La experiencia del Censo Combinado 2023 de Uruguay

13 de marzo de 2025

Unidad Registros Administrativos
rraa@ine.gub.uy

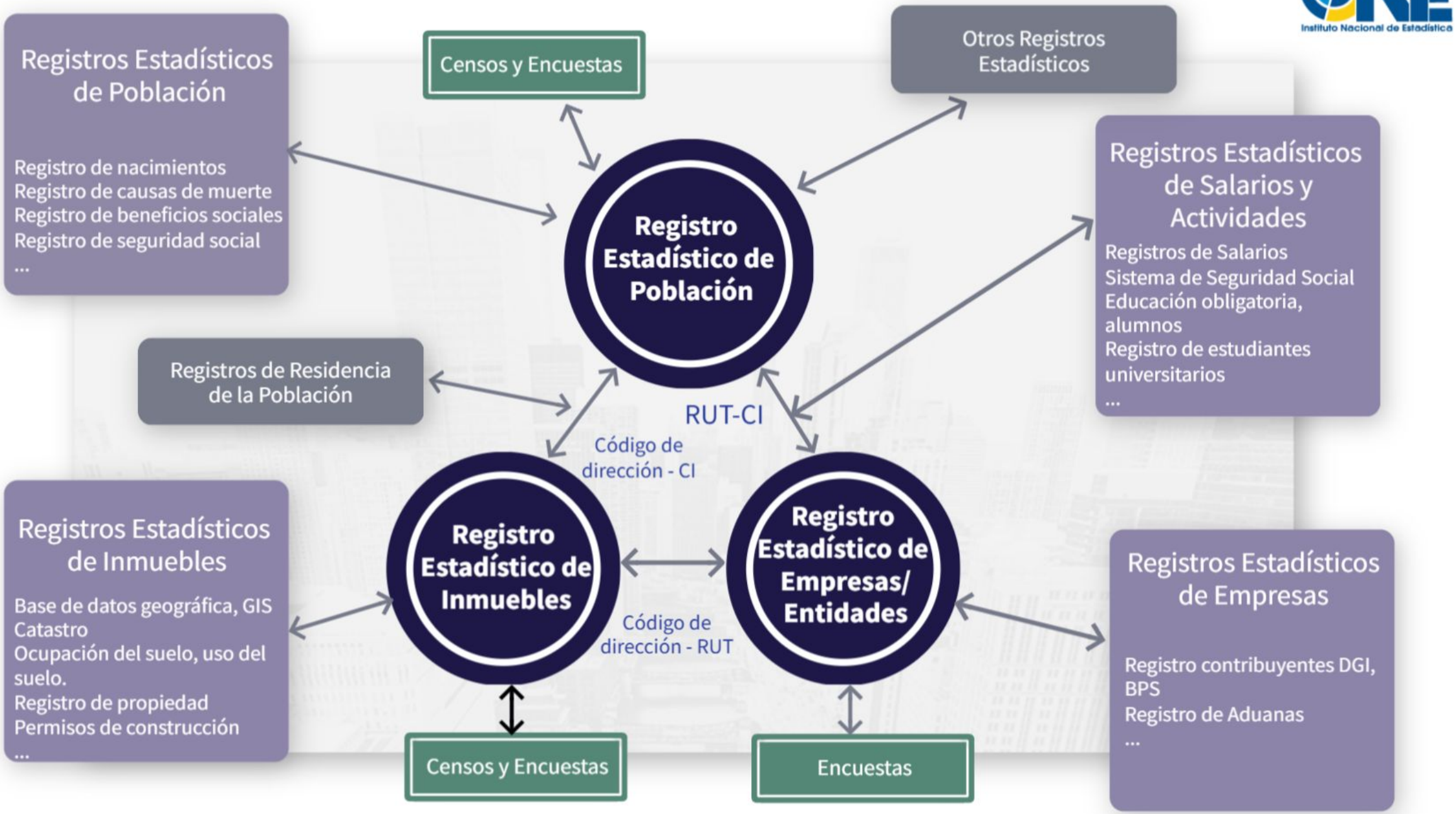
Uso de RRAA en el Censo 2023 para complementar la enumeración basada en cuestionarios

Federico Segui
fsegui@ine.gub.uy

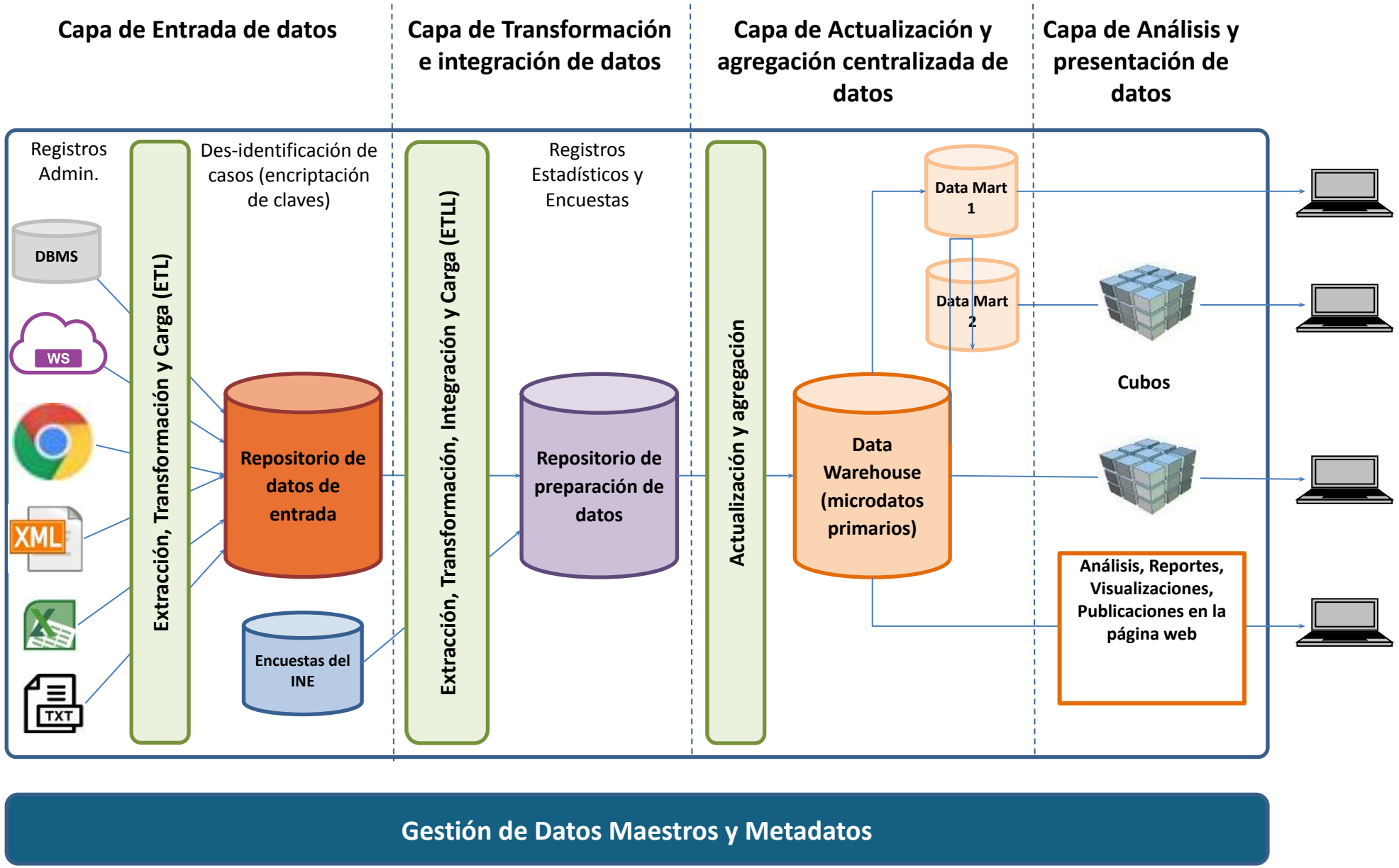
Presentación - Piloto de Censo basado en registros administrativos 2023

Metodología del Censo Combinado 2023

Sistema Integrado de Registros Estadísticos y Encuestas - SIREE



Arquitectura del Data Warehouse Geo-Estadístico



Evaluación externa del piloto de censo basado en registros (Stat Norway, UNSD)

- **The United Nations Statistics Division (UNSD) commends the Instituto Nacional de Estadística (INE) of Uruguay for its groundbreaking work in conducting a pilot register-based census alongside the traditional 2023 population and housing census.** This innovative approach, coupled with Uruguay's well-developed system of administrative registers, holds potential for enhancing the accuracy and timeliness of population statistics. The pilot allows for valuable comparisons with the traditional method, further solidifying the foundation for a forward-looking census system.

Taking advantage of the availability of data from administrative registers (whose coverage has been deemed good by the comparison study), INE should consider using the data from the registers to validate data from the traditional census and vice versa. Validation exercises can help identify discrepancies between census and administrative data. Specifically, the validation exercise could be used impute missing data (data from administrative registers can be used to fill in missing information for non-responding households or individuals in the traditional census) and improve coverage (by comparing register data with census data, potential undercounts in the census can be identified and potentially rectified).

Microdatos del censo 2011 y anteriores

Cuestionarios censales

RRAA viv. col. y otros casos

Imputación de unidades (personas)

No se hacía nada

Moradores ausentes / Rechazos / No respuesta

Omisión propiamente dicha

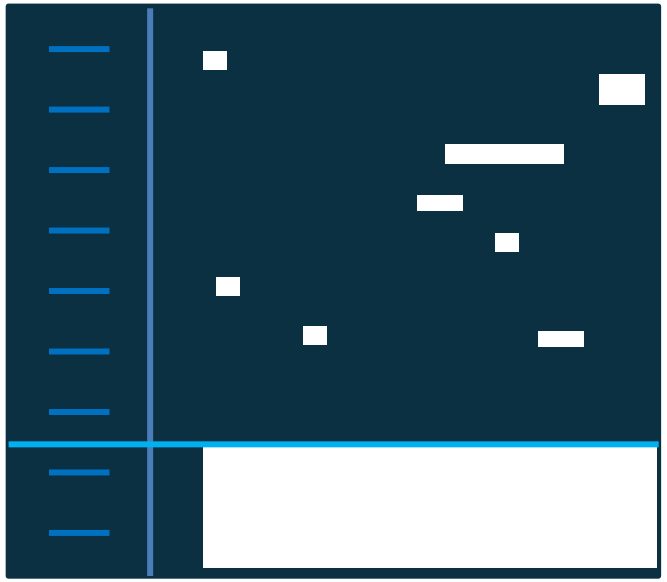
Censo Combinado 2023: Fuentes de datos separadas

