COLOMBIA POTENCIA DE LA VIDA

DANE DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA 70 AÑOS

Integración de Encuestas con Registros Administrativos

Octubre de 2023

Contenido

1. Antecedentes
2. Etapas para la integración de encuestas con registros administrativos
3. Ejemplo de integraciones de Registros con Encuestas

Antecedentes en el DANE

2020

- **COVID-19** trajo consigo grandes retos en términos de producción estadística.
- Reducción importante del formulario tradicional GEIH.
- Utilizar RRAA de ayudas institucionales y pensiones como otras fuentes de ingreso para producir las estimaciones de pobreza monetaria y desigualdad.

2021

Uso de RR.AA. en distintas áreas temáticas:

- RR.AA. de la Superintendencia de Notariado y Registro (SNR) y la Unidad Administrativa Especial de Catastro Distrital (UAECD) para el calculo del Índice de Precios de la Propiedad Residencial (IPPR).
- Generación del Registro Estadístico de Relaciones Laborales (RELAB) desde el registro de seguridad social (PILA).
- Registro de instituciones educativas y matriculados (Formulario C-600) y Sistema Integrado de Matrícula (SIMAT) para variables de educación para el cálculo de pobreza multidimensional.

2022

- Estimaciones de pobreza monetaria para población Victima en Colombia.
- Modelos de áreas pequeñas.
- Cálculo de ingreso disponible e indicadores de pobreza y desigualdad.

Fuente: DANE.

¿Por qué integrar registros y encuestas?
Ventajas y desafíos

- Complementa las unidades de análisis de la población y de las variables.
- Apoyar la imputación de elementos sin respuesta o con valores atípicos de un cuestionario.
- Enriquecer la encuesta con nuevas variables u adicionar opciones de respuestas de las preguntas.
- Realizar contraste de información, para analizar los sesgos de respuesta y errores de medición.
- Construir modelos de áreas pequeñas.

Ventajas:

1. Bajos costos relativos en la producción de información estadística.
2. Mayores niveles de desagregación geográfica y mayor frecuencia en el reporte de los datos.
3. Reducción de la carga de realizar una encuesta.
4. Permite reducir los errores de medición propios de las variables más sensibles.

Desafíos:

1. Registros Administrativos con baja cobertura de la población.
2. Hay Registros Administrativos con una actualización no periódica.
3. Los registros pueden no estar estandarizados.
4. Limitaciones en la capacidad técnica y tecnológica para procesar e integrar los datos
5. Gestión de proveedores de Registros Administrativos

Fuente: DANE.

Etapas para la integración de encuestas con registros administrativos

Fuente: DANE.

Unidades de análisis homogéneas
Adaptación de la información

Inspeccionar los errores, inconsistencias, duplicados y cobertura del registro. Verificar la frecuencia de la distribución de observaciones y la consistencia.

Un reto es decidir la unidad de análisis más precisa.
No depende solamente del RR.AA. empleado, sino también del propósito de la integración y de las fuentes de datos a utilizar.

Importante tener en cuenta: los conceptos del área temática y el contraste con estadísticas descriptivas generadas de la integración.

- Ejemplo programa de transferencias monetarias *Familias en Acción* (DPS), un caso de una base de datos por persona-transacción, pero que en la práctica se debe ver como una base de datos hogar-transacción.

Fuente: DANE.

Integración de los microdatos
Métodos

Emparejamiento determinístico

- Llave de integración conjunta única (Tipo y número identificación)
- Integración exacta de las variables
- Tasa de falsos positivos suele ser baja
- Emparejamiento en etapas – Validación fonética

