrativo; calcular y presentar los indicadores de calidad que permitan evaluar estas transformaciones e integraciones, y presentar el resultado final que satisface la necesidad estadística planteada en el objetivo.

Es fundamental que la documentación incluya discusiones y argumentos que respalden las decisiones tomadas a lo largo del procedimiento de reemplazo. Esto facilita que el usuario final defina si los resultados estadísticos son aprovechables para su objetivo y proporciona el contexto metodológico adecuado para mostrar las limitaciones y las fortalezas de la operación de reemplazo. Además, permite dirigir de manera más efectiva las discusiones de naturaleza metodológica que puedan surgir durante la ejecución del reemplazo. En particular, la documentación respalda la determinación del calendario de publicaciones, la definición de la relación entre las nuevas estadísticas y las anteriores, los plazos previstos para cada fase del procedimiento desde la recepción de los insumos hasta la publicación, la identificación de los riesgos de la operación y cómo mitigarlos en el futuro.

Los instrumentos del reemplazo también comprenden los códigos o programas utilizados en el procesamiento de la información, estos códigos deben mantener un formato que permita leerlos fácilmente, incluyendo comentarios y títulos descriptivos para cada sección. Lo anterior, de forma que las ejecuten personas externas al equipo y puedan replicar los resultados. En ciertos casos, será recomendable aplicar procedimientos de espejo para asegurar que el procesamiento de los datos sea conforme a los esperado. La ejecución de los códigos puede estar orientada por manuales que describan la función de cada código haciendo más fácil la reproducibilidad de los resultados y la

producción de los mismos indicadores en el futuro.

Por otro lado, los datos crudos, intermedios y finales se almacenan adecuadamente para asegurar respuestas oportunas a las preguntas de los usuarios o para garantizar la continuidad de la producción, especialmente ante cambios en el equipo de trabajo. Además, es importante tener en cuenta la versión de los registros utilizados, ya que en el futuro puede haber una versión actualizada del mismo registro que genere diferencias en los resultados al replicar los ejercicios. También hay que considerar que en las presentaciones de resultados se debe incluir una nota que indique la versión del registro utilizada, ya que pueden producirse cambios o actualizaciones en los registros tras la publicación de los resultados. Esta versión indicará que no se han tenido en cuenta cambios posteriores a esa fecha.

Adicionalmente, los instrumentos del reemplazo son un elemento indispensable para la construcción de los metadatos que registran los remplazos realizados en la base. Los metadatos son el conocimiento acumulado de los procesos necesarios para llegar al resultado final y son estos los que orientan a las personas para entender el cómo se configura la información y cómo utilizarla correctamente.

2.10. Ejercicio de ingreso disponible

En esta sección se presenta a manera de ejemplo de la aplicación de los lineamientos descritos, un ejercicio de la integración de la gran encuesta de Hogares y los registros tributarios de la DIAN 2019, con el propósito de calcular el ingreso disponible de las personas en Colombia (ingresos del hogar luego de descontar los impuestos). Este cálculo permite realizar análisis más precisos sobre la distribución del ingreso y refinar el índice de Gini, lo que en ultimas permite desarrollar políticas públicas y programas sociales.

Objetivo y justificación

Dentro de los lineamientos de los países miembros de OCDE, se encuentra reportar ciertos indicadores relacionados con la desigualdad en el ingreso y la pobreza. Entre estos indicadores se encuentra el ingreso disponible, que es una medida que permite observar el ingreso del que disponen los hogares para respaldar sus gastos de consumo y ahorro, luego de descontar los impuestos, a lo largo de un período de referencia (OECD, 2017; Cabrera, 2011).

Para poder informar sobre el ingreso disponible en Colombia, es fundamental contar con datos sobre los impuestos directos pagados tanto en relación con el ingreso como con la propiedad. Por este motivo para calcular este indicador se integran una fuente primaria con un registro administrativo, pues se utiliza la información recopilada a través de la Gran Encuesta Integrada de Hogares (GEIH) y registros administrativos de la DIAN sobre el pago de impuesto a la renta, con el fin de realizar correcciones sobre la variable que captura la GEIH sobre la declaración de renta.