Cuestionarios censales

Personas

Cuestionario censal

RRAA viv. col. y otros casos



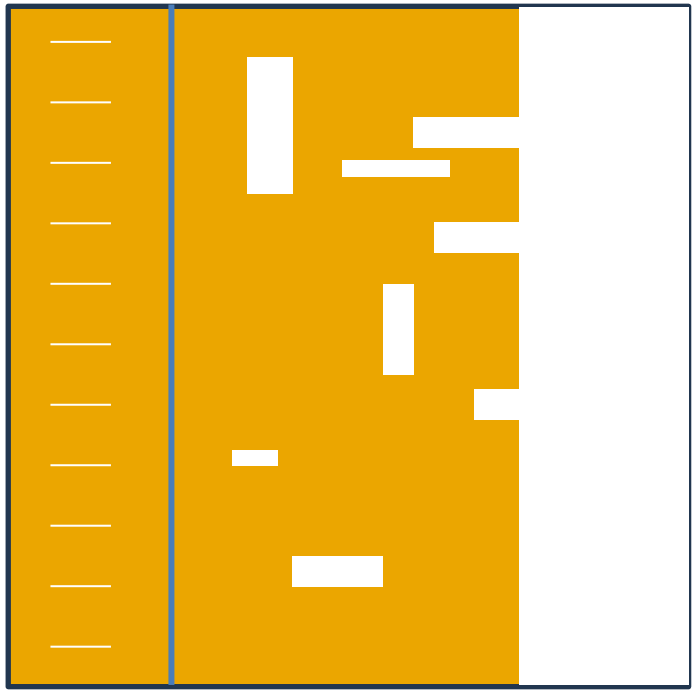
Edad, sexo, lugar de residencia

Otras variables

Características

Registros Administrativos

Personas



Edad, sexo, lugar de residencia

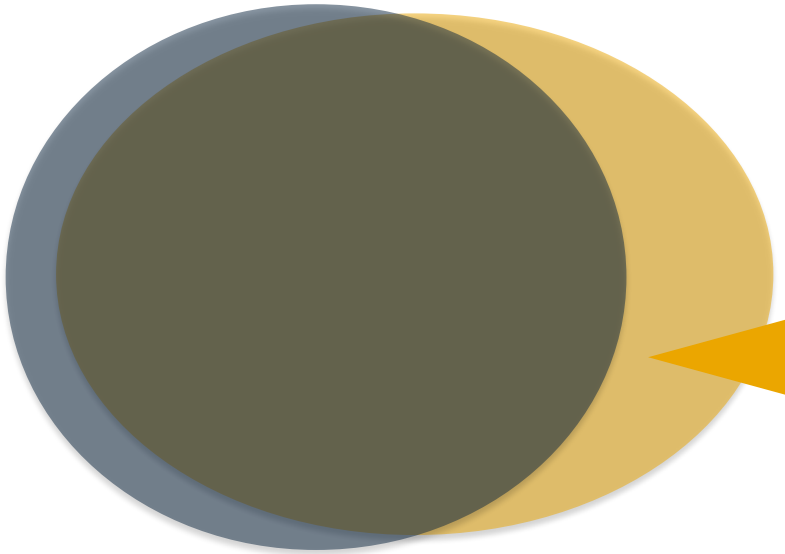
Otras variables

Características

URUGUAY INE Instituto Nacional de Estadística

Combinando cuestionarios censales y registros administrativos

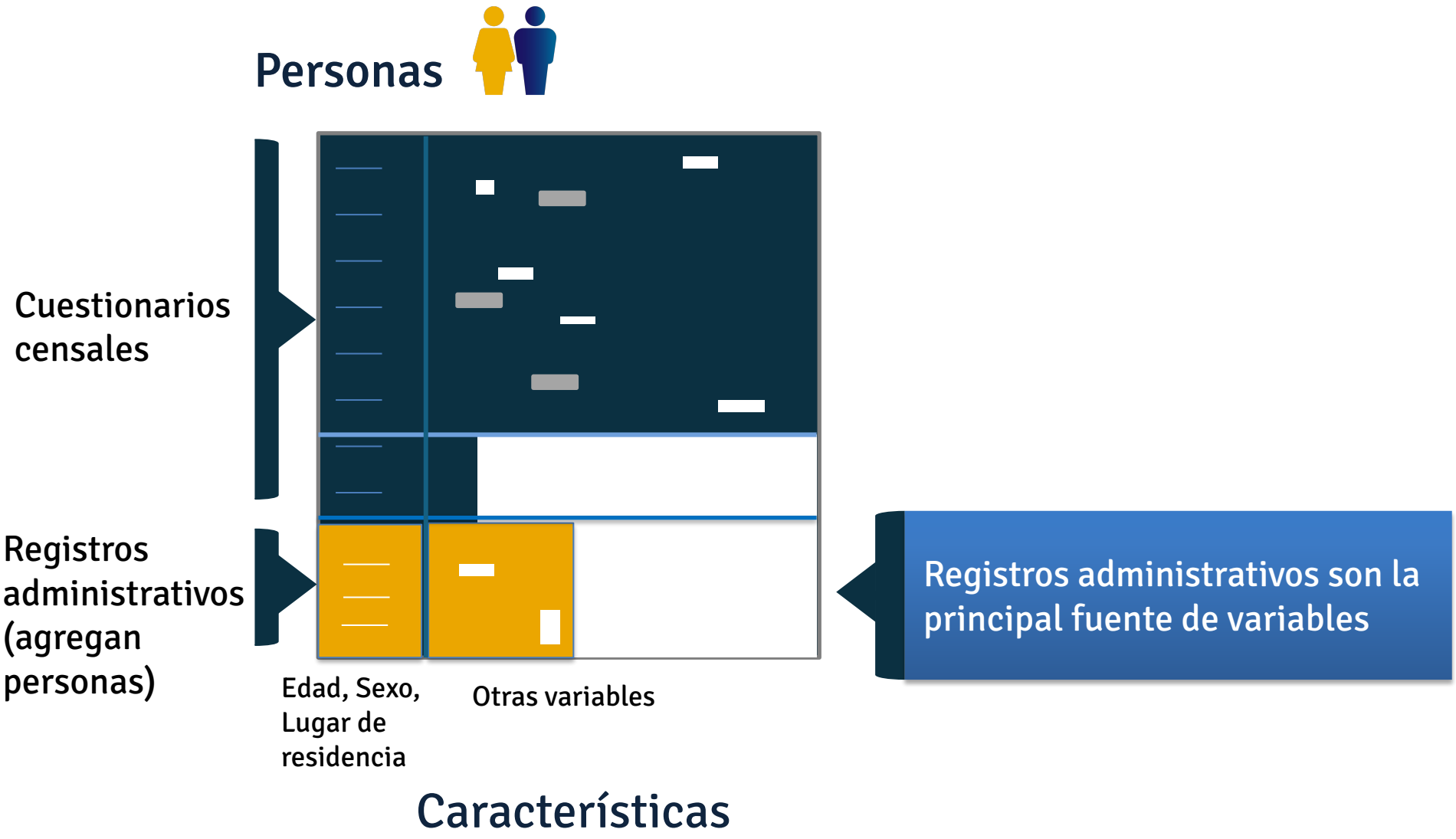
Cuestionarios censales



Registros de población

Registros administrativos que se agregan a los microdatos del censo para contar a las personas no enumeradas por cuestionario

Microdatos del censo 2023: Combinación de fuentes



Microdatos del Censo Combinado 2023

Conteo de población

Características

Creación del Registro Estadístico de Población Residente (REPoR)

hub_de_personas
6.000.000

→ Ubicación geográfica (dirección de residencia)

- Servicio de electricidad
- Servicios de salud
- SIIAS (protección social)

- Nivel de dirección (70%)
- Nivel de calle
- Municipio
- Ciudad
- Departamento
- País

→ Registro Estadístico de Lugar de Residencia (administrativa)

→ Estatus de las personas

- Nacimientos/ defunciones
- Migración (E/S)
- Señales de Vida

1. Residente
2. Emigrante
3. Fallecido
4. Sin señales de vida

→ Registro Estadístico de Población Residente

3,5 millones

Paso 0. Actualización y Revisión del REPoR

1. Actualización del *Registro Estadístico de Población Residente* (*REPoR*) con Señales de Vida (SDV). Fecha censal 31/5/23
   a. SIIAS (+60 RRAA). (Sistema de Información Integrada del Área Social – MIDES).
   b. Salud: RUCAF, Vacunados Covid y Covid+
   c. Educación: ANEP, UDELAR.
   d. Activos, jubilados y pensionistas BPS.
   e. Servicios activos UTE.
   f. E/S migraciones (más de 180 días en UY, en los últimos 12 meses).
   g. Otras fuentes

Paso 1. Vinculación Censo-HUB_Personas

Enumeración Censal

(3.138.800)

Matching determinístico por CI y se verifica otras variables (Nombres, Fec.Nac.)
+
Matching probabilístico en el resto de los casos.

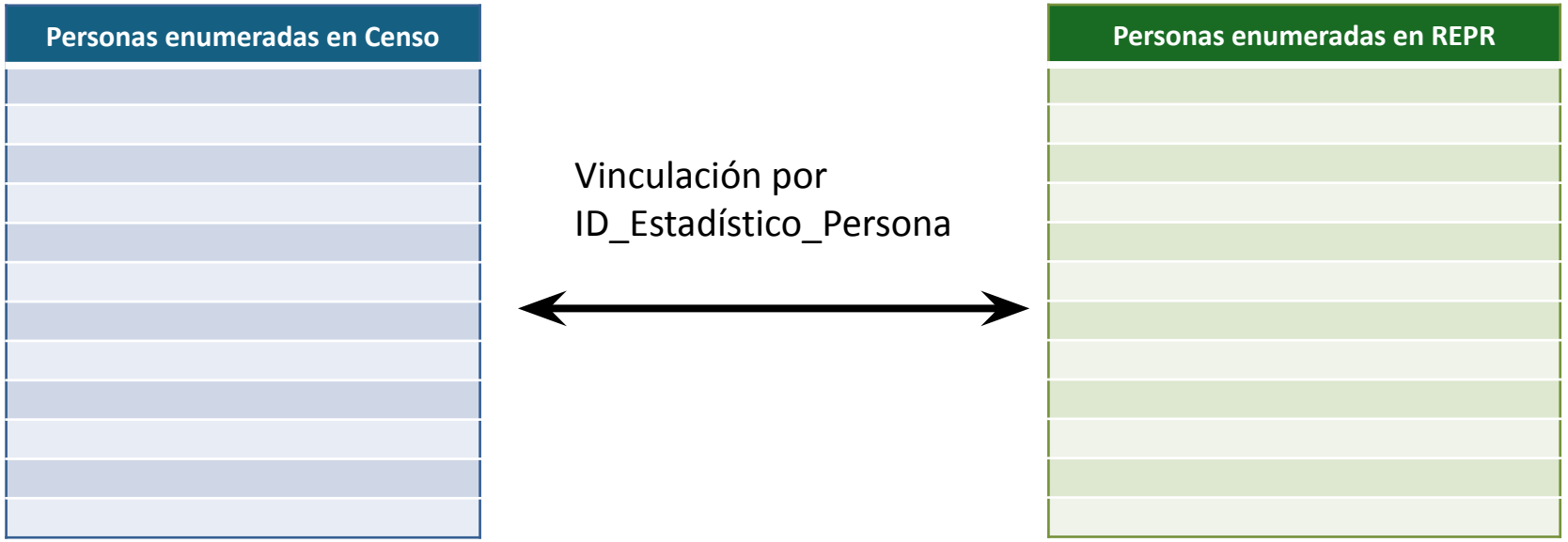
Casos con CI únicas 2.854.168
99,55% matching x CI

Se “pega” ID_Estadístico_Persona y Fec. Nac.

HUB-Personas

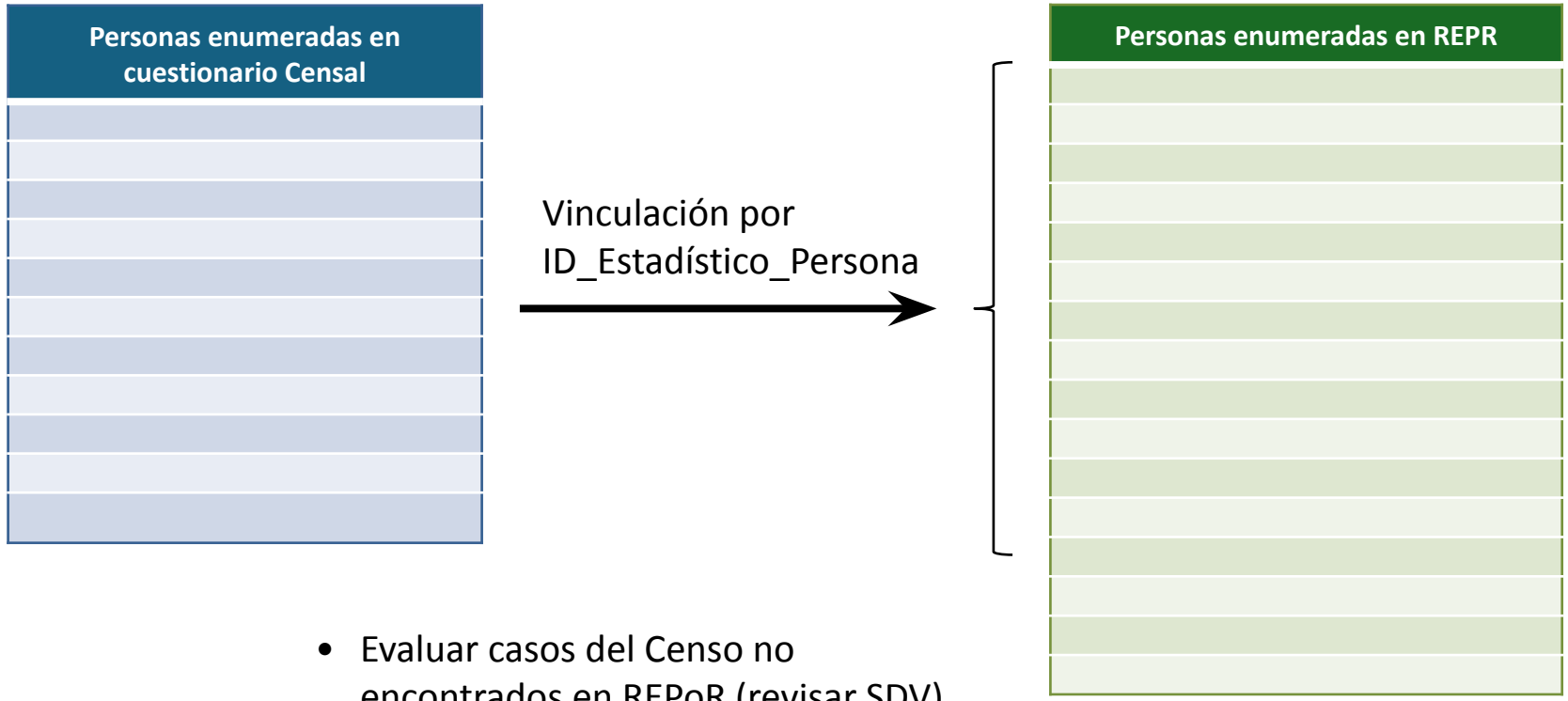
(Tabla maestra de documentos de identidad, 6.000.000 de casos)