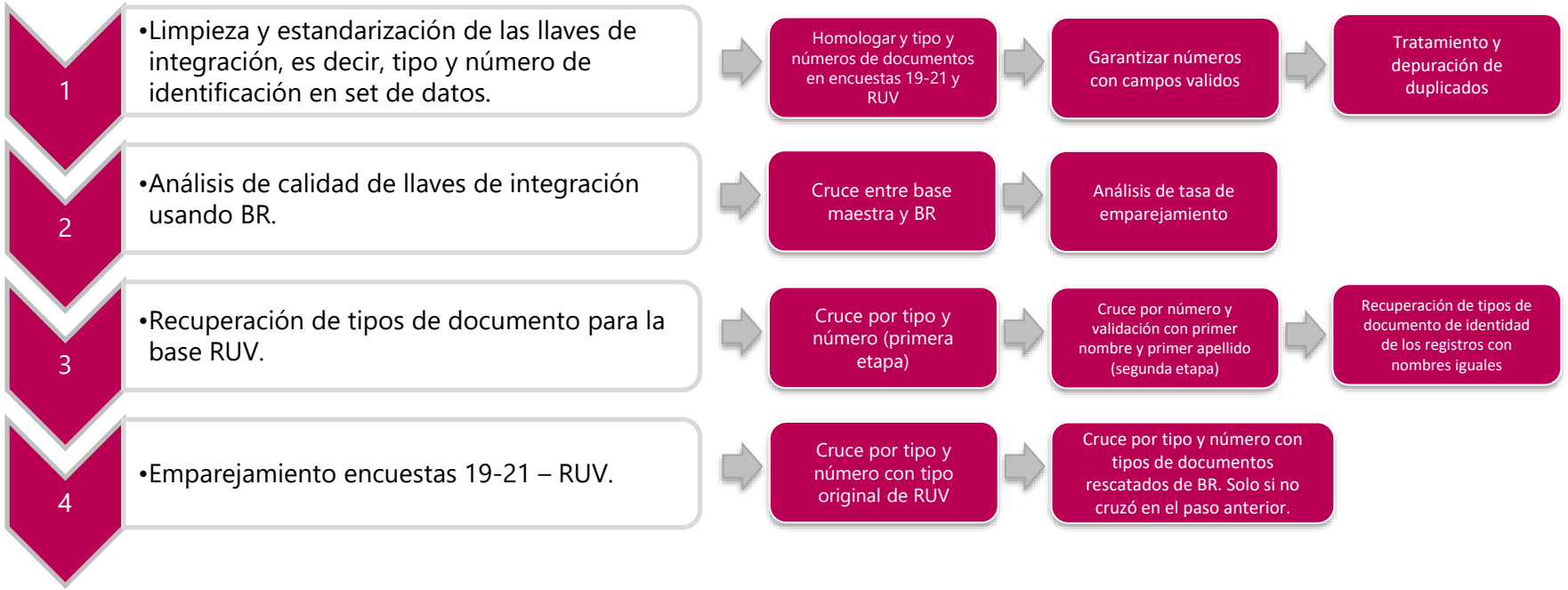
Emparejamiento probabilístico

- Llaves de integración son cuasi-identificadores (Nombres)
- Costo computacional más alto
- Comparación de cadenas de caracteres (Levenshtein, Jaro-Winkler)
- Blocking. Emparejamiento probabilístico como segunda instancia

Fuente: DANE.

Registro único de víctimas – RUV

Protocolo de integración del RUV y la GEIH



Tasas de pobreza y desigualdad - Imputación Coldeck

Integración del registro de seguridad social+ RRAA de ayudas + GEIH

Desafío:

- Durante el período comprendido entre marzo - julio la recolección enfrento algunos retos, el primero fue el cambio de la entrevista directa (cara a cara) a un operativo telefónico. Y el segundo es la validez del operativo de recuento realizado en diciembre debido a la reducción del cuestionario.

Objetivo:

- Utilizar registros administrativos para mejorar la GEIH y reducir los posibles sesgos de cobertura (falta de respuesta) y precisión.

Fuentes:

- Registro de seguridad social– pensionados –PILA
- Ayudas del Gobierno i.e FA; JA; CM
- Encuestas a hogares- GEIH
- RUV
- Remesas

Metodología

Input

- Vincular el registro de prensiones con GEIH
- Vincular el registro de ayudas con la GEIH

Armonización en unidad estadística y variable de tiempo

- El concepto de hogar es diferente en la encuesta y los RRAA
- El período de referencia de la encuesta y los RRAA es diferente

Evaluación de la calidad de insumos y salida

- Evaluar las diferencias en cobertura y no respuesta en las diferentes fuentes Validamos la consistencia entre lo que se reportó en la encuesta y en los RRAA por un mismo hogar

Resultados y validación

- Utilizamos el 2019, como periodo de referencia para validar los resultados en un periodo sin estado pandémico, y comprobar si la integración sesga los resultados oficiales.

