URUGUAY

INE

Instituto Nacional de Estadística

Documento para uso de los microdatos

2023 Censo URUGUAY

INTRODUCCIÓN

El presente documento tiene como objetivo especificar algunas características de las matrices de datos del Censo 2023, con el objetivo de facilitar el uso de los microdatos.

En particular se especificará las unidades de análisis, universo de estudio, fuente de procedencia de los datos, variables fundamentales, codificación de las variables abiertas, generación de nuevas variables, anonimización y versión de los microdatos.

UNIDAD DE ANÁLISIS Y UNIVERSO DE ESTUDIO

Las unidades de análisis del Censo 2023 son las viviendas, hogares y personas.

El censo registró a los residentes de todo el país, ya sea en hogares particulares, hogares colectivos o situación de calle.

CENSO COMBINADO

El Censo 2023, para la enumeración de las personas, se llevó a cabo mediante dos procedimientos:

Enumeración por cuestionario, llevada a cabo en el operativo de campo, mediante los métodos CAPI, CATI y CAWI[1].

Enumeración administrativa, a partir de registros administrativos de distintas fuentes gubernamentales.

- Enumeración por cuestionario

La enumeración por cuestionario, implicó dos cuestionarios distintos: uno completo y otro básico[2]. El cuestionario Completo fue aplicado a las viviendas particulares ocupadas con moradores presentes y a las siguientes viviendas colectivas: Pensión u otra casa de hospedaje, Residencial de estudiantes, Casa de peones, Internado religioso, Otra colectiva.

El cuestionario básico se aplicó a las personas en situación de calle, al resto de las viviendas colectivas, y a los hogares relevados en la fase de recuperación.

- Enumeración administrativa

Las personas provenientes de registros administrativos cuentan con la información de sexo, edad, si es nacional o extranjero, asistencia actual al sistema educativo, nivel educativo al que asiste actualmente, si es jubilado o pensionista y se incluyen como residentes

[1] CAPI (entrevistas personales asistidas por computadora); CAWI (entrevistas web asistidas por computadora); CATI (entrevistas telefónicas asistidas por computadora).
[2] Se pregunta sobre sexo, edad, migración, educación y fecundidad.

habituales en viviendas censadas cuando la información sobre la composición del hogar es confiable (personas no censadas de hasta 15 años, cuya madre, según el certificado de nacido vivo, ha sido censada). En caso contrario, si se pudo geo referenciar la dirección, entonces se asigna a la zona censal correspondiente o se presenta información geográfica a nivel de departamento, localidad o municipio, y no conforman un hogar.

MATRIZ DE DATOS DISPONIBLES

- Matriz de Viviendas: cada fila representa una vivienda, (particular o colectiva), ocupada o desocupada.

- Matriz de Hogares: cada fila representa un hogar, (particular o colectivo), en una vivienda ocupada con moradores presentes.

- Matriz de Personas: cada fila representa una persona, enumerada por cuestionario censal o de forma administrativa, residente en una vivienda particular, colectiva o situación de calle; o no asociada a una vivienda (una parte de la enumeración administrativa). La matriz de datos de Personas del Censo 2023 se compone de un 89,7% de enumeraciones por cuestionario censal y un 10,3% de enumeraciones administrativas, y contiene 3.499.451 registros. Por lo tanto, la matriz de personas contiene la cantidad de población estimada para el año 2023.

- Matriz Ampliada: cada fila representa una persona, enumerada por cuestionario censal o de forma administrativa, residente en una vivienda particular, colectiva o situación de calle. Donde se le asocia la información relacionada a la vivienda y el hogar que integra.

  Aquellas personas provenientes de la enumeración administrativa, que no se los pudo asociar a un hogar, no contarán con información en las variables de vivienda y hogar.

INFORMACIÓN SOBRE ALGUNAS VARIABLES

- Fuente_Ext

Permite identificar, en la matriz de personas, la fuente de la enumeración: enumeración por cuestionario o enumeración administrativa.

- Universo

Permite identificar el componente del universo al cual pertenecen las personas. Esta variable asume cuatro valores: Particulares; Colectivos; Situación de Calle; Desconocido.

Desconocido: refiere a la población censada de forma administrativa a la que no se pudo asociar a un hogar.

- CUESTIONARIO_COMPLETO

Permite identificar a aquellas personas que se les aplicó el cuestionario completo.

- CUESTIONARIO_BÁSICO

Permite identificar a aquellas personas que se les aplicó el cuestionario básico.

- DIRECCIÓN_ID

Es la variable clave, en la matriz de vivienda, que vuelve única cada dirección.

- VIVID

Código correlativo de identificación de la vivienda dentro de la dirección. Vale siempre 1, excepto cuando en la dirección coexisten una vivienda particular con una vivienda colectiva.

- HOGID

Código correlativo de identificación del hogar dentro de la vivienda

- PERID

Código correlativo de identificación de la persona dentro del hogar.

