ISSN 1680-8789

SERIE

ESTUDIOS ESTADÍSTICOS 111

Encuestas de posenumeración censal

Fundamentos estadísticos para su diseño y análisis

Andrés Gutiérrez
Giovany Babativa-Márquez

NACIONES UNIDAS

CEPAL

Gracias por su interés en esta publicación de la CEPAL

NACIONES UNIDAS

CEPAL

Si desea recibir información oportuna sobre nuestros productos editoriales y actividades, le invitamos a registrarse. Podrá definir sus áreas de interés y acceder a nuestros productos en otros formatos.

Deseo registrarme

Conozca nuestras redes sociales y otras fuentes de difusión en el siguiente link:

https://bit.ly/m/CEPAL

SERIE

ESTUDIOS ESTADÍSTICOS 111

Encuestas de posenumeración censal

Fundamentos estadísticos para su diseño y análisis

Andrés Gutiérrez
Giovany Babativa-Márquez

NACIONES UNIDAS

CEPAL

Este documento fue elaborado por Andrés Gutiérrez, Asesor Regional en Estadísticas Sociales de la División de Estadísticas de la Comisión Económica para América Latina y el Caribe (CEPAL), y por Giovany Babativa-Márquez, Consultor de la misma División.

Se agradece especialmente al equipo del Centro Latinoamericano y Caribeño de Demografía (CELADE)-División de Población de la CEPAL por sus valiosos comentarios y aportes a la versión final de este documento.

Publicación de las Naciones Unidas
ISSN: 1680-8789 (versión electrónica)
ISSN: 1680-8770 (versión impresa)
LC/TS.2025/117
Distribución: L

Impreso en Naciones Unidas, Santiago
S.2500781[S]

Esta publicación debe citarse como: Gutiérrez, A. y Babativa-Márquez, G. (2026). Encuestas de posenumeración censal: fundamentos estadísticos para su diseño y análisis. *Serie Estudios Estadísticos* (111) (LC/TS.2026/117). Comisión Económica para América Latina y el Caribe.

Índice

Resumen **5**

Introducción **7**

I. Marco conceptual: el sistema de estimación dual **11**
A. Planteamiento del problema 11
B. Condiciones y supuestos 12
C. Inferencia 16

II. Planeación: cuestionario, muestreo y operativo **19**
A. Diseño del cuestionario 20
B. El diseño de muestreo 23
C. Operativo de recolección 29

III. Procedimientos y enumeraciones **31**
A. La muestra P y la muestra E 31
B. Procedimientos para la reconstrucción de los hogares 32
C. Clasificación de las enumeraciones 34

IV. Emparejamiento estadístico **39**
A. Preprocesamiento 41
B. Indexación 48
C. Comparación 49
D. Clasificación 52
E. Evaluación 53
F. Ejemplo aplicado 55

V. Estimadores en el Sistema de Estimación Dual **59**
A. Estimadores de muestreo 59
B. Otros estimadores del SED 64
C. Estimadores basados en modelos log-lineales 71

D. Comparación entre los estimadores ... 73

VI. Modelamiento estadístico ... 75
A. Creación de ponderaciones para estimar la cobertura censal ... 76
B. Modelamiento de las probabilidades mediante regresión logística ... 77
C. Ejemplo aplicado ... 79