En este ejercicio se evalúa la información de pago de impuestos recogida en la GEIH y la disponible en los registros administrativos, específicamente los reportados por la DIAN, quien es la entidad que recoge y

administra los registros relacionados con impuestos.

Con la información disponible se busca realizar un contraste de la variable de pago de impuesto de renta que se recoge en la GEIH con la que se encuentra en el registro administrativo del 2019; lo anterior, para evaluar la viabilidad de hacer una imputación sobre personas que realizaron pago de impuesto de renta en 2022, utilizando características y variables claves que permitan identificar las personas que declaran renta.

Información

Registros tributarios

Desde 2021, la GEIH incluyó nuevas preguntas sobre el pago de impuestos, gracias a la actualización del marco muestral con el Censo poblacional realizado en Colombia en 2018. Entre las preguntas que se agregaron, se encuentran las siguientes:

Para asalariados e independientes (15 años o más):

¿Le descontaron retención en la fuente a lo que ganó el mes pasado en este empleo?, ¿Cuánto?
Para toda la población dentro de la PET (15 años o más):
¿Es propietario de una o varias propiedades inmuebles?
Durante los últimos doce meses, ¿cuánto pagó por impuesto predial de su(s) propiedad(es)?
Durante los últimos doce meses, ¿cuánto pagó por impuesto de valorización de su(s) propiedad(es)?
Durante los últimos doce meses, ¿realizó el pago de impuesto de vehículos?
Durante los últimos doce meses, ¿realizó el pago de impuestos a la renta y complementarios?
Durante los últimos doce meses, ¿realizó el pago de impuestos a ganancias en juegos de azar, chances, loterías, indemnizaciones, liquidaciones, venta de propiedades, acciones, vehículos, etc.?

Adicionalmente, el DANE cuenta con los registros del formulario 210 de la Dirección de Impuestos y Aduanas Nacionales (DIAN), correspondiente a la "declaración de renta y complementario de personas naturales y asimiladas residentes y sucesiones ilíquidas de causantes residentes", para 2017, 2018 y 2019. Lo anterior indica que no está disponible un registro administrativo de la DIAN actualizado, y por su naturaleza, es difícil contar con esto, ya que los reportes de declarante del año vencido cierran aproximadamente a mediados de octubre del presente, y pueden tener pagos extemporáneos el resto del año.

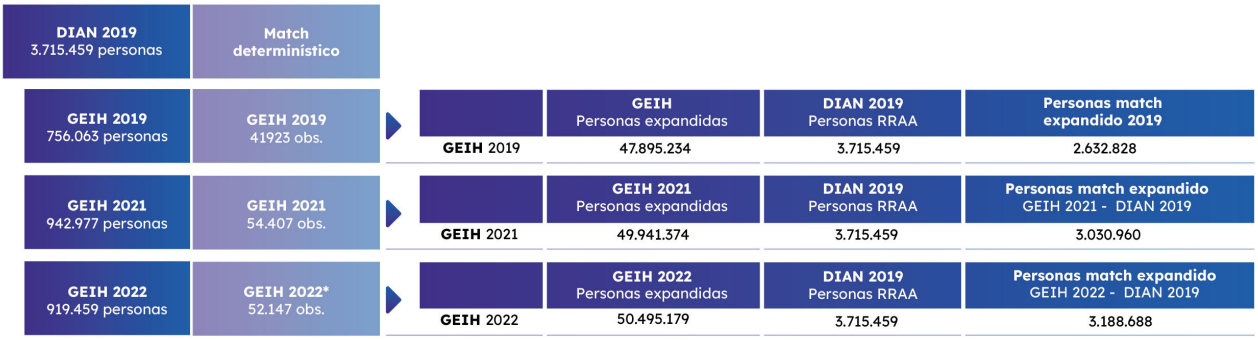
Cruce entre la GEIH y el registro administrativo de la DIAN:

Para realizar el cruce entre la GEIH y el formulario 2010 de la DIAN es necesario realizar dos procesos distintos descritos en las secciones anteriores de ese documento. Primero es necesario hacer rutinas de limpieza sobre las variables que son las llaves de integración, en este caso tipo y número de documento. El segundo proceso, consiste en realizar una verificación sobre la variable de tipo de documento, cruzando ambas bases con la base BDUA que sirve como referencia para esa variable.