Cálculo de ingreso disponible

Integración del registro de seguridad social+ RRAA DIAN+ GEIH

Desafío:

- Rezago en la información recibida por parte de la DIAN
- Cambio de esquemas impositivos en el país

Objetivo:

- Medir indicadores de desigualdad usando el ingreso disponible en cambio del ingreso corriente

Fuentes:

- Registro de seguridad social– pensionados –PILA
- Ayudas del Gobierno i.e FA; JA; CM
- Encuestas a hogares- GEIH
- Formulario 210 – DIAN
- RELAB

Metodología

Input

- Vincular el registro de pagos a seguridad social con GEIH
- Vincular el registro de ayudas con la GEIH
- Vincular el registro de impuestos con la GEIH

Armonización en unidad estadística y variable de tiempo

- El período de referencia de la encuesta y los RRAA es diferente

Evaluación de la calidad de insumos y salida

- Evaluar las diferencias en cobertura y no respuesta en las diferentes fuentes Validamos la consistencia entre lo que se reportó en la encuesta y en los RRAA por un mismo hogar

Resultados y validación

- Utilizamos el 2019, como periodo de referencia para validar los resultados en un periodo sin estado pandémico, y comprobar si la integración sesga los resultados oficiales.

Cruce DIAN 2019 - GEIH 2022

Comparativo nueva pregunta de la GEIH vs Declarantes DIAN

- El 85,3% de las personas que aparecen en el formulario 210 de la DIAN de 2019 dijeron no haber declarado renta o no pagar impuesto de renta, **por lo cual la pregunta de la GEIH no es informativa.**

Nuevos declarantes

	Declarantes DIAN	
Declarantes GEIH	Si	Missing
Si	13,4%	**0,3%**
No	**85,3%**	75,8%
NS/NR	**0,8%**	0,3%
Menores de 15 años	0,5%	23,6%

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

Cruce DIAN 2019 - GEIH

No todas las personas que hacen match pagan un valor positivo de impuesto de renta, pueden declarar en 0.

DIAN 2019
** personas

Match determinístico

GEIH 2019
756.063 personas

GEIH 2019
41.952 obs.

- Del cruce realizado entre la Gran Encuesta Integrada de Hogares (2019) y el formulario 210 de la DIAN para 2019, se obtiene una cobertura (expandida) del 70% del recaudo de impuestos de renta.

	GEIH Personas expandidas	DIAN 2019 Personas RRAA	Personas match expandido 2019
GEIH 2019	48.910.936	**	**

Fuente: DANE: GEIH 2019 marco 2005. RELAB. DIAN: Formulario 210

Componentes del ingreso disponible

Ingreso de los Empleados

- ✓ Salarios y honorarios
- ✓ Otros ingresos del empleo (Bonos en dinero y en especie, bienes subsidiados o gratis proporcionados por el empleador, indemnización por despido, etc.).
- ✓ Pagos de seguridad social por enfermedad.

Ingreso de capital y propiedades

- ✓ Ingresos por activos financieros (netos de costos), ingresos por activos no financieros (netos de costos) y regalías.
- ✓ Pagos recibidos de fondos individuales voluntarios de pensión y seguros de vida.
- x Ganancias de capital

Ingreso del trabajo independiente

- ✓ Ganancias y pérdidas de empresas no constituidas.
- ✓ Bienes producidos para consumo propio (autoconsumo)
- ✓ Otros.