VII. Conclusiones ... 83

Bibliografía ... 85

Serie de Estudios Estadísticos: números publicados ... 88

Cuadros

Cuadro 1 Distribución conjunta de la población censada y encuestada ... 12
Cuadro 2 Matriz de conteos del sistema de estimación dual ... 13
Cuadro 3 Categorización de personas para el emparejamiento entre el Censo y la EPC ... 22
Cuadro 4 Ejemplo de los miembros de un hogar según su residencia en la fecha del censo y de la EPC ... 33
Cuadro 5 Tipos de enumeraciones erróneas ... 36
Cuadro 6 Direcciones y localización geográfica asociada ... 42
Cuadro 7 Codificación de caracteres en el algoritmo Soundex ... 44
Cuadro 8 Codificación de nombres con algunos algoritmos de codificación fonética ... 46
Cuadro 9 Codificación de apellidos con algunos algoritmos de codificación fonética ... 47
Cuadro 10 Clasificación de registros basados en la codificación Soundex ... 53
Cuadro 11 Registros simulados entre la EPC y el censo ... 55
Cuadro 12 Codificación Metaphone de nombres y apellidos para los registros simulados ... 56
Cuadro 13 Clasificación de registros entre la EPC y el censo ... 57
Cuadro 14 Matriz inicial de estimadores necesarios para el SED ... 61
Cuadro 15 Matriz de estimadores de Petersen para el SED ... 63
Cuadro 16 Matriz de estimadores del modelo loglineal con interacción para el SED ... 73
Cuadro 17 Resumen de las ventajas de algunos estimadores del SED ... 74
Cuadro 18 Coeficientes de regresión estimados para el modelo de enumeraciones correctas ... 80
Cuadro 19 Coeficientes de regresión estimados para el modelo de emparejamiento ... 81

Diagrama

Diagrama 1 Flujo general del proceso de emparejamiento ... 41

Resumen

Sin duda, la operación estadística más exigente que realizan los Institutos y Oficinas Nacionales de Estadística en los países son los censos de población y vivienda, los cuales se diseñan, entre otros fines, para empadronar y enumerar a todos y cada uno de los habitantes de un país. Millones de personas son contadas y entrevistadas en un levantamiento que, desde cualquier punto de vista, pretende ser exhaustivo y moviliza a miles de personas en campo. Un operativo de esta naturaleza siempre está expuesto a diversas fuentes de error y, por ende, es necesario evaluar la calidad de los indicadores resultantes, siendo uno de los más importantes el número total de habitantes del país. Las encuestas de posenumeración censal cumplen este propósito y, además, se utilizan para realizar correcciones a dichos conteos. Mucho se ha escrito sobre este tipo de encuestas, pero la literatura carece de un documento que sintetice todos los procesos estadísticos que deben considerarse para diseñar y analizar este tipo de operaciones. Este documento busca llenar ese vacío metodológico con una aproximación técnica robusta y actualizada que siente las bases para su planeación: desde el diseño de muestreo, pasando por el emparejamiento de sus registros con los del censo de población, hasta la definición de estimadores apropiados basados en el sistema dual y su posterior modelamiento estadístico.

Introducción

Las encuestas de posenumeración censal (EPC, también conocidas internacionalmente como PES, por sus siglas en inglés) constituyen un estudio complementario a los censos de población y vivienda, cuyo objetivo principal es evaluar la cobertura y la calidad de la información recolectada en el censo. Su implementación permite detectar posibles errores en la enumeración, tales como omisiones, duplicaciones o clasificaciones incorrectas, lo cual contribuye a mejorar la precisión de los resultados censales y, a su vez, ofrece insumos fundamentales para el diseño de futuras operaciones estadísticas.

Las EPC son un insumo adicional para los procesos de conciliación demográfica, ya que permiten ajustar los conteos censales mediante su integración con estadísticas vitales, registros administrativos y encuestas especializadas. Al proporcionar estimaciones más precisas del tamaño y la estructura de la población, mejoran la coherencia de las series y fortalecen tanto las estimaciones como las proyecciones demográficas. De este modo, no solo corrigen las cifras censales en el corto plazo, sino que también aportan información confiable para la planificación de políticas públicas y el análisis de la dinámica poblacional a largo plazo. Al identificar y cuantificar los errores de cobertura, estas encuestas refuerzan la credibilidad de las cifras censales que sustentan indicadores esenciales en áreas como pobreza, educación, salud, igualdad de género y vivienda adecuada. Con ello, contribuyen a que los datos nacionales reflejen con mayor precisión la realidad poblacional, asegurando que ningún grupo quede invisibilizado en las estadísticas oficiales.