El PERID solo aplica a las personas censadas mediante cuestionario censal y al conjunto de personas censadas por enumeración administrativa que se pudo asociar a un hogar.

- AGREGADO_A_HOGAR_CENSADO

Esta variable permite identificar las personas censadas por enumeración administrativa que se logró asociar a un hogar censado.

- MADRE_PERID

Para el conjunto de personas censadas de forma administrativa que se logró asociar a un hogar, la variable MADRE_PERID indica el número de persona que corresponde a la madre[3].

- ID_CENSO

Código correlativo de identificación de la persona censada por cuestionario censal o de forma administrativa.

[3] Las personas provenientes de registros administrativos integradas a un hogar censado, son personas de hasta 15 años cuya madre, según el certificado de nacido vivo, ha sido censada.

OCUPACIÓN

- Ocupación

Se disponibiliza la primera versión de la codificación de la variable PERAL06: (¿Qué tareas realiza en ese trabajo?, pregunta abierta que se aplica a los censados por cuestionario que cuentan con trabajo), se realizó de acuerdo al Clasificador Uniforme de Ocupaciones (CIOU-08)[4], de forma automática y se presenta a un dígito.

Para la codificación automática, se utilizó un algoritmo del tipo SVM (Support Vector Machine – algoritmo de aprendizaje supervisado). El entrenamiento y validación del algoritmo se realizó utilizando los datos del Censo 2011, obteniendo un precisión global del 79 % y logrando una codificación del 82% de los casos del Censo 2023. Se continúa trabajando para aumentar la cobertura de codificación, por lo tanto, a futuro se disponibilizará una nueva versión de la variable.

GENERACIÓN DE NUEVAS VARIABLES

La matriz de datos de personas incorpora nuevas variables, con información para el total de la población estimada, excepto la población en situación de calle (3.495.947 personas), con el objetivo de enriquecer los datos del Censo 2023.

Las variables que se agregan a la matriz son las siguientes:

- ASISTENCIA (Asistencia actual a un centro de enseñanza)
- NIVELEDU_ACT (Nivel educativo de la población que asiste actualmente a un centro educativo)
- JUB_PEN (jubilado o pensionista)
- EXTRANJERO (Nacido en el exterior)

Para la generación de las variables ASISTENCIA (Asistencia actual a un centro de enseñanza),
NIVELEDU_ACT (Nivel educativo de la población que asiste actualmente a un centro educativo), y JUB_PEN (jubilado o pensionista), se utilizaron tres estrategias:

- Información proveniente de registros administrativos (para la población censada de forma administrativa[5]).
- Información proveniente del cuestionario censal (para la población censada mediante cuestionario censal).
- Imputación (para la población censada mediante cuestionario o de forma administrativa, cuando no existe información en la fuente original).

---

[4] https://www.gub.uy/instituto-nacional-estadistica/comunicacion/publicaciones/clasificador-internacional-uniforme-ocupaciones-rev-4

[5] Se realizaron ejercicios de validez de la información proveniente de registros, comparando las respuestas de la población censada por cuestionario, en las variables de interés, con su información en registros: obteniendo coincidencias por encima del 90%.

La imputación estadística de las variables seleccionadas se realizó utilizando el método de imputación del vecino más cercano (k-Nearest Neighbors Imputation, KNN Imputation). Este método permite reemplazar los valores faltantes en un conjunto de datos utilizando valores de observaciones similares.

El método consiste en tres pasos: primero se calcula la similitud entre las observaciones con dato faltante y sin dato faltante utilizando una métrica de distancia basada en las variables disponibles. Luego para cada observación con valores faltantes, se identifica la observación más similar (vecino más cercano) según la distancia calculada. Una vez emparejados los datos, el valor ausente se reemplaza por el valor donado por el vecino más cercano.

Para el censo 2023 el proceso se llevó a cabo en etapas, donde en cada una se definió un grupo de variables disponibles y otro de variables a imputar. Para mantener la coherencia en la relación entre las variables, aquellas pertenecientes al mismo conjunto fueron imputadas de manera conjunta, utilizando el mismo donante.

Variables como el sexo, la edad y la localización geográfica tienen un peso determinante en la elección del vecino más cercano, por lo cual fueron utilizadas en todas las etapas de la imputación.

Finalmente, para cada variable agregada se realizó un testeo de similitud con el objetivo de descartar cambios estadísticamente significativos con la distribución de la variable relevada mediante cuestionario censal, a partir de la incorporación de la población enumerada administrativamente, junto con la imputación de valores.

Las variable EXTRANJERO (Nacido en el exterior), se conforma con información proveniente de registros administrativos (para la población censada de forma administrativa), y con información proveniente del cuestionario censal (para la población censada mediante cuestionario censal). No se llevó a cabo ningún proceso de imputación estadística[6].

La variable MUNICIPIO_PAIS identifica a las personas, hogares y viviendas contenidas en cada uno de los 125 municipios existentes en el país, en el área urbana y rural[7].