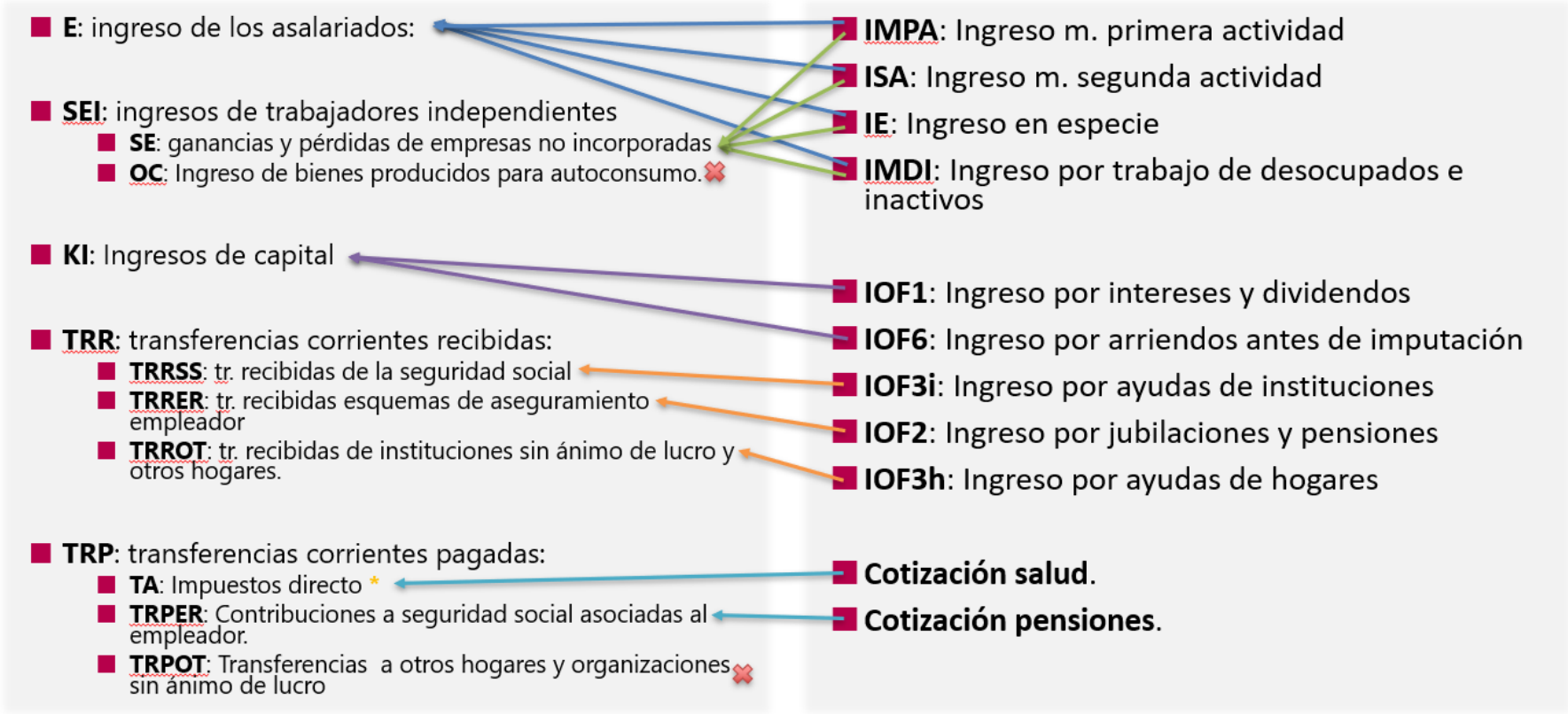
Transferencias corrientes recibidas

- ✓ Transferencias de seguridad social
  - Beneficios de accidente e incapacidad
  - Beneficios por vejez
  - Beneficios de desempleo
  - Licencia de maternidad (paternidad)
  - Subsidios familiares
  - Subsidios de vivienda
  - Otros subsidios
- ✓ Transferencias de esquemas de aseguramiento asociados al empleador
- ✓ Transferencias recibidas de instituciones sin ánimo de lucro y otros hogares.

Transferencias corrientes pagadas

- ✓ Impuestos directos al ingreso y a la propiedad
  - x Impuestos a las ganancias de capital.
- ✓ Contribuciones a seguridad social.
- ✓ Contribuciones a esquemas de aseguramiento social relacionados con el empleo.
- ✓ Transferencias a otros hogares y a organizaciones sin ánimo de lucro.

Generación de las variables



Indicadores de validaciones de la integración

Chequeos después de la integración

- Indicadores de cobertura vía factores de expansión.
- Comparación de distribuciones. Resultante vs original.
- Validaciones haciendo uso de variables cuasi-identificadoras.
- Regresiones para estimación de falsos positivos y falsos negativos.
- Detección de verdaderos positivos vía métodos de clusterización basados en Machine Learning.

Fuente: DANE.

COLOMBIA POTENCIA DE LA VIDA

DANE
DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA
70 AÑOS

Integración de Encuestas con Registros Administrativos

Octubre de 2023

/DANEColombia
@DANE_Colombia
@DANEColombia
/DANEColombia