Por último, en el marco de la Agenda 2030, las EPC permiten evaluar la cobertura censal con un nivel de desagregación demográfica y territorial esencial para el seguimiento del principio de "no dejar a nadie atrás". Al analizar la información por sexo, edad, etnia, región o condición socioeconómica, estas encuestas identifican posibles brechas en la representación de grupos vulnerables, lo que facilita la toma de decisiones informadas para el diseño de políticas públicas y programas focalizados. Además de la evaluación general, las EPC facilitan el análisis del impacto de diversos factores que influyen en la calidad de un censo. Entre ellos se destacan la movilidad poblacional, las estrategias de recolección de datos y el desempeño de los enumeradores en el terreno. De esta manera, los resultados de la encuesta no solo permiten corregir o ajustar los resultados derivados de la información censal, sino que también genera conocimiento sobre los elementos que condicionan la efectividad del operativo.

Los objetivos de una encuesta de posenumeración censal incluyen la estimación de los errores de cobertura, mediante la comparación entre los resultados del censo y los de la encuesta para cuantificar las discrepancias. Asimismo, buscan identificar y analizar los componentes de cobertura, como las omisiones y duplicaciones, así como otras posibles clasificaciones erróneas. Este análisis también se nutre de la diferenciación por grupos demográficos, considerando variables como edad, sexo, etnia, región y condición socioeconómica. Por ende, los indicadores principales de la encuesta se derivan de modelos de captura y recaptura, que determinan no sólo el análisis de la encuesta, sino su diseño metodológico. Entre los más relevantes se encuentran:

- El error neto de cobertura que mide la diferencia entre la población real y la censada.
- Las tasas de omisión y de inclusiones erróneas, que permiten dimensionar las personas no contadas o mal clasificadas en relación con la población de referencia.
- La tasa de emparejamiento, que refleja la proporción de registros de la EPC correctamente vinculados con los del censo.
- La tasa bruta de error de cobertura, que resume, en términos relativos, la magnitud conjunta de las omisiones y de las inclusiones erróneas con respecto al tamaño de la población censada.

El éxito de este tipo de estudios dependen de varios factores que se basan en la apropiación de técnicas estadísticas diversas que determinarán la solidez del diseño muestral probabilístico en la selección de la muestra de áreas, de la eficiencia en la logística de recolección de la información en campo, de la calidad del sistema de emparejamiento de registros entre la EPC y el censo, de la pertinencia de los métodos estadísticos de estimación empleados y de la robustez del modelamiento de la omisión encontrada. Estos factores garantizan que los resultados sean confiables, representativos y útiles para evaluar la calidad del censo y para orientar la planificación de futuras rondas censales.

A nivel internacional, se han establecido lineamientos detallados para la realización de censos de población y vivienda, incluyendo un conjunto de principios y recomendaciones específicamente dedicados a las encuestas de posenumeración censal (United Nations, 2025). Este marco normativo describe los objetivos fundamentales de estas encuestas, sus componentes metodológicos básicos y las buenas prácticas para asegurar estimaciones confiables del error de cobertura. Una referencias muy completa en el plano conceptual es el trabajo de Whitford y Banda (2002), quienes ofrecen una introducción estadística estructurada al diseño de las encuestas de posenumeración, discutiendo los principales enfoques y supuestos detrás de los métodos de emparejamiento y los desafíos más frecuentes que enfrentan las oficinas nacionales de estadística al implementar estas investigaciones. En el contexto regional, Chackiel (2010) presenta una síntesis valiosa sobre el uso de las encuestas de posenumeración como herramientas para evaluar la calidad censal en América Latina. Asimismo, el documento de CEPAL (1999) recopila experiencias de varios países que implementaron evaluaciones directas e indirectas de cobertura. Más recientemente, Borges y Queiróz (2025) documentaron la experiencia brasileña, aportando evidencia sobre los aspectos metodológicos y estrategias operativas que contribuyeron a robustecer la medición de la omisión censal en Brasil.

En este sentido, es importante señalar que los resultados de una EPC no están concebidos para modificar o editar los microdatos del censo, sino para aportar información independiente que permita evaluar su cobertura y calidad. Su contribución principal es ofrecer estimaciones agregadas que fortalecen la interpretación de los resultados censales y apoyan la planificación estadística. Más relevante aún, las EPC cumplen un rol complementario dentro de un marco analítico más amplio: no reemplazan los procesos de conciliación demográfica ni las estimaciones y proyecciones de población. Estos ejercicios de mayor alcance permiten reconstruir la dinámica poblacional y garantizar coherencia entre censos, registros administrativos y encuestas, algo que va más allá del objetivo de una EPC.