Paso 2. Comparar direcciones Censo-REPoR



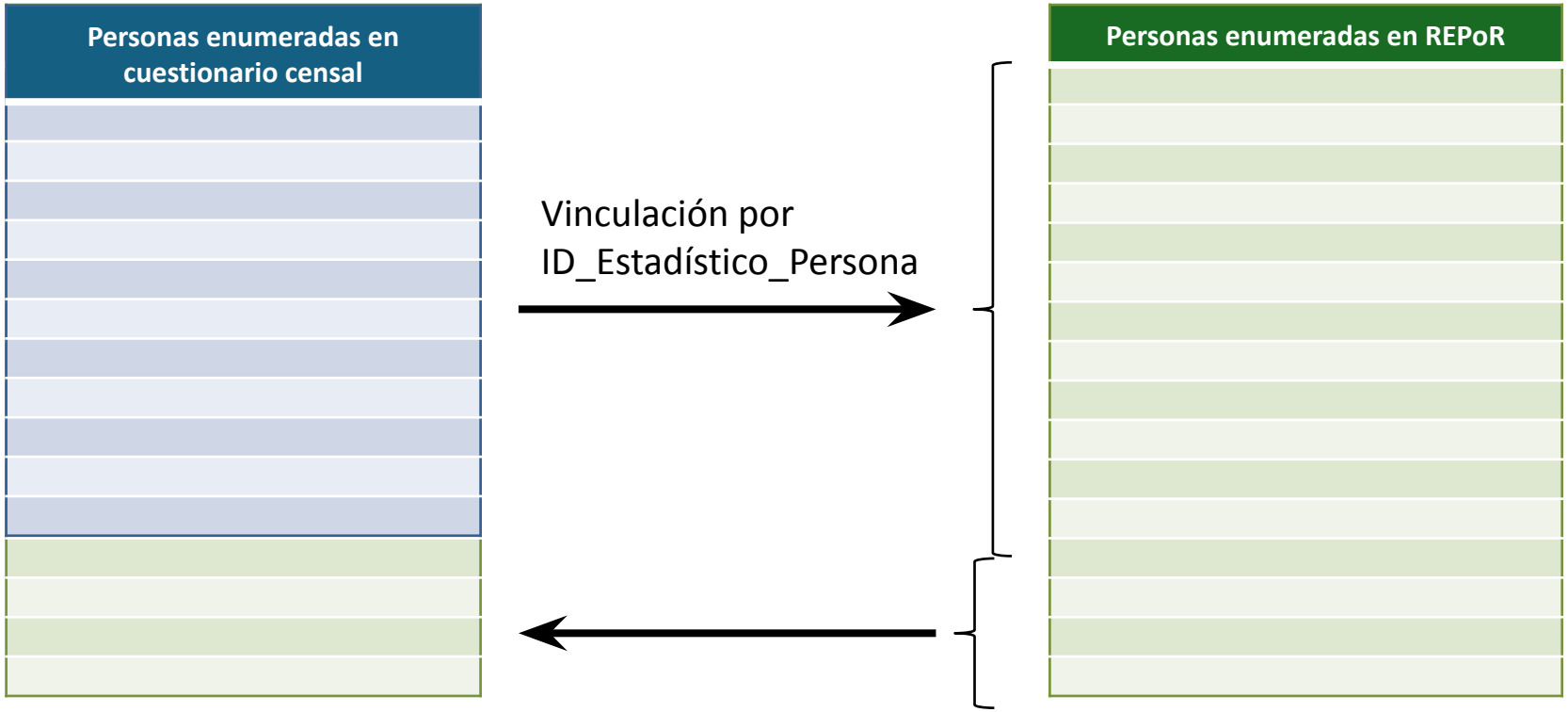
- Vincular cada persona y verificar si la dirección coincide.
- Si coincide, asignar un “peso” mayor a la fuente de esa dirección en el REPoR.
- Una vez re-ponderadas las fuentes de direcciones, volver a ejecutar la asignación de direcciones del REPoR con base en esta re-ponderación.
- Volver a comparar direcciones con Censo.

Paso 3. Vinculación Censo-REPoR



- Evaluar casos del Censo no encontrados en REPoR (revisar SDV)
- Los casos no vinculados del REPoR con Censo son los casos de “enumeración administrativa” a agregar a los microdatos finales del censo (360.000 aprox.)

Paso 4. Agregar personas del REPR al Censo



- Los casos no vinculados del REPoR con Censo son los casos de “enumeración administrativa” que se agregan a los microdatos del censo (360.000 aprox.)

HUB-Personas
6.000.000

SDV

Residentes	1
Fallecidos	2
Baja lógica (más de 112/105 años)	3
Emigrantes	4
Sin SDV	5

REPoR-31/5/2023

Id_Persona	Indice de Residencia
321	1
465	0,90
742	0,85
784	0,76
801	0,66
875	0,60
892	0,55
331	0
521	0
211	0,33

Pob. estimada
3.499.451

Paso 4.1 Agregar personas del REPoR al Censo

- Menores de 14 años no censados, se agregan al hogar del censo donde aparece censada la madre. Se utiliza el CNV para obtener la CI del menor y su madre.
- Primero, se asignan a direcciones del censo con estado “no respuesta”.
- El resto, se asignan a nivel de áreas geográficas (zonas, municipios o localidades).

Registro de población: Hub personas y vinculación de documentos

Antonella Vignolo
avignolo@ine.gub.uy

SIREE: Sistema Integrado de Registros Estadísticos y Encuestas

Registros estadísticos de población:
- Nacimientos
- Fallecimientos
- Migración
- Seguridad social

Registro de lugar de residencia de la población

Registros estadísticos de salarios y actividades:
- Cotizantes
- Sistema de seguridad social
- Educación
- Profesionales

Registros estadísticos de empresa:
- Contribuyentes DGI
- Banco de Previsión Social
- Registro de aduanas

Registros estadísticos de inmuebles:
- Sistema Único de Direcciones
- Catastro
- Permisos de construcción
- Servicios públicos

Registro base de población

Registro base de inmuebles

Registro base de empresas

Id ubicación- Id persona

Dirección - Documento

Id persona - Id empresa

Documento - RUT

Dirección - RUT

Id ubicación - Id empresa

Registro Base de Población

El registro base de población pretende registrar a todas las personas nacidas o que residen permanente o temporalmente en el país.

Cada persona debe tener asignado un Id Estadístico Persona, para esto es necesario vincular su/s documento/s de identidad a un mismo Id en la tabla ***hub_personas***.

id_estadistico_persona	id_pais_documento	id_tipo_documento	documento	id_fuente	id_estado_dato
123456789	27	3	XXXXXXX	25	1
123456789	553	2	YYYYYYY	27	1
123456789	1	1	WWWWW	9	2

Proceso de creación del hub personas

Primer hub_personas - SIIAS

Problemas de calidad en docs.

Envío masivo DNIC

Creación hub_personas_dnic

Segundo hub_personas

Actualización

Creación del Hub Personas 2.0.

R1

+

R2

+

R3

+

R4

- Limpieza
- Eliminación de duplicados
- Selección de registros según nulos

Tabla maestra de docs.

SPLINK Data linkage at scale

Creación de clusters

Vinculación determinística

nombres + apellidos + nacimiento

documento

Vinculación probabilística

nombres + apellidos + nacimiento

Hub personas

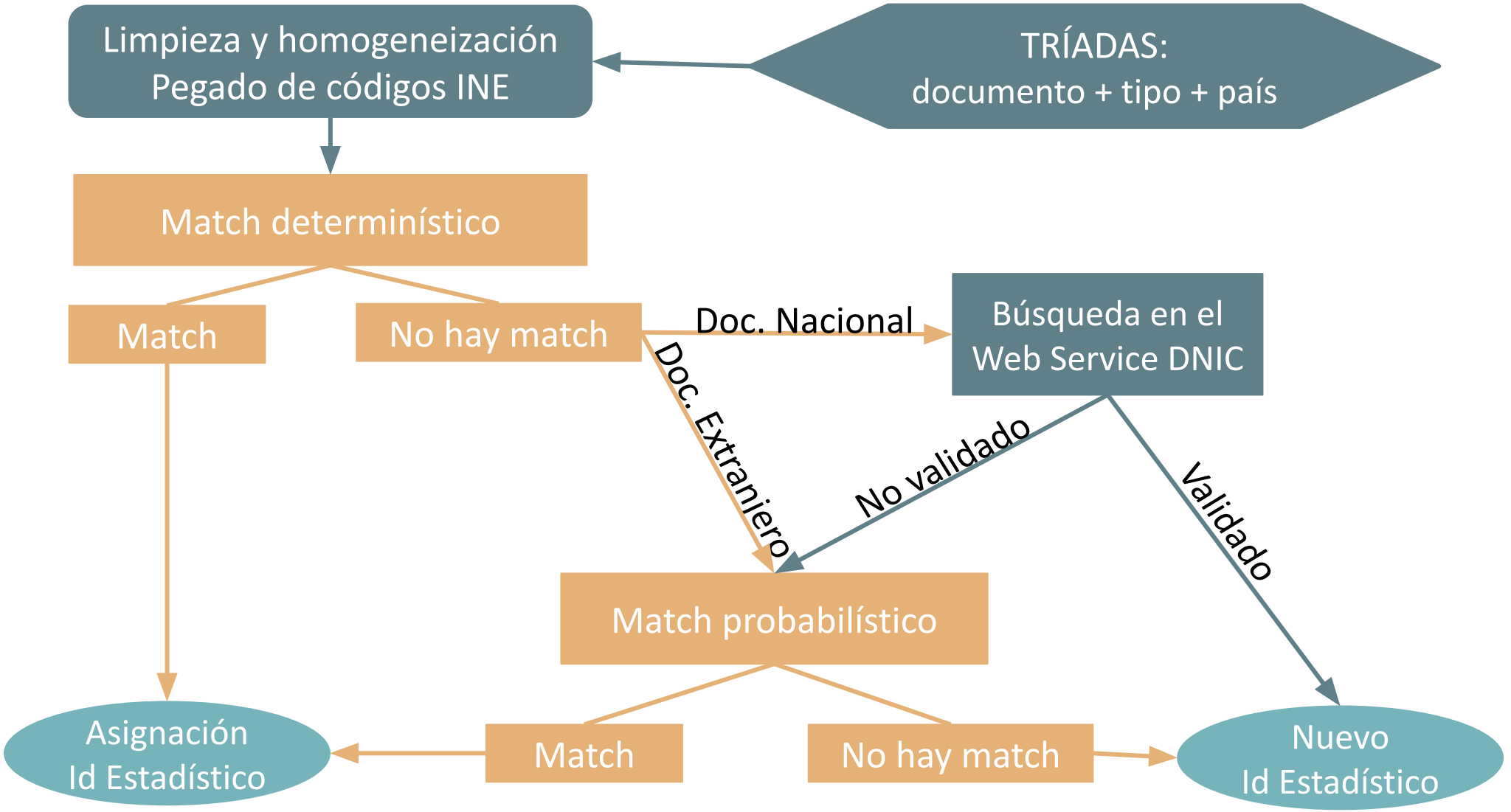
SPLINK Data linkage at scale

Corrección
Falsos Negativos

Descartan
Falsos Positivos

URUGUAY INE Instituto Nacional de Estadística

Proceso de asignación del Id Estadístico



Principales fuentes de información

- Dirección Nacional de Identificación Civil (DNIC)
- Ministerio de Salud Pública
  - Certificado de nacido vivo
  - Certificado de defunción
  - Cobertura de salud
- Ministerio de Desarrollo Social
  - Sistema Integrado del Área Social
  - Prestaciones sociales

Principales esquemas del registro de población

- Actividad laboral
- Educación
- Seguridad Social
- Migración
- Otros registros de población
  - Vacunados COVID-19
  - Ex privados de libertad
  - Personas en viviendas colectivas

Herramientas utilizadas

Data Integration

Apache Superset™

Migraciones, estatus migratorio

Esteban Cardoso
ecardoso@ine.gub.uy

OBJETIVO

El objetivo principal del registro estadístico migratorio es identificar la situación migratoria de las personas relacionadas con Uruguay. Como objetivo adicional, el registro pretende generar estadísticas migratorias relevantes, tales como el cálculo de inmigrantes y emigrantes por año, así como otras variables asociadas al perfil del migrante.

FUENTES DE INFORMACIÓN

Principales fuentes de datos:

- Entradas y salidas, Dirección Nacional de Migración (registro de todos los movimientos de entrada y salida del país), dependencia del Ministerio del Interior.
- Fuente reciente: Emigrantes pensionistas en el extranjero (14.000).

- Solicitudes de residencia y residencias concedidas:

Ministerio de Relaciones Exteriores.
Dirección Nacional de Migración, Ministerio del Interior (a partir de 2023 única fuente de residencias).

URUGUAY INE Instituto Nacional de Estadística

VARIABLES:

Variables relevantes de entradas y salidas:

- Nombres y apellidos.
- Tipo de movimiento (entrada o salida).
- Fecha del movimiento.
- Puesto migratorio.
- Medio de transporte.
- Nacionalidad

Conceptos y metodología, un proceso abierto y de aprendizaje continuo..

Pregunta central:

¿Cómo definir conceptual y metodológicamente un inmigrante/emigrante utilizando R.R.A.A.?
¿Utilizar o no un criterio temporal? ¿Cuál?

Reuniones con países que cuentan con un padrón demográfico por el cual los ciudadanos tienen incentivos a declarar su residencia dado que muchos servicios están asociados a su empadronamiento.

Imputación del estatus mediante un criterio temporal

Dos cortes temporales para cada año. Fecha de inicio / fecha de fin.
(la fecha puede estar sujetos a cambios, por ejemplo probar 1 de junio año n / 1 de junio año n + 1)

Persona 1.

S = Salida.
E = Entrada.

Se suman los tramos positivos y se suman los tramos negativos, se pasan a valor absoluto y si los tramos positivos son mayores que los negativos entonces la persona será un **residente**, de lo contrario será un **no residente**.

Persona 2. Agregando un año más.

PROCESO

1) Tabla con todos los movimientos migratorios.

2) Vinculación de los n documentos de una persona.

3) Tabla que contabiliza todos los días de entrada y salida del país de cada persona.

4) Rellenar el resto de los años según los movimientos existentes.

5) Calcular las personas en cada categoría migratoria.

Conceptos principales

- Las categorías utilizadas se generan mediante una imputación realizada a partir de Registros Administrativos teniendo el tiempo de estadía en el país como una dimensión fundamental para su definición.