Al analizar el cruce entre la GEIH y el formulario 210 de la DIAN, se encontró que el cruce expandido (usando factores de expansión censo 2018) entre el último registro administrativo disponible (2019) con la Gran Encuesta Integrada de

Hogares, llega a un valor similar de las personas que realmente declararon renta en 2019 (3.715.459). Específicamente, en 2019, utilizando la GEIH que se basa en el marco muestral del censo de 2005, cruzaron 41.923 personas, que expandidas representaban a 2.632.828 personas, logrando así una cobertura expandida del 70,9% del recaudo de impuestos de renta. Por su parte en 2022, cruzaron 52.147 personas, las cuales expanden a 3.188.688 (ver Figura 1).

Figura 2. Resultado del cruce DIAN 2019 - GEIH 2019/2022



Fuente: DANE - GEIH 2022. Cálculos propios.

Al comparar los registros administrativos de declaración de renta de personas naturales (Formulario 210 de la DIAN) con la nueva pregunta de autoreporte de la GEIH, se encontró que sólo 7.018 personas (sin expandir) dijeron haber declarado renta y efectivamente declararon renta en 2018 (ver Figura 2), otras 2.260 personas dijeron haber declarado renta, pero no estaban en el registro, lo cual puede indicar que son parte de los nuevos declarantes de renta entre 2020 y 2022. Sin embargo, 44.412 personas afirmaron no haber declarado renta, pero sí estaban en el registro de declarantes de renta de 2019. Es poco probable que un porcentaje tan alto de declarantes de renta haya salido de la base en solo tres años, lo que sugiere que estas personas forman parte del subreporte que se observa en la GEIH.

Finalmente, otras 267 personas se encuentran en el registro de declarantes de renta de la DIAN, pero no en la GEIH por su edad, ya que la pregunta en la encuesta de hogares se realiza sólo para los mayores a 15 años, mientras que en la actualidad en la DIAN no hay un mínimo de edad para declarar impuestos. En registros anteriores de la DIAN, los padres solían reportar los ingresos de los menores, pero actualmente no hay un mínimo de edad para declarar.

Figura 3. Comparativa pregunta auto - reporte y cruce DIAN 2019 - GEIH 2022

Nuevos declarantes

Cruce DIAN 2019 - GEIH 2022

	Declarantes DIAN		
Declarantes GEIH	**Si**	**Missing**	**Total**
Si	7.018	2.260	9.278
No	44.412	657.253	701.665
NS/NR	450	2.338	2.788
Menores de 15 años	267	205.461	205.728
Total	**52.147**	**867.312**	**919.459**

Declararon ante la DIAN y dijeron no haber realizado el pago del impuesto de renta

Fuente: DANE - GEIH 2022. Cálculos propios.