Este documento no pretende ser una guía operativa, sino una herramienta de consulta para el estadístico que esté involucrado en este tipo de encuestas, ya sea poque está diseñando una EPC, o porque debe proveer estimaciones de la calidad del censo usando los datos de ambos levantamientos. En cualquiera de los casos, es necesario tener un entendimiento profundo de la base inferencial que rige este tipo de levantamientos. Y es que el sistema de estimación dual, basado en los métodos de captura y recaptura, dista del paradigma tradicional de las encuestas de hogares centrado únicamente en el diseño de muestreo.

Sin embargo, el procesamiento típico de una EPC debe tener en cuenta el emparejamiento con los registros del censo, siendo este el pilar fundamental del sistema de estimación dual. Sin emparejamiento no existe la posibilidad de evaluar la calidad del censo, y este proceso, además, descansa en metodologías estadísticas avanzadas. Finalmente, la aplicación de los estimadores correctos es fundamental para que la inferencia sea apropiada y se pueda realizar un análisis estadístico robusto para proveer resultados concluyentes acerca de la calidad del censo.

Siguiendo este planteamiento, el primer capítulo presenta las bases inferenciales de las EPC: el sistema de estimación dual (SED). Es importante que antes de diseñar o analizar una encuesta de este tipo, el investigador tenga un completo dominio y entendimiento de los supuestos que rigen los métodos de captura y recaptura. En este capítulo se presentan las condiciones bajo las cuales es posible realizar una inferencia apropiada, las cuales a su vez determinan las condiciones logísticas que se deben tener en cuenta al momento de la planificación de una EPC. Asimismo, se presenta la racionalidad de los estimadores del SED, junto con sus propiedades estadísticas.

El segundo capítulo del documento gira en torno a la planificación de este tipo de encuestas, la cual depende directamente de los objetivos que se quieran conseguir con la investigación. Este capítulo aporta los elementos fundamentales en cuanto al cuestionario, al diseño de muestreo, la determinación de los tamaños de muestra y al operativo de recolección. En cada una de estas secciones se aportan recomendaciones importantes que redundarán en una correcta vinculación de los registros con los del operativo censal para su correspondiente análisis, siempre teniendo en mente con el cumplimiento de las condiciones planteadas por el SED.

El tercer capítulo aborda dos conceptos claves para definir los errores de cobertura en el censo; estos son la muestra P (de posenumeración independiente) y la muestra E (de registros censales). En este sentido se plantean los conceptos apropiados para definir cuándo una enumeración es correcta o errónea, así como los fundamentos del emparejamiento desde una perspectiva procedimental orientada a reconstruir los hogares tal como existían en el momento de la recolección censal.

El cuarto capítulo introduce los elementos básicos para el emparejamiento estadístico a partir de un flujo de trabajo general que inicia con la limpieza de los datos, seguida de la estandarización de textos, la búsqueda en bloques, la clasificación y su evaluación correspondiente. Mediante ejemplos, el lector encontrará un análisis general de la metodología aplicada en cada etapa, cuyo propósito es estandarizar los conjuntos de datos para minimizar los errores de emparejamiento.

Habiendo realizado el emparejamiento de ambos conjuntos de datos, el capítulo cinco presenta diferentes estimadores de muestreo en el sistema de estimación dual con la correspondiente aplicación de los factores de expansión y la metodología adecuada para estimar las varianzas y los correspondientes intervalos de confianza. Además, aborda la incorporación de modelos estadísticos que aportan una mayor flexibilidad cuando algunos de los supuestos del sistema de estimación dual no se cumplen.

Con el propósito de obtener una comprensión más profunda de los factores asociados a la omisión censal, el sexto capítulo se centra en la construcción y análisis de modelos estadísticos basados en regresión logística. Estos modelos permiten identificar las características demográficas, sociales y territoriales que influyen en la probabilidad de que una persona u hogar sea omitido en el censo, proporcionando así una herramienta valiosa para mejorar la cobertura en futuras operaciones censales.

El séptimo capítulo presenta las principales conclusiones y recomendaciones del documento, destacando los aprendizajes metodológicos y las implicaciones prácticas de los resultados obtenidos, con miras a fortalecer los procesos de evaluación y planificación censal.