El procedimiento de asociación de las personas a cada municipio se realizó mediante tres procedimientos:

- Para las observaciones del área urbana que cuentan con información sobre la zona censal, se recurre a la tabla de correspondencia zona censal - municipio.
- Para las observaciones del área rural, que cuentan con las coordenadas de geolocalización (X;Y), se llevó a cabo un geoproceso a los efectos de identificar los puntos rurales que caen dentro de los polígonos de los distintos municipios[8].

[6] El ejercicio de validez de la información proveniente de registros, comparando las respuestas de la población censada por cuestionario, (lugar de nacimiento), con su información en registros: obteniendo coincidencias por encima del 99,3%.

[7] La capa de municipios sobre la que se reprocesaron los datos censales 2023 fue generada en el marco del Grupo de Trabajo de Límites Administrativos (GTLA), liderado por IDE e integrado por Corte Electoral, OPP, IGM, DINOT e INE. La capa contempla 125 municipios y surge de la capa de series electorales 2024 aprobada por la Corte Electoral y Circular 10889

[8] Los límites de los polígonos que conforman cada municipio, en las áreas rurales, no coinciden con las unidades geoestadísticas.

- Para las personas que no cuentan con información geográfica a nivel de zona censal para el área urbana y coordenadas para el área rural[9], se llevó a cabo un proceso de imputación.

La imputación se realizó sobre 63 mil personas, y se utilizó el método estadístico del vecino más cercano con las variables sexo, edad, departamento, localidad, asiste a un centro educativo, nivel educativo que cursa, cantidad de hijos nacidos vivos, año de nacimiento del último hijo, año de nacimiento del primer hijo, trabaja actualmente y es jubilado o pensionista.

Se controló que las personas sean imputadas dentro de municipios pertenecientes a su localidad y/o departamento, y se mantengan las proporciones originales de ciertas variables estructurales (como sexo, edad, nivel educativo y condición de ocupación).

Luego de la imputación se realizó un testeo de similitud para determinar si la imputación de la variable MUNICIPIO_PAIS, produjo un cambio estadísticamente significativo en la distribución de los variable antes y después de la imputación. El resultado de la prueba Chi cuadrado fue de $\chi 2$= 15.472 con 15.252 grados de libertad (df). El valor-p obtenido fue 0.10, lo que indica que las distribuciones en las categorías son similares desde un punto de vista estadístico.

Para cada una de las variables mencionadas en esta sección, es posible identificar la procedencia del dato (cuestionario censal, registros administrativos o imputación), mediante las variables que comienzan con la expresión TIPO.

MICRODATOS ANONIMIZADOS

El Secreto Estadístico garantiza que la información recopilada por el Instituto Nacional de Estadística será difundida preservando el anonimato de la fuente. La Ley 16.616 obliga a tratar los datos individuales proporcionados por la fuente de información con la más absoluta confidencialidad.

En ningún caso ni circunstancia, es posible acceder a información estadística con un grado de desagregación tal que se vulnere el secreto estadístico y se individualice e identifique a la fuente de información.

El archivo de microdatos del Censo 2023, divulgados, fue sometido a un proceso de disociación de datos personales, muchas veces llamado anonimización, siendo la finalidad de dicho proceso impedir que, a partir de una información o de una combinación de informaciones, se logren identificar individuos.

[9] Mayoritariamente casos censados de forma administrativa, que pueden estar a nivel de departamentos o localidad sin zona censal.

Los métodos de protección de los datos que se han utilizado son básicamente de agregación, recodificación de datos y supresión de algunas variables. No se han utilizado perturbaciones de los valores.

Considerando lo antes expuesto, se han aplicado las siguientes restricciones a los microdatos a difundir:

- Se eliminan las variables referidas a las unidades geoestadísticas (sección, segmento, zona) y barrio.
- Las localidades con menos de 100 habitantes se presentan agrupadas junto a las áreas rurales.
- Se eliminan las variables sobre tenencia de documentos, identidad de género, mes y año de nacimiento, área de estudio.
- La edad se presenta en dos variables: agrupada en tramos quinquenales para toda la población, y abierta en edades simples para las personas residentes en localidades de 10 mil o más habitantes.
- Las variables referidas a la fecha de nacimiento del primer y último hijo se presenta agrupada en tramos quinquenales (del año de nacimiento).
- La variable lugar de nacimiento se presenta agrupada en tres categorías (en este departamento, en otro departamento, en otro país). Se eliminan las variables referidas a localidad.
- En las preguntas sobre lugar de residencia anterior se eliminan las variables referidas a localidad.
- La variable lugar de estudio y lugar de trabajo se presenta agrupada en tres categorías (en este departamento, en otro departamento, en otro país). Se eliminan las variables referidas a localidad.

VERSIÓN 02_2025 DE LOS MICRODATOS

Por último, se disponibiliza la primera versión de los microdatos anonimizados, denominados VERSIÓN 26_02_2025.

A medida que se continúen enriqueciendo los microdatos del Censo 2023, se divulgarán y anunciarán nuevas versiones.