- Emigrantes: Aquellas personas con documento uruguayo que en un año calendario estuvieron más de 6 meses días fuera del país.
- Residentes inmigrantes: Aquellos extranjeros que en un año calendario permanecieron 6 meses o más en nuestro país.
- Residentes uruguayos: Aquellas personas que habiendo nacido en el país permanecieron 6 meses o más en él en un año calendario.

Paso 2: Tabla generada: Fact_migraciones

entradas_salidas → Asignación del id_estadístico → **Fact_migraciones:**
procedencia_destino
Tipomov
id_pais_residencia
id_tiempo

id_estadistico_persona	procedencia_destino	tipomov	inspectoria	puesto	clasemigratoria	id_tiempo
7.015.640	GUALEGUAYCHU	SALIDA	9	21	1	20.120.331.034.600
7.015.640	GUALEGUAYCHU	ENTRADA	9	21	1	20.120.403.223.600
7.015.640	SANTIAGO DE CHILE	SALIDA	2	8	1	20.130.324.115.500
7.015.640	[NULL]	ENTRADA	2	8	1	20.130.402.111.000
7.015.640	[NULL]	SALIDA	8	20	1	20.150.328.214.600
7.015.640	[NULL]	ENTRADA	8	20	1	20.150.404.231.000
7.015.640	BUENOS AIRES	SALIDA	7	17	1	20.190.720.100.000
7.015.640	BUENOS AIRES	ENTRADA	7	17	1	20.190.722.185.500

URUGUAY INE Instituto Nacional de Estadística

Paso 3 y 4: Tabla generada: Fact_aux_residencias

Fact_migraciones

Cálculo de los días dentro y fuera del país de cada persona. Se suma una columna 'weird' que indica null si no hay un itinerario incoherente y 365 si lo hay.

fact_aux_residencias:
In
Out
Weird
año

id_estadistico_persona	days_dif_sum_in	days_dif_sum_out	days_dif_sum_weird	last_tipomov	first_tipomov	anio
7.015.640	362	3	[NULL]	ENTRADA	SALIDA	2.012
7.015.640	355	8	[NULL]	ENTRADA	SALIDA	2.013
7.015.640	355	8	[NULL]	[NULL]	[NULL]	2.014
7.015.640	357	7	[NULL]	ENTRADA	SALIDA	2.015
7.015.640	357	7	[NULL]	[NULL]	[NULL]	2.016
7.015.640	357	7	[NULL]	[NULL]	[NULL]	2.017
7.015.640	357	7	[NULL]	[NULL]	[NULL]	2.018
7.015.640	362	2	[NULL]	ENTRADA	SALIDA	2.019
7.015.640	365	0	[NULL]	[NULL]	[NULL]	2.020
7.015.640	365	0	[NULL]	[NULL]	[NULL]	2.021
7.015.640	365	0	[NULL]	[NULL]	[NULL]	2.022

Limitaciones

- Sabemos que existe un sub-registro en el control de las entradas y salidas, lo que genera que tengamos personas con itinerarios de viaje incoherentes que provoca dificultades a la hora de imputar la cantidad de días de estadía.

- Los puestos migratorios que no tienen sistemas de registro biométricos (todos excepto el Aeropuerto Internacional de Carrasco. La situación cambió para fines de 2024, la DNM impulsó un nuevo sistema de registro digital) nos han generado problemas ya que un error en el registro del documento o la fecha de nacimiento provoca errores en la asignación del id estadístico.

- Al realizar la imputación de estadía en un año calendario podemos tener casos de personas cuya estadía supere los 6 meses en el país pero no en el mismo año, por lo que bajo esta metodología esa persona no se imputa como inmigrante. Por ejemplo: un extranjero que llega al Uruguay a mediados de Julio y se va el 30 de Junio pero del año siguiente, no será catalogado como inmigrante en ninguno de los dos años al no haber permanecido 182 días o más en el país en un año calendario.

¿Cómo superar estas limitaciones que tienen los registros de entradas y salidas?

- Utilizar solamente el aeropuerto Internacional de Carrasco como registro confiable para medir la emigración (a partir de 2025 podremos utilizar todos los puestos migratorios adheridos al nuevo sistema SIGMU, Sistema Integral de Gestión Migratoria).

- Utilizar las residencias concedidas para medir el flujo de inmigrantes.

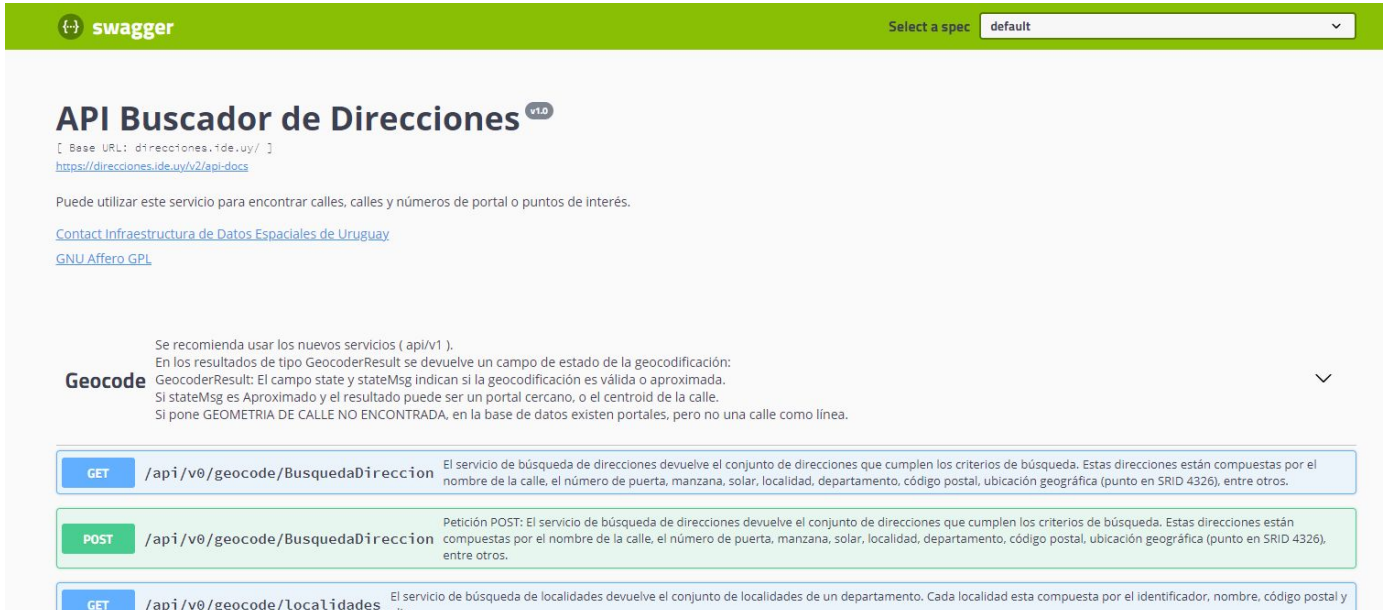
- Combinar el registro de migración con la **metodología de Señales de Vida**.

Lugar de residencia administrativa

Kevin Roelsgaard
kroelsgaard@ine.gub.uy

Sistema Único de Direcciones (SUDIR)

El Sistema Unificado de Direcciones es una base de datos única de direcciones actualizadas. Su objetivo es asegurar que todos los hogares y negocios cuenten con una dirección en un formato interoperable con un código nacional único.
Actualmente estamos utilizando servicios desarrollados por la Infraestructura de Datos Espaciales (IDE) para codificar las direcciones en los registros administrativos que recibimos. Esto nos permite estandarizar los nombres de las calles, las cuales están escritas de diferentes maneras, a un ID de calle. (https://direcciones.ide.uy/swagger-ui.html)



Identificación de fuentes

En un principio identificamos todas las fuentes donde había datos de una dirección asociada a una persona.
UTE (Compañía de electricidad)
SIIAS (Sistema de Información Integrada del Área Social proveniente del Ministerio de Desarrollo Social)
RUCAF (Registro Único de Cobertura de Asistencia Formal)
CNV (Certificados de Nacido Vivo)
ECH (Encuesta Continua de Hogares)

Se están terminando de incorporar **constancias de domicilio** e información parcial de **alquileres**.

Como siguiente paso, identificamos los niveles a los cuales podemos determinar la residencia de una persona del registro cuando la direccion viene de un registro administrativo, estos son: a nivel de departamento, localidad, calle y por último calle+ número de puerta (le llamamos dirección externa)

	id_ubic_geo	id_departamento	id_localidad	id_direccion_externa	id_fuente	id_direccion_interna	direccion_sin_codificar	id_calle
951	5.206.781	16	16220	407.589	9	[NULL]	MALDONADO 1907	8.795
952	5.206.782	16	16220	407.628	9	[NULL]	MALDONADO 1904	8.795
953	5.206.783	16	16220	407.971	9	[NULL]	MALDONADO 1959	8.795
954	5.206.784	16	16220	407.947	9	[NULL]	MALDONADO 1948	8.795
955	5.206.785	16	16220	407.929	9	[NULL]	MALDONADO 1913	8.795
956	5.206.786	16	16220	407.799	9	[NULL]	MALDONADO 1973	8.795
957	5.206.787	16	16220	407.742	9	[NULL]	MALDONADO 1939	8.795
958	5.206.788	16	16220	407.513	9	[NULL]	MALDONADO 2161	8.795
959	5.206.789	16	16220	407.893	9	[NULL]	MALDONADO 2270	8.795
960	5.206.790	16	16220	407.833	9	[NULL]	MALDONADO 2212	8.795
961	5.206.791	16	16220	407.718	9	[NULL]	MALDONADO 2282	8.795
962	5.206.792	16	16220	407.687	9	[NULL]	MALDONADO 2281	8.795
963	5.206.793	16	16220	408.104	9	[NULL]	MALDONADO 2336	8.795
964	5.206.794	16	16220	408.008	9	[NULL]	MALDONADO 2187	8.795
965	5.206.795	16	16220	407.872	9	[NULL]	MALDONADO 2271	8.795
966	5.206.796	16	16220	408.145	9	[NULL]	MALDONADO 2174	8.795
967	5.206.797	16	16220	407.790	9	[NULL]	MALDONADO 2112	8.795
968	5.206.798	16	16220	2.999.491	9	[NULL]	MALDONADO 2226	8.795
969	5.207.123	01	01020	5.853.040	24	[NULL]	LUIS FRANZINI 764	8.827
970	5.207.124	01	01020	5.853.048	24	[NULL]	LUIS FRANZINI 1319	8.827
971	5.207.125	01	01020	5.853.060	24	[NULL]	LUIS FRANZINI 1427	8.827
972	5.207.126	01	01020	5.853.066	24	[NULL]	LUIS FRANZINI 1440	8.827
973	5.207.127	01	01020	5.853.078	24	[NULL]	LUIS FRANZINI 1537	8.827
974	5.207.128	01	01020	5.853.089	24	[NULL]	LUIS FRANZINI 2675	8.827
975	5.207.129	01	01020	5.853.096	24	[NULL]	LUIS FRANZINI 9981	8.827
976	5.207.292	16	16220	414.401	9	[NULL]	DURAZNO 1566	8.862
977	5.207.293	16	16220	414.668	9	[NULL]	DURAZNO 1621	8.862
978	5.207.294	16	16220	414.065	9	[NULL]	DURAZNO 1505	8.862
979	5.207.295	16	16220	414.192	9	[NULL]	DURAZNO 1592	8.862
980	5.207.296	16	16220	414.063	9	[NULL]	DURAZNO 1572	8.862
981	5.207.297	16	16220	414.274	9	[NULL]	DURAZNO 1384	8.862
982	5.207.298	16	16220	414.536	9	[NULL]	DURAZNO 1622	8.8[illegible]
983	5.207.299	16	16220	414.482	9	[NULL]	DURAZNO 1678	8.862

Conteo de nuestro de registro de población agrupado por la especificidad de su dirección asociada.

Nivel de especificidad	Count	%
Uruguay	5684	0.16
Departamento	40593	1.15
Localidad	603900	17.09
Calle	430360	12.18
Dirección externa	2452125	69.41

¿Cómo decidimos asignar una dirección sobre otra si una persona tiene múltiples?