I. Marco conceptual: el sistema de estimación dual

Este capítulo tiene como propósito definir las condiciones bajo las cuales es apropiado aplicar el sistema de estimación dual, así como los supuestos necesarios para que este método produzca estimaciones insesgadas y precisas.

A. Planteamiento del problema

Para realizar un análisis estadístico adecuado de las encuestas de posenumeración censal (EPC), concebidas como instrumentos destinados a medir la omisión censal, es fundamental comprender que todo el proceso inferencial se sustenta en el sistema de estimación dual, núcleo metodológico de este enfoque. Dicho sistema tiene su origen en los modelos de captura y recaptura desarrollados desde el siglo XVII, con formulaciones modernas a partir de Petersen (1896), Lincoln (1930) y Schnabel (1938), así como su primera aplicación a eventos vitales humanos en el estudio clásico de Sekar y Deming (1949).

Considere una población humana U, de tamaño N, el cual es fijo pero desconocido y es precisamente el parámetro de interés sobre el cual se requiere una inferencia precisa. En una primera instancia, se supone que se realiza un censo de la población en un momento específico en el tiempo, y que el censo intenta enumerar a cada individuo. Sin embargo, por diversas razones, algunos individuos no son contados en el censo. Por ende, la diferencia entre el conteo censal y N se denotará como el error de cobertura.

Una de las principales complicaciones del error de cobertura es que su magnitud no puede determinarse únicamente a partir de los datos del censo. Para cuantificar este error, es imprescindible disponer de información adicional, la cual se obtiene generalmente mediante una encuesta por muestreo aplicada a la misma población objetivo. Esta encuesta se realiza habitualmente después del censo, utilizando el mismo período de referencia temporal y permite estimar la magnitud del error de cobertura censal proporcionando así una medida más precisa y ajustada de la población real.

Inicialmente, supóngase que la encuesta constituye una enumeración completa de toda la población, es decir, que cada individuo ha sido registrado, aunque en la práctica esto nunca ocurre[1]. Esta suposición, aunque idealizada, resulta fundamental para poder explicar y comprender con claridad las propiedades estadísticas del sistema de estimación dual, ya que permite analizar su comportamiento bajo condiciones conocidas y controlar los elementos esenciales del modelo.

B. Condiciones y supuestos

El modelo del error de cobertura descansa bajo un número de supuestos que son imprescindibles a la hora de utilizar una encuesta de posenumeración como instrumento fiable para la medición del error de cobertura en un censo. A continuación, se realiza un listado exhaustivo de ellos.

Cierre poblacional

Este supuesto, también conocido como cerramiento demográfico, plantea que la población U es cerrada y de tamaño fijo N. En la práctica, esto implica que el período de referencia del censo está bien definido; es decir que el censo se lleva a cabo en un intervalo de tiempo específico y claramente establecido. Este período es crucial para garantizar que todos los datos recolectados se refieran a la misma fecha o intervalo de tiempo, evitando así inconsistencias y errores en la estimación de la población. Como consecuencia, se asume que no existen incorporaciones ni pérdidas durante el período de referencia, es decir, que no ocurren nacimientos ni defunciones ni cualquier tipo de migración; de modo que no se agregan ni se restan individuos a la población.

Estructura multinomial

Con este supuesto se asume que el evento conjunto de que un individuo esté o no esté en el censo y esté o no en la encuesta se puede modelar correctamente usando una distribución multinomial justo como lo muestra el cuadro 1.

Cuadro 1
Distribución conjunta de la población censada y encuestada

	En la encuesta	Fuera de la encuesta	Total
En el censo	p_{11}	p_{12}	p_{1+}
Fuera del censo	p_{21}	p_{22}	p_{2+}
Total	p_{+1}	p_{+2}	1

Fuente: Elaboración propia.