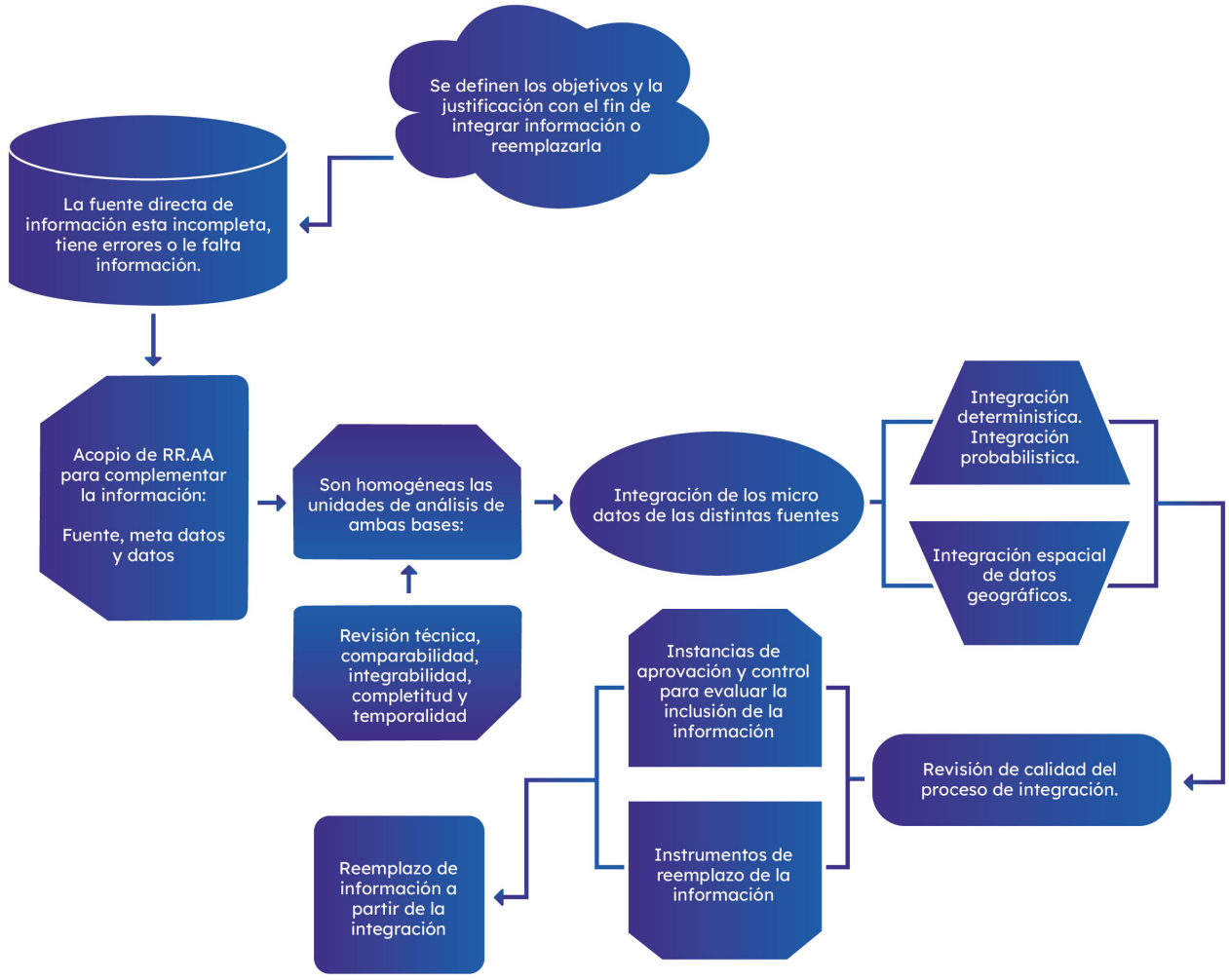
Los resultados comparativos entre el registro y el auto reporte indican que el 85,2% de las personas incluidas en el formulario 210 de la DIAN en 2019 afirmaron no haber declarado renta o no pagar impuesto de renta en la GEIH (Tabla 2), lo que sugiere que la pregunta de la GEIH no resultó ser informativa y que el registro tributario es la mejor fuente de información para corregir la declaración de renta.

Con lo anterior, se concluye que el auto reporte de los encuestados en la GEH no es informativo de los declarantes de renta en el país y aunque se cuenta con registros administrativos para corregirla, el calendario tributario establecido en el país no permite que la información de la DIAN llegue a tiempo para realizar la corrección y reporte. Por lo tanto, se podría emitir un reporte temporal en cada año con la información de la GEIH, mientras se espera la disponibilidad del registro para su corrección definitiva.

3. Conclusiones

A lo largo del documento se pueden identificar varios procedimientos que permiten integrar RR. AA. o integrar RR. AA. con fuentes directas o encuestas, realizando controles de calidad y generando instancias de control y aprobación. También pudimos analizar métodos de reemplazo y métodos que permiten analizar la calidad de la integración. Por último, analizamos métodos de integración determinística, probabilística y geográfica. A continuación, presentamos un flujograma que sintetiza las diferentes etapas del proceso.