Fuente	id_rdu	Año	Total	Coincidencias	No coincidencias	% Coincidencias	% No coincidencias
9	2	2019	17684	9022	8662	51.02	48.98
9	2	2020	16459	8613	7846	52.33	47.67
9	2	2021	15630	8736	6894	55.89	44.11
9	2	2022	13714	8480	5234	61.83	38.17
9	2	2023	3019	1904	1115	63.07	36.93
9	3	2019	14528	9745	4783	67.08	32.92
9	3	2020	14709	9811	4898	66.70	33.30
9	3	2021	15297	10327	4970	67.51	32.49
9	3	2022	14033	10108	3925	72.03	27.97
9	3	2023	5132	3544	1588	69.06	30.94
9	4	2019	8819	5912	2907	67.04	32.96
9	4	2020	9196	6161	3035	67.00	33.00
9	4	2021	9359	6488	2871	69.32	30.68
9	4	2022	8467	6480	1987	76.53	23.47
9	4	2023	2998	2238	760	74.65	25.35
9	5	2022	4110	2637	1473	64.16	35.84
9	5	2023	452777	339380	113397	74.96	25.04
11	6	2022	215	154	61	71.63	28.37
11	6	2023	1472503	1009125	463378	68.53	31.47
29	7	2022	9489	4663	4826	49.14	50.86
29	7	2023	830848	668773	162075	80.49	19.51
33	8	2021	11056	9632	1424	87.12	12.88
33	8	2022	21437	19058	2379	88.90	11.10
33	8	2023	8911	8346	565	93.66	6.34

Instituto Nacional de Estadística

En este momento estamos asignando IDs de direcciones en función de la más específica que tenemos.
Pero ¿qué sucede cuando dos fuentes diferentes tienen direcciones externas distintas para la misma persona?
Este es un ejemplo de ese caso. La idea es también tener un orden de prioridad para nuestras fuentes, por lo que comparaciones como la anterior que se realizó contra la información proveniente del censo, nos están ayudando a decidir cuáles son más confiables.

id_estadistico_persona	id_ubic_geo	id_tiempo	especificidad	id_fuente	id_rdu
6.805.080	1.238.316	20.220.301	4	9	5
6.805.080	1.719.455	20.220.301	4	9	4

Se define prioridad de asignación de dirección:

ID RDU	Descripción	Prioridad
1	Colectivas	8
2	CNV hijos	6
3	CNV Madres	5
4	CNV Padres	4
5	RUCAF	3
6	SIIAS	7
7	UTE	2
8	ECH	1

(Otras fuentes: Vacunaciones)

PRÓXIMOS PASOS

- Actualmente, además del separador de direcciones y mejores usos de los WebServices de la IDE, se está explorando otras herramientas que no sean basados en elegir una dirección solo porque es la más específica, o porque es la más prioritaria, sino usar algún modelo, donde además de la lista de direcciones se incorporen otras variables como por ejemplo usar la tabla de vinculaciones donde podemos poner como variable la dirección del cónyuge, hijos o padres de esa persona, para poder determinar su dirección con mayor precisión.

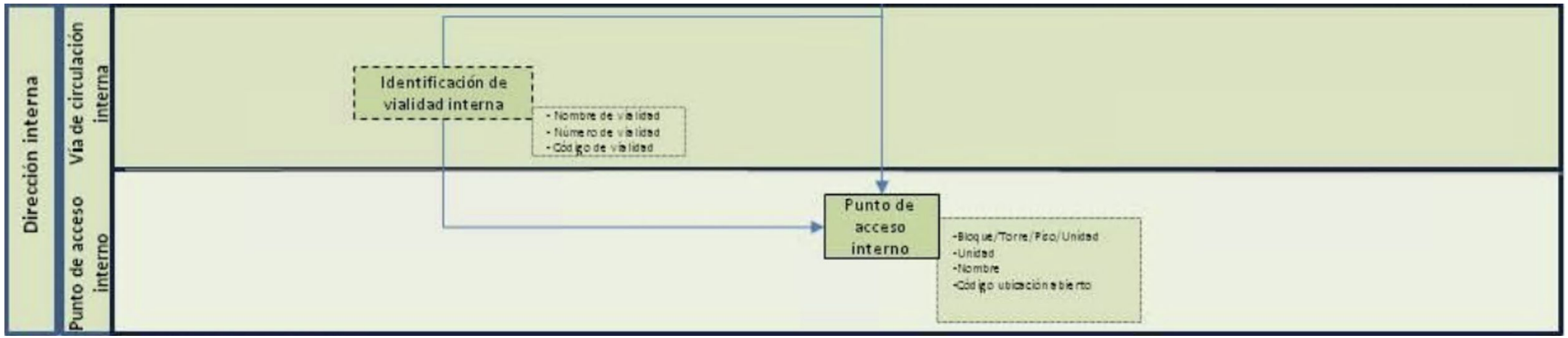
- Este ejercicio se llevó a cabo con departamento y localidad, donde se vieron mejoras en los porcentajes de coincidencia.

% aciertos (sobre el total de vinculados con censo)			
	Actual	Nuevo	
Depto	91.54	93.44	
Localidad	72.49	79.40	

Direcciones internas

- Se están identificando patrones y convenciones de escribir direcciones internas en textos de direcciones provenientes de las distintas fuentes para luego pasar a su correspondiente codificación y así poder vincular entre fuentes.

- Para esto se sigue el modelado de la IDE:



Record Linkage

Lucas Pescetto
lpescetto@ine.gub.uy

Vinculación de registros

Determinístico vs Probabilístico

coincidencia exacta en los campos especificados

+ computacionalmente eficiente y fácil de interpretar
- sensible a datos de mala calidad y faltantes
- no considera incertidumbre

coincidencias flexibles basadas en similitudes parciales

+ robusto frente a datos de mala calidad y faltantes
+ incorpora incertidumbre
- computacionalmente más costoso

URUGUAY INE Instituto Nacional de Estadística

Splink

SPLINK
Data linkage at scale

- librería de Python utilizada para vinculación determinística y probabilística
- dos tipos de vinculación:
  1. **de-duplicación (1 tabla)**
  2. **vinculación de datos de distintas fuentes (2 tablas)**
- alto rendimiento computacional y visualizaciones interactivas

URUGUAY INE Instituto Nacional de Estadística

Splink

➔ Procedimiento

1. pre-procesamiento de la/s tablas y análisis exploratorio
2. reglas de bloqueo
3. estimación de los parámetros
4. selección de umbral y predicción
5. visualización y evaluación de los resultados

1. pre-procesamiento de la/s tablas y análisis exploratorio

tabla 1: personas censadas (3.1 m)

tabla 2: hub personas DNIC (6.1 m)

variables en común:

- C.I. (documento uruguayo)
- fecha de nacimiento
- primer nombre, segundo nombre, primer apellido, segundo apellido.
- departamento (división geográfica)
- sexo
- documento extranjero
- país del documento extranjero

1. pre-procesamiento de la/s tablas y análisis exploratorio

a. separación y estandarización de nombres

hub personas

nombres		apellidos	
1° nombre	2° nombre	1° apellido	2° apellido
nombres + apellidos			

censo

1° nombre	2° nombre	1° apellido	2° apellido

+ limpieza
+ estandarización

1. pre-procesamiento de la/s tablas y análisis exploratorio

b. departamento

- Se combinan algunas divisiones geográficas

c. documento

- Se descartan documentos genéricos o repetidos muchas veces

2. reglas de bloqueo

- total de comparaciones > 19.154.000.000

1. documento uruguayo
2. primer nombre, segundo nombre, primer apellido, segundo apellido
3. año de nacimiento, primer nombre, primer apellido
4. departamento, año de nacimiento, primer nombre y primera letra del primer apellido
5. departamento y fecha de nacimiento

3. estimación de los parámetros

a. definición de las comparaciones

¿match exacto? **¿campos similares?**

surname_l	surname_r	comparison_level	interpretation
Rob	Rob	Exact match	great match
Rob	Jane	All other	bad match
Rob	Robert	All other	bad match, this comparison has no notion of nicknames

b. estimación

4. selección de umbral y predicción

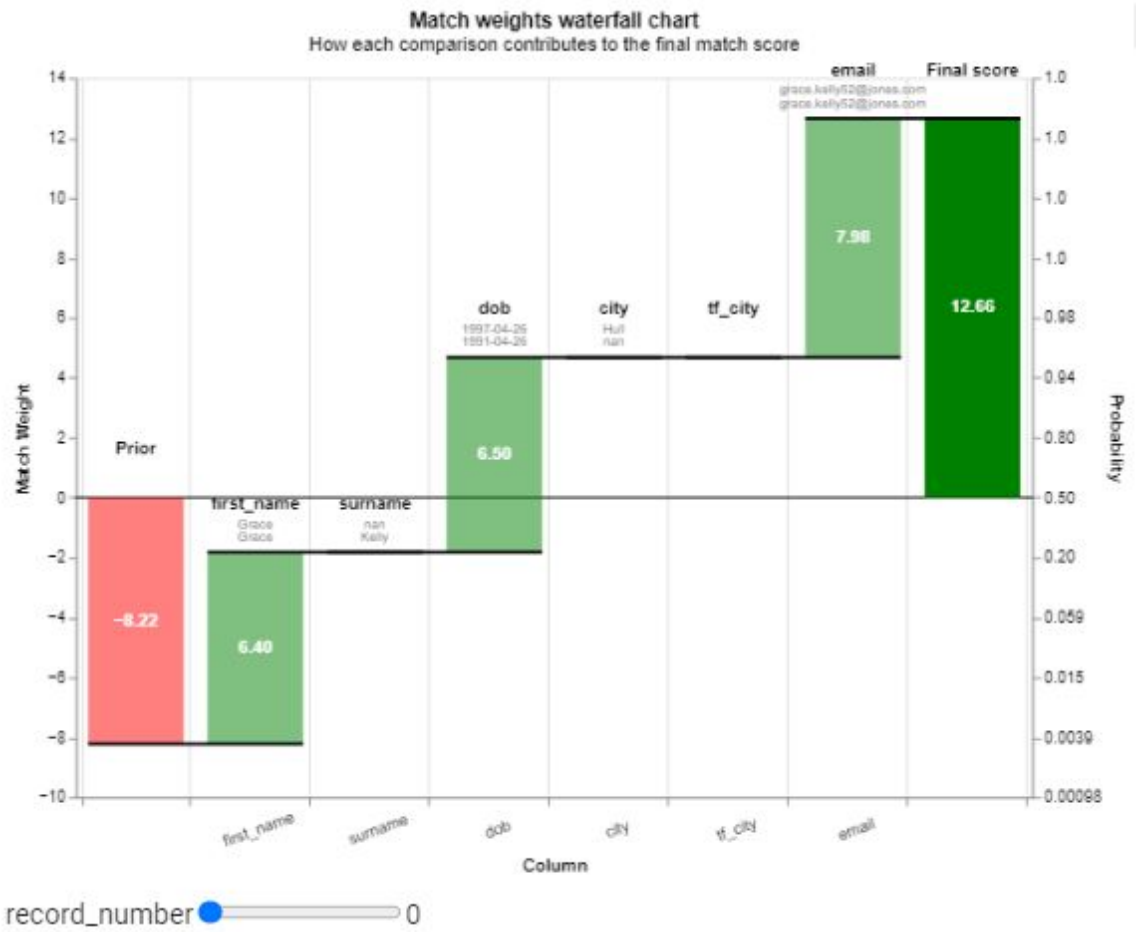
↑ umbral
=
↑ falsos negativos

↓ umbral
=
↑ falsos positivos

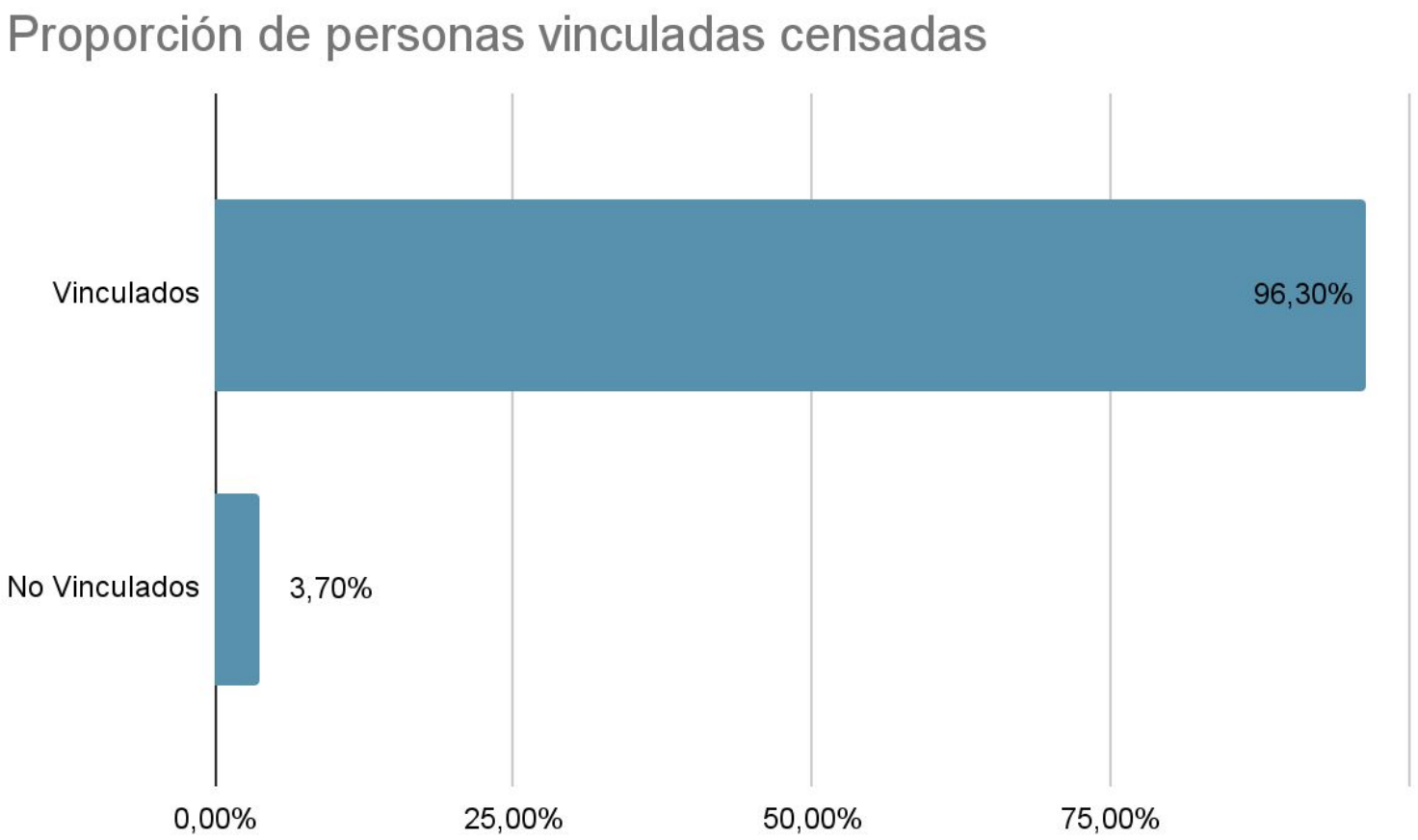
- Se pueden utilizar reglas lógicas para decidir qué considerar como match:
  1. probabilidad_match > 0.999
  2. probabilidad_match ∈ (0.9, 0.999);
     (coinciden los documentos y el documento es único); o **(coincide la fecha de nacimiento y primer_nombre y primer_apellido son similares)**

5. visualización y evaluación de los resultados

- ¿etiquetas?
- ¿reglas lógicas?