En donde p_{11} denota la probabilidad de que un individuo sea encontrado en el censo y en la encuesta, p_{12} denota la probabilidad de que un individuo sea encontrado en el censo, pero no en la encuesta, p_{21} denota la probabilidad de que un individuo no sea encontrado en el censo, pero sí en la encuesta, p_{22} denota la probabilidad de que un individuo no sea encontrado ni en el censo, ni en la encuesta. Asimismo, en términos de las probabilidades marginales, se definen las siguientes cantidades: p_{+1} es la probabilidad de que un individuo sea correctamente encontrado en el censo. Finalmente, p_{+1} es la probabilidad de que un individuo sea correctamente encontrado en la encuesta.

[1] Es importante señalar que, en la realidad, las encuestas solo cubren una fracción de la población. En los capítulos siguientes se introducirán los métodos y ajustes necesarios para adaptar la inferencia estadística a situaciones reales, garantizando que los estimadores producidos por la encuesta sean lo más precisos y confiables posible.

Esto significa que cada individuo tiene la posibilidad de encontrarse en cualquiera de los cuatro estados definidos por la tabla anterior, pero que, al momento de la recolección de los datos, sólo puede ser clasificado en uno de ellos, sin posibilidad de pertenecer a más de un estado simultáneamente.

Independencia autónoma

Este supuesto plantea que el censo y la encuesta se generan como resultado de N ensayos mutuamente independientes. Cada ensayo representa a un individuo de la población real U. A partir de la recolección de los datos, se obtiene la clasificación indicada en el cuadro 2.

Cuadro 2
Matriz de conteos del sistema de estimación dual

	En la encuesta	Fuera de la encuesta	Total
En el censo	N_{11}	N_{12}	N_{1+}
Fuera del censo	N_{21}	N_{22}	N_{2+}
Total	N_{+1}	N_{+2}	N_{++}

Fuente: Elaboración propia.

Note que $N_{ab} = \sum_{k \in U} x_{kab}$, donde x_{kab} es una variable aleatoria dicotómica que señala si el individuo k pertenece a la celda (a, b) de la tabla $(a, b = 1, 2, +)$. Bajo este esquema inicial, se tienen las siguientes consideraciones:

- El conteo del censo N_{1+} se considera observable. Como se verá más adelante, esta cantidad debe estar depurada de las enumeraciones erróneas y de imputaciones.
- Los valores N_{11}, N_{12} y N_{21} se consideran observables a partir de los datos de la encuesta y el emparejamiento con el censo.
- Los valores N_{22}, y el tamaño de la población de interés $N = N_{++}$, se consideran desconocidos y deben estimarse con base en el modelo.

Nótese que, bajo este marco conceptual, el conteo del censo N_{1+} define una variable aleatoria con media $E(N_{1+}) = N \cdot p_{1+}$ y varianza $V(N_{1+}) = N \cdot p_{1+} \cdot (1 - p_{1+})$.

Independencia causal

Según este supuesto, se considera que el evento de ser incluido en el censo es independiente del evento de ser incluido en la encuesta. Como consecuencia de este supuesto, la razón de productos cruzados de probabilidades, conocida tradicionalmente como Razón de Odds, satisface la siguiente relación:

$$\frac{p_{11} \cdot p_{22}}{p_{12} \cdot p_{21}} = 1$$

Este resultado se obtiene porque la probabilidad conjunta de que un individuo se ubique en una celda específica de la matriz de conteos puede factorizarse de la siguiente manera:

$$p_{11} = Pr(\textit{individuo está en el censo y en la encuesta}) = p_{1+} \cdot p_{+1}$$

Similarmente, se tiene que

$$p_{12} = p_{1+} \cdot (1 - p_{+1})$$
$$p_{21} = (1 - p_{1+}) \cdot p_{+1}$$
$$p_{22} = (1 - p_{1+}) \cdot (1 - p_{+1})$$

Sustituyendo adecuadamente en la razón de productos cruzados, se confirma que:

$$\frac{p_{11} \cdot p_{22}}{p_{12} \cdot p_{21}} = \frac{p_{1+} \cdot p_{+1} \cdot (1 - p_{1+}) \cdot (1 - p_{+1})}{p_{1+} \cdot (1 - p_{+1}) \cdot (1 - p_{1+}) \cdot p_{+1}} = 1$$

Por otro lado, la dependencia causal, como señala Bureau (2022), es un fenómeno que ocurre cuando la inclusión o exclusión de un individuo en el censo influye en su probabilidad de ser incluido en la encuesta. Este tipo de dependencia puede generar sesgos en los datos y afectar la calidad de las estimaciones, lo que a su vez puede comprometer la validez de las conclusiones derivadas de estos estudios. Por ello, es fundamental implementar estrategias que mitiguen este riesgo y aseguren la independencia operativa entre ambos sistemas.