Figura 4. Flujograma para integrar RR. AA. y fuentes directas



Fuente: elaboración propia.

Cabe anotar que el aprovechamiento estadístico de los RR. AA. también involucra nuevos tipos de riesgos que se diferencian de los riesgos de las operaciones estadísticas de encuestas. Entre estos riesgos, es posible encontrar que los microdatos necesarios para las integraciones no estén disponibles en el momento requerido, problemas de cobertura, cambios en el modelo de datos de un año a otro, cambios de normatividad o dirección estratégica que los proveedores de los RR. AA. y la falta de cooperación de las entidades administradoras de los RR. AA. para garantizar la calidad, entre otros.

Adicionalmente, si bien este documento no incluye información sobre las demandas potenciales de recursos de tecnologías de la información como capacidad de procesamiento, almacenamiento, acceso, conectividad y software, los requerimientos deben hacerse según los recursos disponibles y la demanda requerida por el reemplazo considerando que la disponibilidad de recursos de tecnologías de la información es un insumo esencial para aplicar el reemplazo de información de encuestas con RR. AA.

Para finalizar, el seguir este procedimiento asegura que la información que se integra desde los RR. AA. tenga los mejores estándares de calidad y permita enriquecer los análisis pensados. Este proceso le puede ser muy útil a las instituciones del SEN pues les permite generar información precisa y de calidad a bajo costo, además permite darle usos diferentes a la valiosa información que generan las instituciones en su actuar misional.

4. Referencias

Angel, S., Disslbacher, F., Humer, S., & Schnetzer, M. (2019). What did you really earn last year? explaining measurement error in survey income data. Journal of the Royal Statistical Society: Series A (Statistics in Society), 182(4), 1411-1437.

Bound, J., Brown, C., Duncan, G. J. and Rodgers, W. L. (1994) Evidence on the validity of cross-sectional and longitudinal labor market data. J. Lab. Econ., 12, 345–368.

CAN (2013), Decisión 780

Congreso de la República de Colombia (2011). Ley 1448 de 2011. Bogotá.

Corte Constitucional de la República de Colombia (2004). Sentencia T-025 de 2004. Bogotá. CEPAL. (2020). Continuity of household surveys after the coronavirus disease (COVID-19) pandemic. ECLAC.

Chun, A., Larsen, M., Durrant, G., & Reiter, J. P. (2021). Administrative Records for Survey Methodology. John Wiley and Sons.

Christen, P. (2012). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection.

DANE. (2018). Metodología de Diagnóstico de los Registros Administrativos para su aprovechamiento estadístico. Bogota D.C: DANE.

DANE. (2021). Anexo técnico No. 2: Construcción del indicador inasistencia escolar del Índice de Pobreza Multidimensional 2020. Bogotá D.C.: DANE.

DANE. (2021). Boletín Técnico: Índice de Precios de la Propiedad Residencial (IPPR) - II trimestre 2021. Bogotá D.C.: DANE.

DANE. (2021). Nota Metodológica: Microdatos de ingresos para la medición de pobreza monetaria y desigualdad 2020. Bogotá D.C.: DANE.

DANE. (2022). Nota metodológica – Uso del RR.AA Registro Único de Víctimas para la estimación del pobreza monetaria y pobreza extrema 2019-2021. Bogotá D.C.: DANE.

DANE. (2023). METODOLOGIA GENERAL DE CUENTAS TRIMESTRALES. Bogotá D.C.: DANE.

DNP (2012), documento CONPES Social, Metodologías oficiales y arreglos institucionales para la medición de la pobreza en Colombia.

DANE (2024). Buenas practicas para la configuración de registros administrativos para su aprovechamiento estadístico en el SEN. Bogotá D.C.:DANE.

EUROSTAT. (2017). Tourism statistics: Early adopters of big data? Statistical Working Papers Sutton, T. (2009, April 1). Gentle gis introduction. https://docs.qgis.org/3.34/pdf/es/QGIS-3.34-GentleGISIntroduction-es.pdf

Presidencia de la República de Colombia (2011). Decreto 262 de 2004. Bogotá.

Presidencia de la República de Colombia (2011). Decreto 1170 de 2015. Bogotá.