Vinculación Censo - RRAA: resultados



Vinculación Censo - RRAA: resultados

Documento	Proporción de vinculados (%)
match exacto	94.2%
similar	0.35%
distinto	0.05%
valor faltante	5.4%

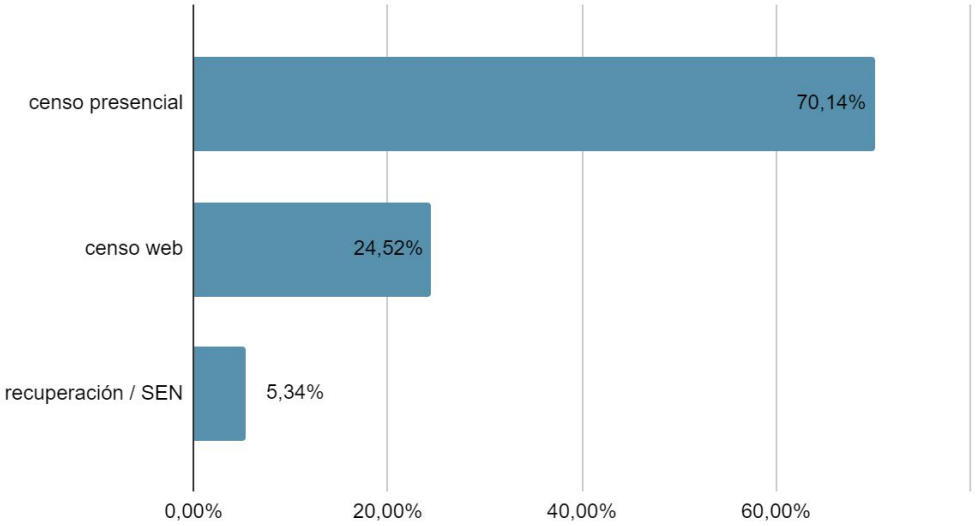
Fecha de nacimiento	Proporción de vinculados (%)
match exacto	95.7%
similar	3.5%
distinto	0.6%
valor faltante	0.04%

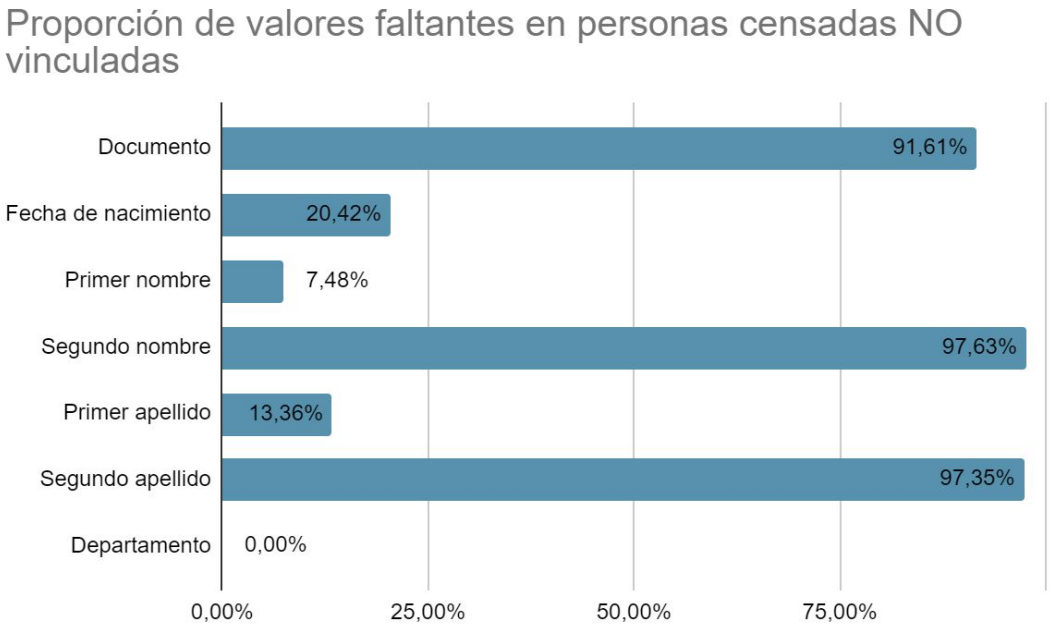
Primer nombre	Proporción de vinculados (%)
match exacto	90.2%
similar	2.9%
distinto	6.8%
valor faltante	0.1%

Primer apellido	Proporción de vinculados (%)
match exacto	94.7%
similar	3.4%
distinto	1.4%
valor faltante	0.5%

URUGUAY INE Instituto Nacional de Estadística

Vinculación Censo - RRAA: resultados

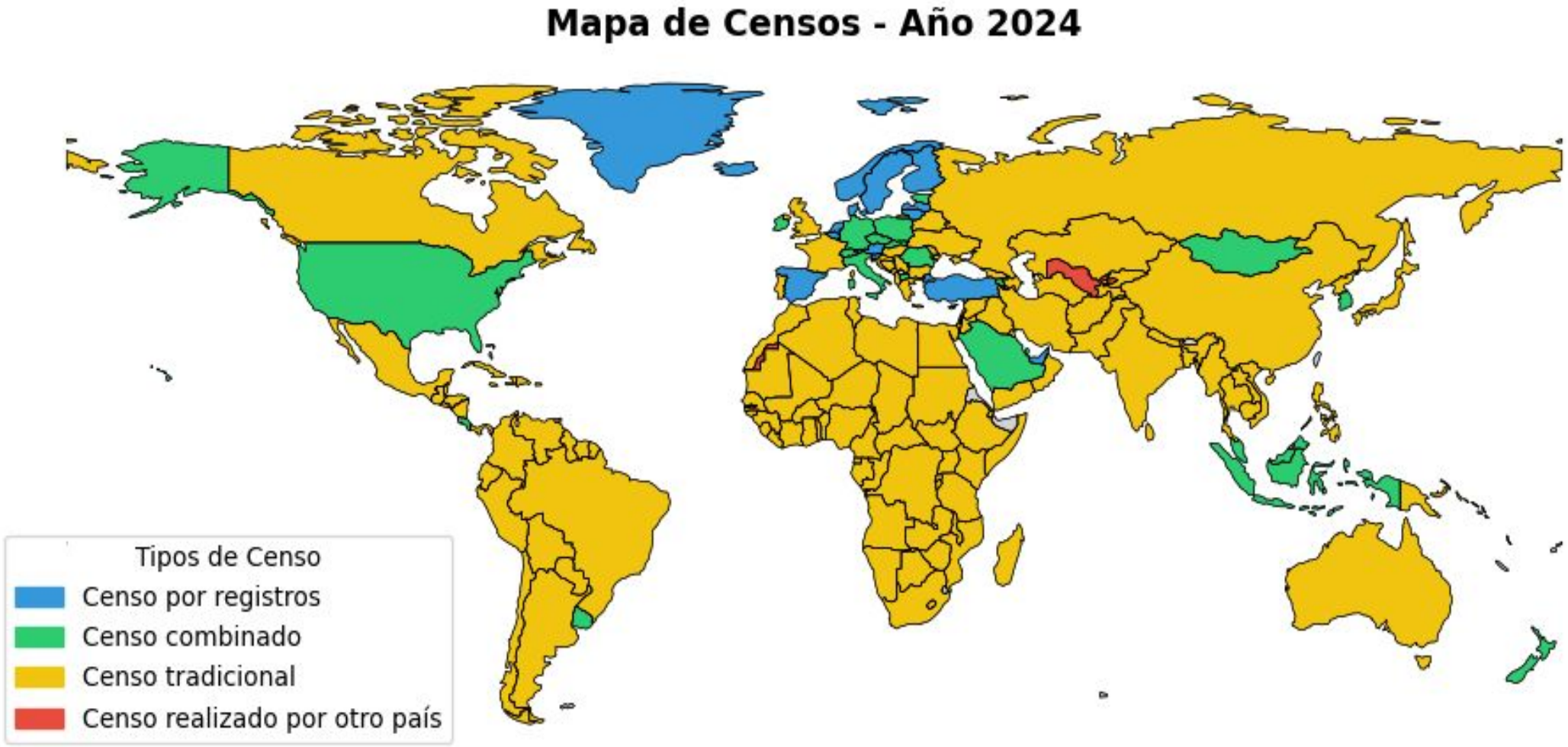




Metodología de Señales de vida e Índice de residencia

Pablo Dubourdieu
pdubourdieu@ine.gub.uy

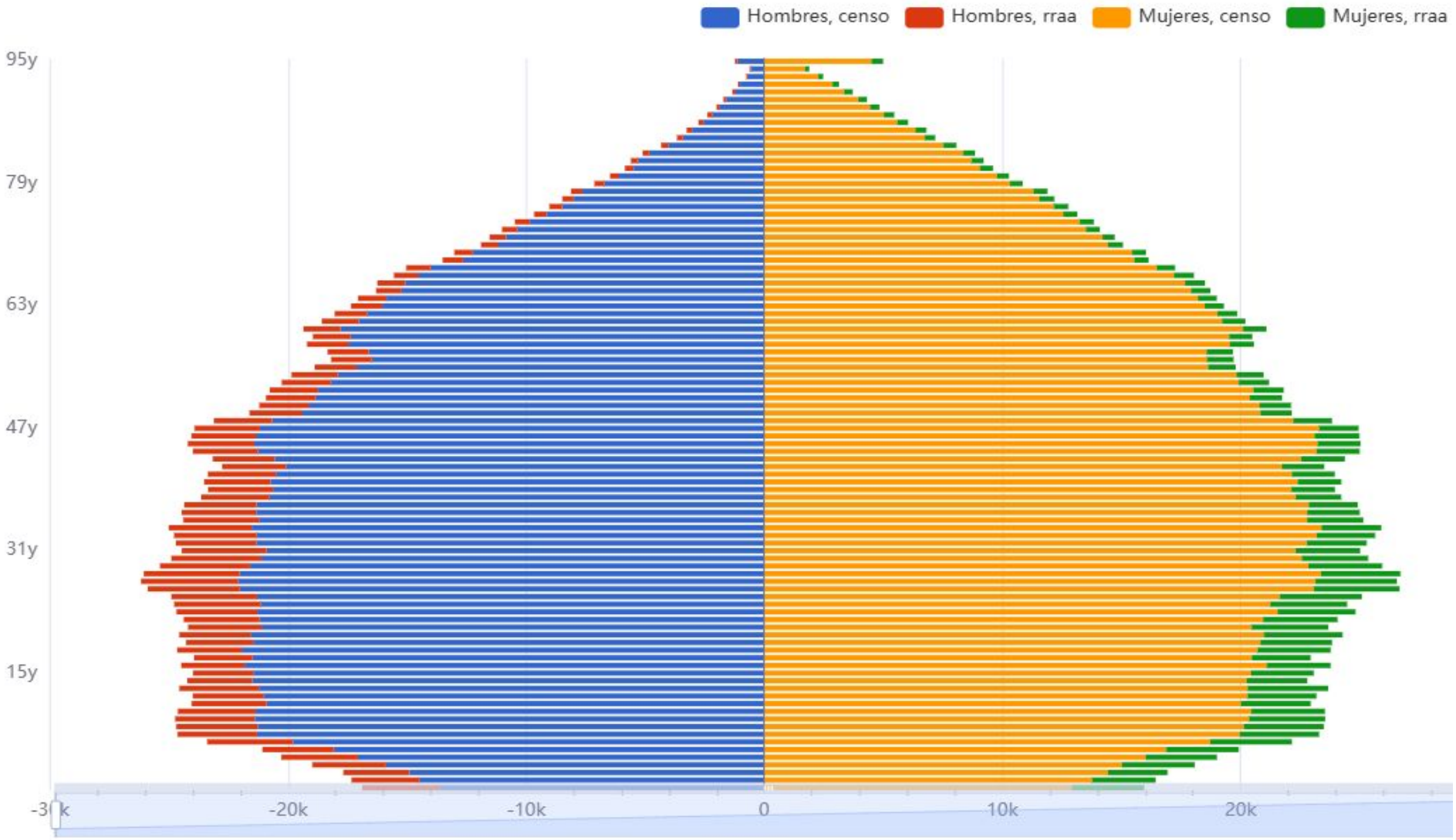
Estado del arte: Experiencias otros países



Censo combinado 2023

- El número final de la población uruguaya surge a partir de distintas estimaciones:
  - Censo + ENEC (Encuesta Nacional de Evaluación Censal)
  - Análisis de la evolución de los componentes demográficos
  - Método CAWI
  - Método UTE
- Se utilizan registros para identificar la población omitida.
- Se seleccionan las personas con mayor probabilidad de ser residente.