Una de las medidas clave para lograr esta independencia operativa es garantizar que el personal involucrado en la recolección de datos de la encuesta no participe en las mismas áreas geográficas o comunidades donde trabajaron durante el censo. Esto reduce la posibilidad de que los encuestadores influyan en las respuestas de los individuos basándose en interacciones previas o en información recopilada durante el censo. Además, al evitar la superposición de personal, se minimiza el riesgo de que los encuestados asocien ambas actividades, lo que podría alterar su disposición a participar o la veracidad de sus respuestas.

Otra estrategia importante es asegurar que las entrevistas de la encuesta se realicen después de que las operaciones del censo hayan concluido en un área específica. Esto permite que los procesos de recolección de datos no se solapen temporalmente, lo que reduce la posibilidad de que los resultados de una actividad afecten directa o indirectamente a la otra. Por ejemplo, si un individuo ha sido contactado recientemente para el censo, podría sentirse menos motivado a participar en la encuesta. Separar prudencialmente ambas operaciones ayuda a mantener la independencia de las respuestas. Finalmente, es importante restringir el acceso del personal del censo a la información sobre la muestra de la encuesta, y viceversa. El personal de la encuesta no debe conocer los resultados del censo durante la recolección de datos, ya que dicha información podría sesgar su enfoque y la interpretación de las respuestas al instrumento.

En América Latina, las EPC no son una realidad en todos los países de la región, y el principio de independencia entre el censo y la EPC puede verse vulnerado por diversos factores operativos. Puede existir dependencia entre los levantamientos debido a que ambos operativos enfrentan dificultades de cobertura similares en asentamientos informales, zonas rurales dispersas o áreas con problemas de seguridad. También puede surgir por la influencia del operativo censal en la EPC, cuando parte del personal de campo ha participado en actividades censales previas. Otro factor es la dependencia en la respuesta de los hogares, quienes suelen recordar la visita censal y repetir información de manera similar en ambas entrevistas. Finalmente, la alta movilidad interna incrementa simultáneamente la probabilidad de omisión en ambas listas. Todas estas situaciones generan correlación entre los procesos de captura y afectan el cumplimiento estricto del supuesto de independencia.

Emparejamiento

Bajo la premisa de este supuesto, se considera que es posible realizar un emparejamiento preciso entre los resultados de la encuesta y los del censo. En otras palabras, se puede identificar de manera exacta y sin errores cuáles individuos registrados en la encuesta también figuran en los registros del censo y

cuáles no. Este emparejamiento correcto resulta crucial para evaluar la cobertura del censo y para ajustar las estimaciones de la población total, garantizando que los datos sean lo más precisos y completos posible.

Como inevitablemente habrá algún grado de no respuesta en el censo y en la encuesta (esto es, que algunos individuos no serán contactados o no proporcionarán la información solicitada), es fundamental recopilar suficiente información auxiliar sobre los no respondientes. Esta información puede incluir datos como nombres, direcciones, fechas de nacimiento y otros identificadores únicos que permitan una correcta identificación de los individuos. En la práctica, se implementan procedimientos específicos para asegurar que la información recopilada sea lo suficientemente detallada y precisa para permitir un emparejamiento exacto entre los datos del censo y los de la encuesta. Este emparejamiento es crucial para evaluar la cobertura del censo y ajustar las estimaciones de la población total.