Decreto 2404 de 2019
Instituto Nacional de Estadística de Uruguay. (2021). Marco conceptual y metodológico del Sistema

Instituto Nacional de Estadística de Uruguay. (2012). Encuesta Longitudinal de Protección Social (ELPS) https://www.elps.org.uy/elps/file/73/1/metodologia-de-ponderacion-elps-ola-1-1.pdf.

Integrado de Registros Estadísticos y Encuestas – SIREE.

Kim, C. and Tamborini, C. R. (2014) Response error in earnings: an analysis of the survey of income and program participation matched with administrative data. Sociol. Meth. Res., 43, 39–72.

Künn, S. (2015). The challenges of linking survey and administrative data. IZA World of Labor.

Little, R., & Rubin, D. (2002). Statistical Analysis with Missing Data. John Wiley and Sons.

Meyer, B., & Mittag, N. (2021). Combining Administrative and Survey Data to Improve Income Measurement. En A. Chun, M. Larsen, G. Durrant, & J. Reiter, Administrative Records for Survey Methodology (págs. 297-322). John Wiley and Sons.

Moore, J. C., Stinson, L. L., & Welniak, E. J. (2000). Income measurement error in surveys: A review. Journal of Official Statistics-Stockholm, 16(4), 331-362.

Rao, Peluru Janardhana; Rao, Kunjam Nageswara & Gokuruboyina, Sitaratnam. (2022). An Experimental Study with Fuzzy-Wuzzy (Partial Ratio) for Identifying the Similarity between English and French Languages for Plagiarism Detection. International Journal of Advanced Computer Science and Applications, Vol. 13, No. 10, 393-401.

Reiter, J. (2021). Assessing Uncertainty When Using Linked Administrative Records. En A. Chun, M. Larsen, G. Durrant, & J. Reiter, Administrative Records for Survey Methodology (págs. 139-154). John Wiley and Sons.

Rothbard, A. (2015). Quality Issues in the Use of Administrative Data Records. En J. Fantuzzo, & D. Culhane, Actionable intelligence: Using Integrated Data Systems to Achieve a More Effective, Efficient, and Ethical Government (págs. 77-104). Palgrave Macmillan.

Statistics Norway. (2012). This is Statistics Norway – an institution that counts. https://www.ssb.no/en/omssb/jobb-i-ssb/hvorfor-jobbe-i-ssb/_attachment/116316?_ts=13ef53ba940

UNECE. (2011). Canberra Group Handbook on Household Income Statistics. Geneva: United Nations.

Unidad para la Atención y Reparación Integral a las Víctimas (2013). Instructivo de caracterización. Bogotá. Recuperado el 20 de noviembre de 2020 de <https://www.unidadvictimas.gov.co/es/publicaciones-periodicas/15760>.

United Nations. Commission économique pour l'Europe, & United Nations. Economic Commission for Europe. (2007). Register-based statistics in the Nordic countries: review of best practices with focus on population and social statistics.

Wallgren, A., & Wallgren, B. (2007). Register-based statistics: administrative data for statistical purposes (Vol. 553). John Wiley & Sons.

Wallgren, A., & Wallgren, B. (2016). Estadísticas basadas en registros: Aprovechamiento estadístico de datos administrativos. México: Instituto Nacional de Estadística y Geografía.

Wallgren, A., & Wallgren, B. (2014). Register-based statistics: statistical methods for administrative data. John Wiley & Sons.

Wallgren, A., & Wallgren, B. (2021). Toward an Integrated Statistical System Based on Registers. Washington D.C.: Banco Interamericano de Desarrollo.

Zhang, L. C. (2011, August). Topics of statistical theory for register-based statistics. In ISI conference, Dublin (pp. 22-26).

Zhang, L.-C. (2021). On the Use of Proxy Variables in Combining Register and Survey Data. En A. Chun, M. Larsen, G. Durrant, & J. Reiter, Administrative Records for Survey Methodology (págs. 3-24). John Wiley and Sons.

www.sen.gov.co


@DANE_Colombia


/DANEColombia


/DANEColombia


@DANEColombia