Pirámide poblacional censo por fuente del dato



El problema

- Residentes con documento nacional
- Residentes sin documento nacional

- Más de 6 millones de documentos (DNIC)
- Debemos clasificarlos en:
  - Residentes
  - Emigrantes
  - Fallecidos

} SDV

Buen registro MSP

La información utilizada

- **Variables Demográficas** de personas con documento nacional (Fecha de nacimiento, sexo, edad, fecha de fallecimiento)
- **Señales De Vida** (Actividad, Educación, Seguridad Social, Entradas y salidas del país, Vacunas*, Colectivos*, etc.)
- **Registros de Ubicación** (UTE, Alquileres, Sistema de Salud, etc.)
- **Información de Migraciones** (Días fuera del país, Cobro de pensiones fuera del país)
- **Censo** (Población censada vinculada) – Variable a predecir.

URUGUAY INE Instituto Nacional de Estadística

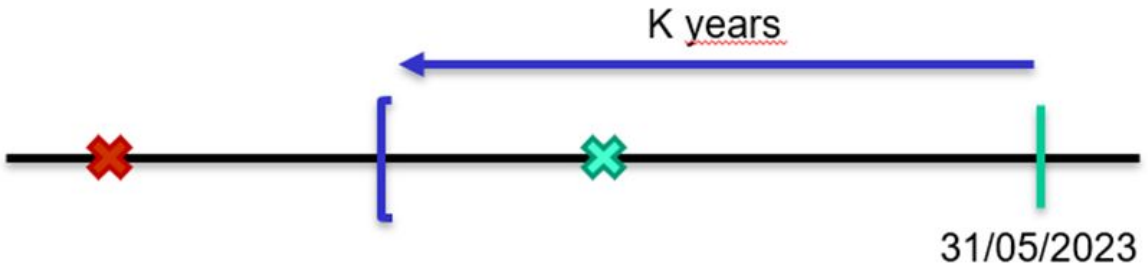
¿Qué es una Señal de Vida?

Una señal de vida es la presencia de un individuo en un registro específico y en un período acotado.

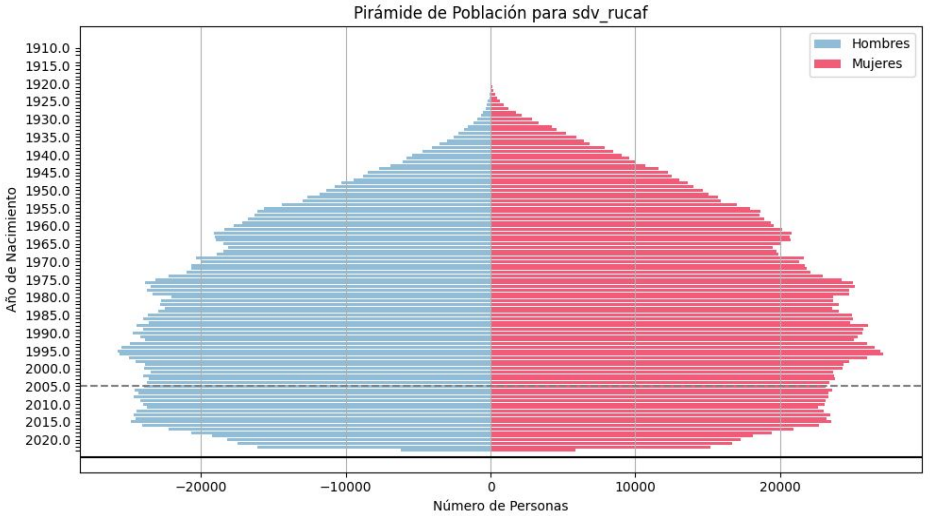
id_persona - id_tiempo - id_sdv

El proceso actual considera 24 señales de vida:

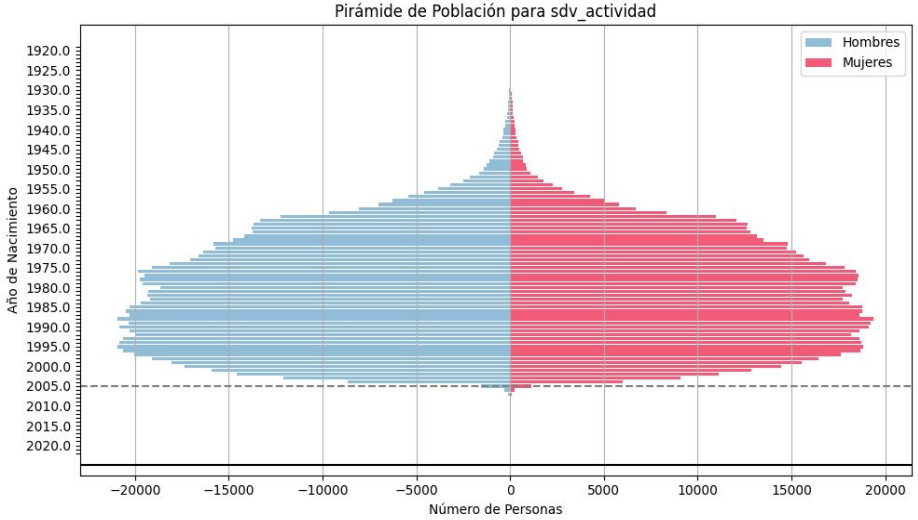
- 13 obtenidas de fuentes directas.
- 8 obtenidas por el SIIAS.
  (Sistema de Información Integrada del Área Social)
- 3 generadas por el censo



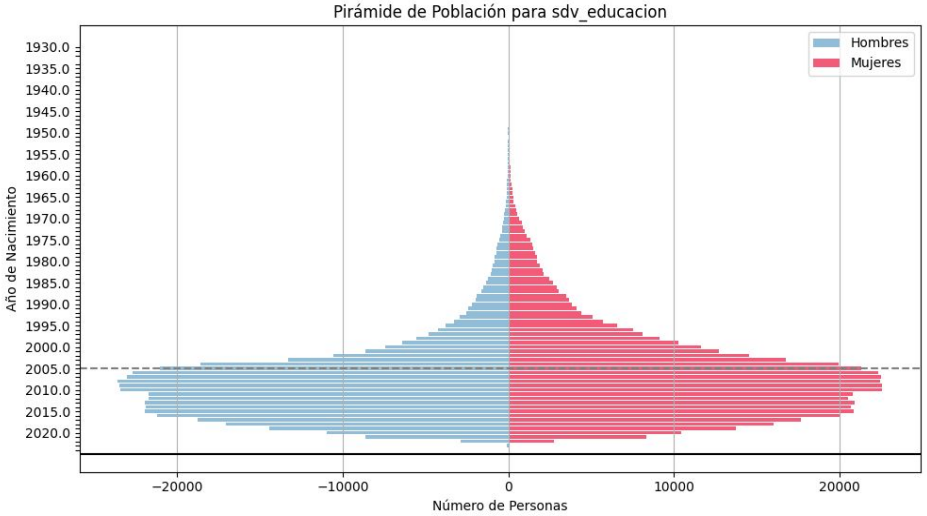
Pirámides de distintas SDV



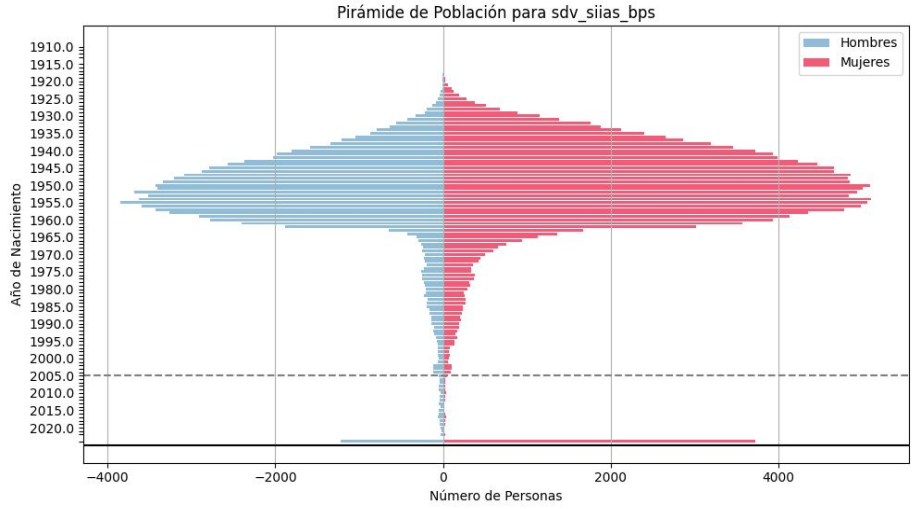
Total de Personas: 3476952 (Hombres: 1673620, Mujeres: 1803332)



Total de Personas: 1570760 (Hombres: 827907, Mujeres: 742853)

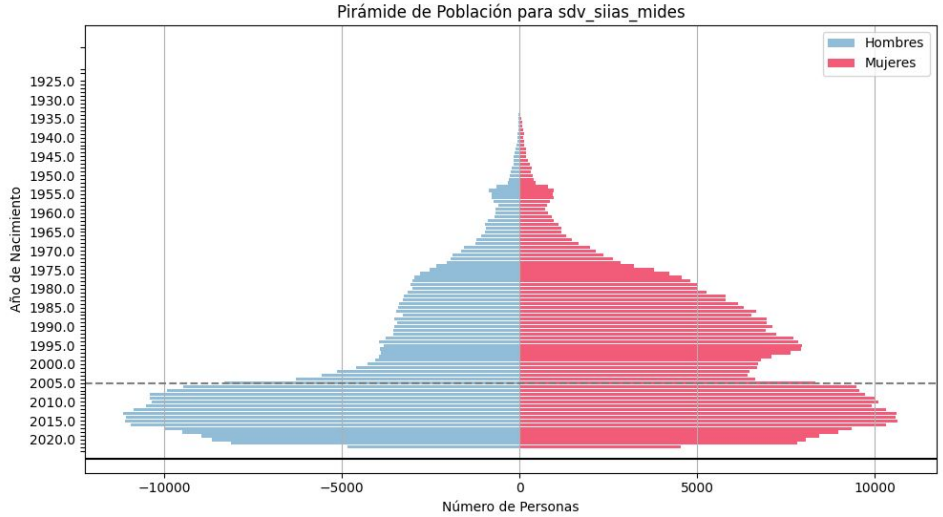


Total de Personas: 966913 (Hombres: 458156, Mujeres: 508757)

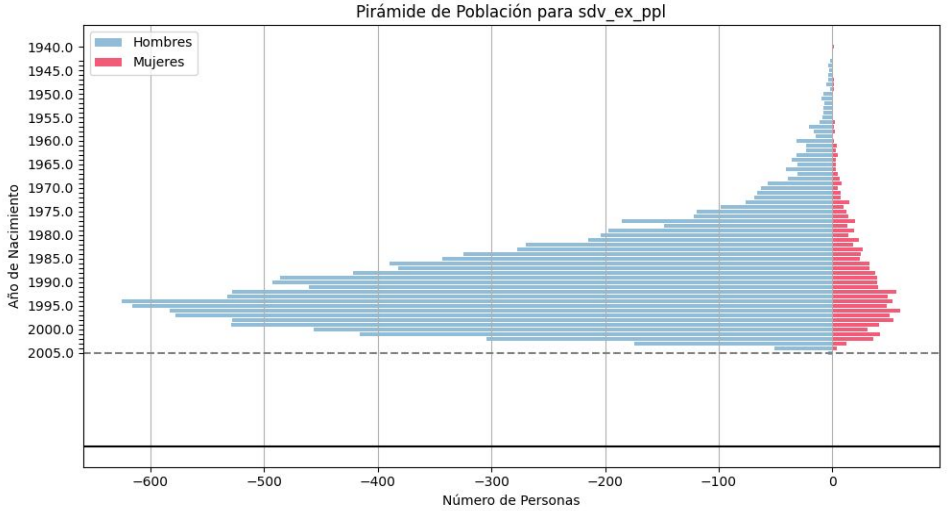


Total de Personas: 235889 (Hombres: 87177, Mujeres: 148712)

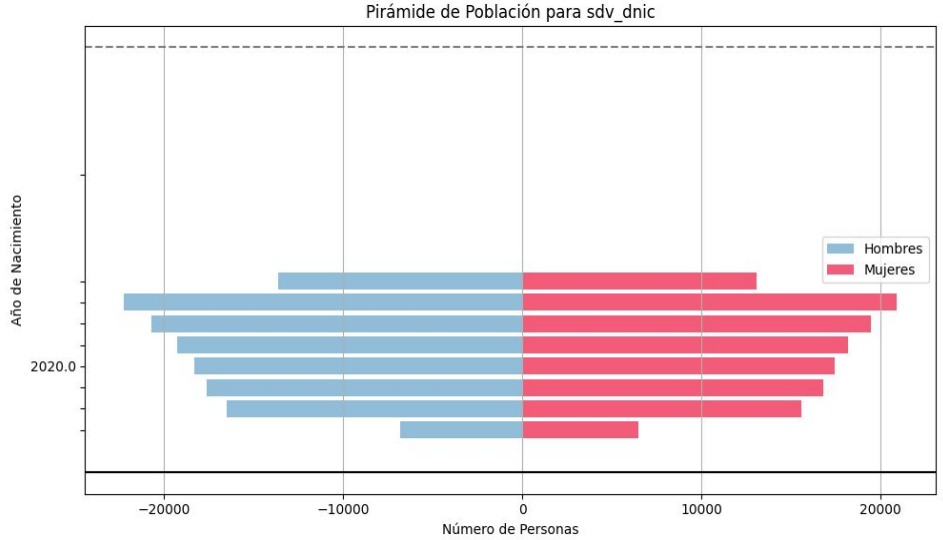
Pirámides de distintas SDV



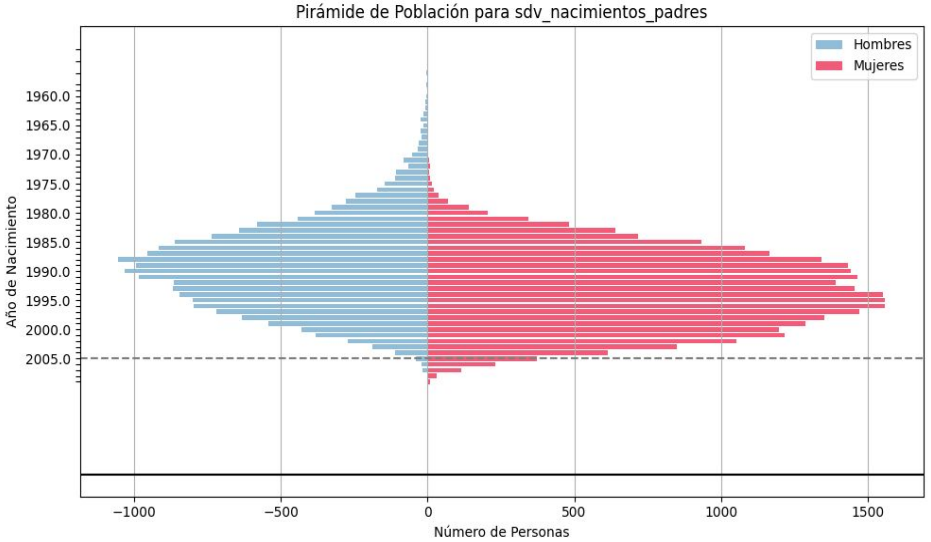
Total de Personas: 707796 (Hombres: 314145, Mujeres: 393651)



Total de Personas: 12878 (Hombres: 11817, Mujeres: 1061)

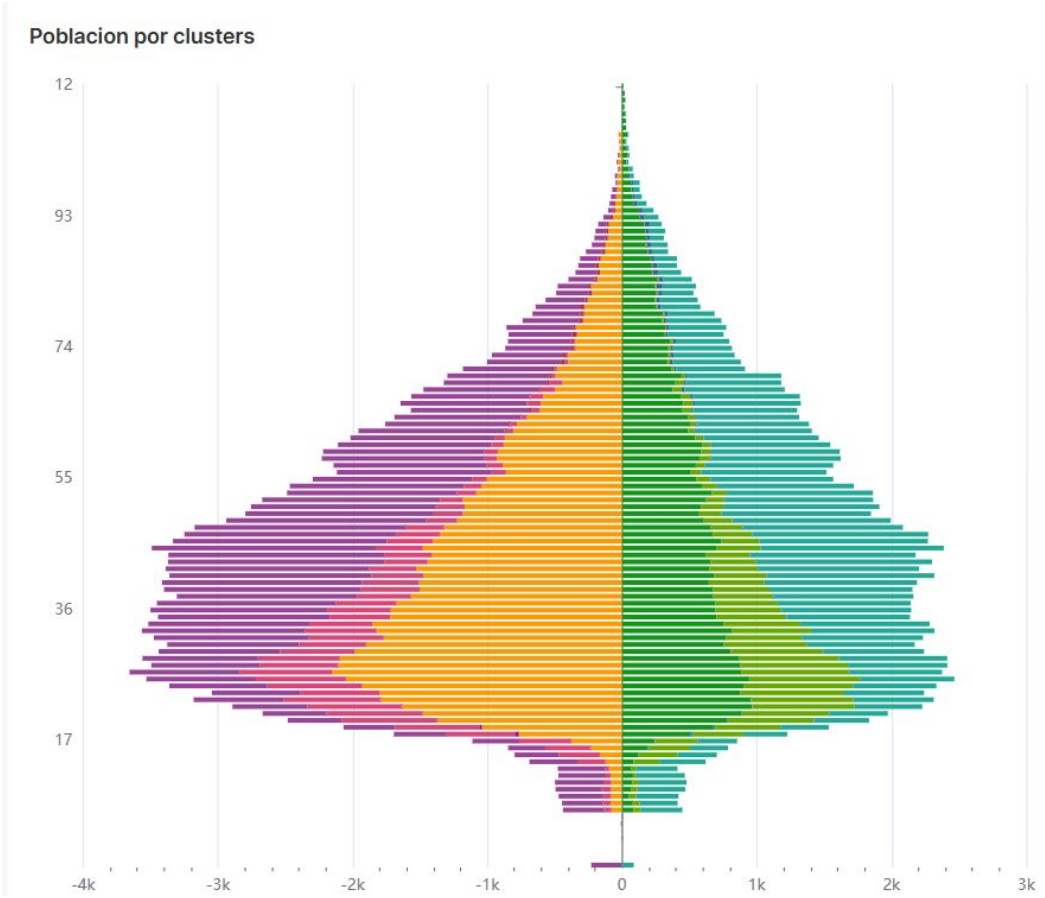


Total de Personas: 263066 (Hombres: 135098, Mujeres: 127968)



Total de Personas: 47767 (Hombres: 18897, Mujeres: 28870)

Análisis por clusters



Resultados Modelo Random Forest

El modelo se ajusta con el objetivo de priorizar la identificación de la población censada.
Esto implica un mayor número de falsos positivos, personas clasificadas cómo residentes que no fueron censadas.

Indicador	Resultados
Accuracy	0,87
ROC AUC	0,91

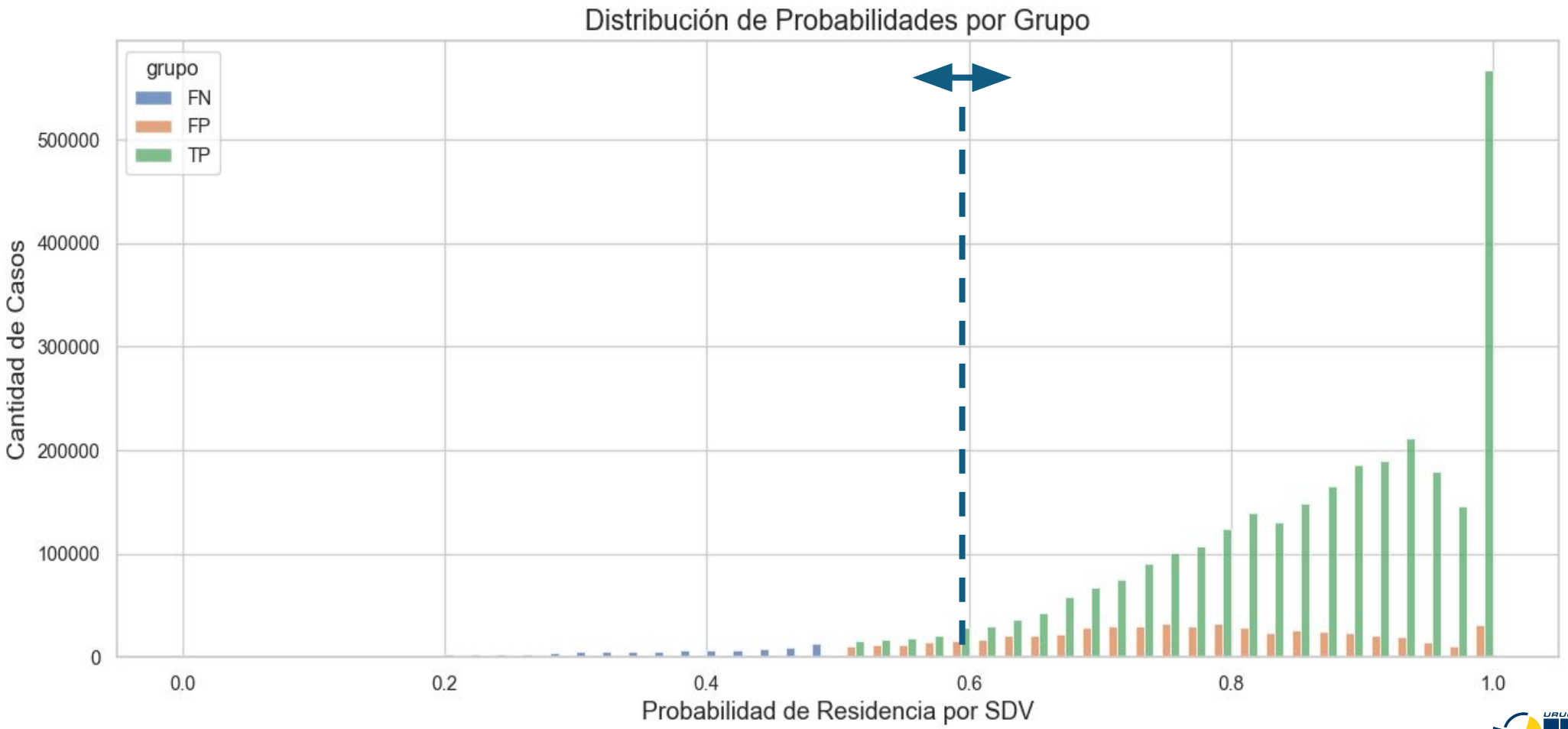
Accuracy:
Proporción de predicciones correctas sobre el total de casos evaluados.
ROC-AUC:
Evalúa la capacidad del modelo para distinguir entre clases.

Censado	Precision	Recall	F1-score
0	0,96	0,71	0,82
1	0,83	0,98	0,90

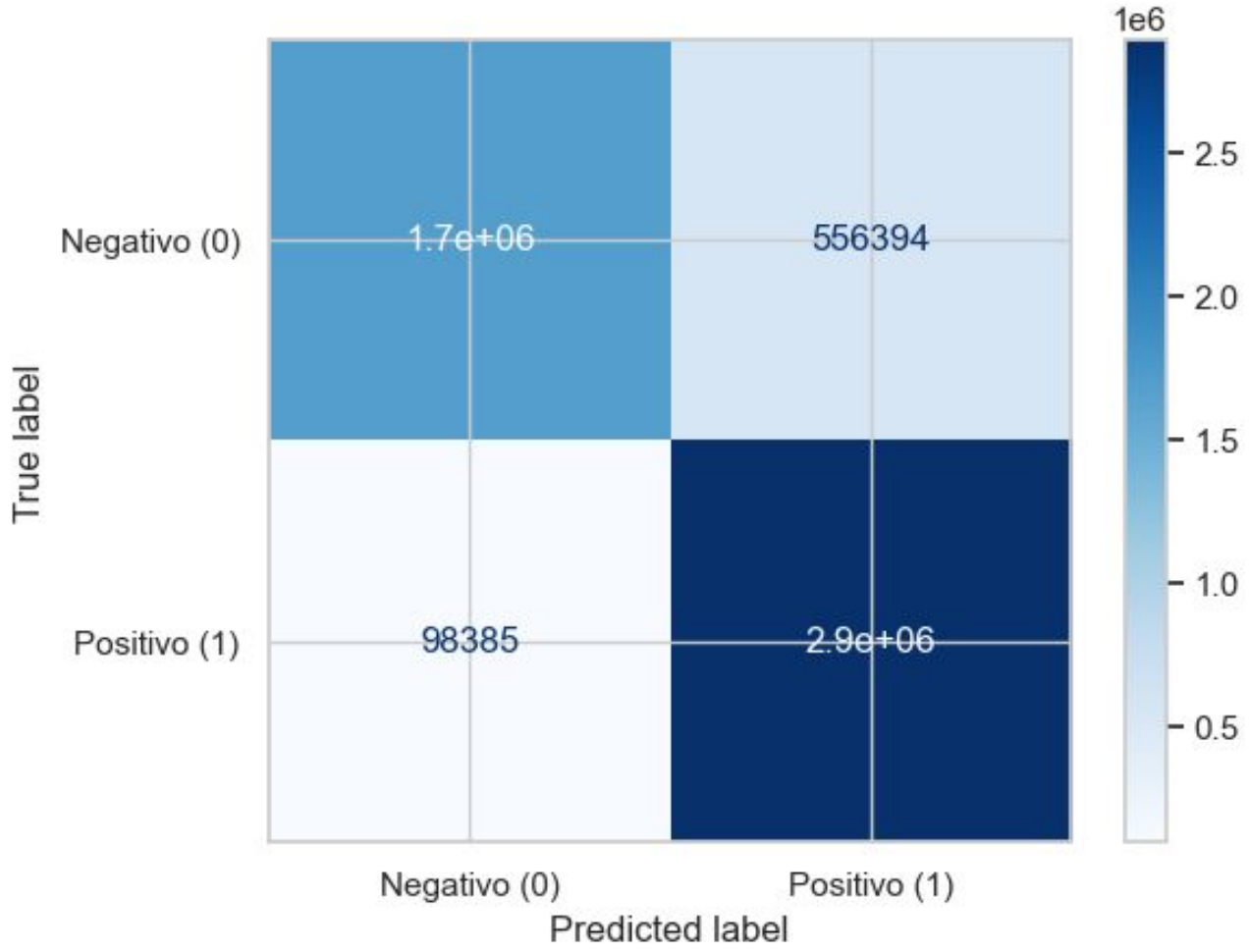
Precision: Indica cuántas de las predicciones positivas fueron realmente correctas.
Recall (Sensibilidad): Mide la capacidad del modelo para detectar correctamente la clase positiva.
F1-Score: Media armónica entre precisión y recall, equilibrando ambos indicadores.

URUGUAY INE Instituto Nacional de Estadística

Probabilidad de residencia



Matriz de confusión



¿Dudas, preguntas, comentarios?