Ausencia de eventos espurios

Conforme a este supuesto, tanto el censo como la encuesta deben estar libres de incidencias espurias o falsas, o que estas hayan sido eliminadas antes de realizar las estimaciones. Esto implica que se han tomado medidas para evitar cualquier tipo de error en el registro de los resultados tanto del censo como de la encuesta. En la práctica, esto significa que se han implementado procedimientos rigurosos para identificar y corregir cualquier anomalía en los datos. Algunos de los eventos espurios más importantes que pueden ocurrir incluyen:

- Duplicaciones en la lista del censo. Esto ocurre cuando un individuo es contado más de una vez en el censo, lo que puede inflar artificialmente el tamaño de la población.
- Registros de casos inexistentes. Estos son registros de individuos que no existen en realidad, pero que han sido incluidos erróneamente en el censo o en la encuesta. Esto puede suceder debido a errores de entrada de datos o malentendidos en la recolección de información.
- Casos no pertinentes. Estos son individuos que no deberían haber sido incluidos en el censo debido a que no cumplen con los criterios del período de referencia. Un ejemplo común es el registro de un individuo que nació después del período de referencia del censo, lo que resulta en una inclusión incorrecta en los datos.

Para asegurar la precisión de las estimaciones, es crucial que estos eventos espurios sean identificados y eliminados antes de proceder con el análisis de los datos.

Ausencia de imputaciones en la base censal

El presente documento se desarrolla bajo el supuesto fundamental de que la base de datos censal utilizada se encuentra libre de imputaciones de individuos o registros completos. Este supuesto es esencial para garantizar la validez de los resultados obtenidos, dado que la inclusión de registros imputados puede introducir sesgos en las estimaciones del total poblacional y en la medición de las omisiones censales. En consecuencia, no se recomienda en ningún caso realizar estimaciones derivadas (particularmente las referidas a los tamaños poblacionales o a indicadores de cobertura) a partir de bases de datos que hayan incorporado imputaciones de individuos completos.

Tal práctica comprometería la integridad de los indicadores de error de cobertura y distorsionaría la interpretación de los resultados de la EPC, cuyo propósito es precisamente evaluar la calidad del operativo censal a partir de información empíricamente observada y verificable.

Posestratificación

La posestratificación es una técnica estadística que permite ajustar las estimaciones de la población dividiéndola en subgrupos homogéneos, basados en variables categóricas. Esta técnica mejora la precisión y la validez de las estimaciones al considerar las diferencias dentro de la población. Por ejemplo, una forma común de posestratificación es por edad. En este caso, la población se divide en diferentes grupos de edad, como niños, adolescentes, adultos jóvenes, adultos de mediana edad y personas mayores. Para cada uno de estos grupos de edad, se realizan estimaciones específicas de la población. Estas estimaciones se basan en los datos recolectados tanto en el censo como en la encuesta. Una vez obtenidas las estimaciones específicas por edad, se agregan para calcular una estimación total de la población.

Bajo este supuesto, se considera posible identificar subgrupos de interés mediante variables demográficas, geográficas o socioeconómicas, como edad, sexo, etnia, nivel educativo o región geográfica. Es fundamental que todas las variables utilizadas para la posestratificación estén correctamente registradas para cada individuo, tanto en el censo como en la encuesta, con el fin de garantizar la validez y precisión de los análisis. Además, es común y recomendable aplicar algún tipo de posestratificación al estimar el tamaño real de la población, ya que contribuye a obtener resultados más precisos y representativos.

C. Inferencia

En esta sección se desarrollan algunos resultados probabilísticos que determinan el tratamiento inferencial del problema en cuestión. El objetivo principal de la inferencia es estimar el tamaño total de la población (N_{++}), a partir de la información combinada de dos fuentes complementarias: el censo y la encuesta. La primera fuente, el censo, captura a N_{1+} individuos, mientras que la encuesta permite identificar a N_{+1} individuos. Al comparar y combinar ambas fuentes se busca obtener una estimación más completa y confiable del total poblacional.

Una de las consecuencias del supuesto de la distribución multinomial, es que el evento de que un individuo sea registrado en alguna de estas fuentes puede modelarse condicionalmente como un proceso estocástico de tipo Bernoulli. Esto significa que los conteos N_{11}, N_{1+}y N_{+1} se pueden tratar como variables aleatorias binomiales, ya que resultan de la suma de múltiples eventos independientes de tipo Bernoulli. Esta formulación permite utilizar propiedades probabilísticas conocidas para derivar estimaciones y evaluar la variabilidad de los resultados, constituyendo la base teórica para la inferencia sobre el tamaño total de la población.

Luego, las siguientes variables aleatorias siguen distribuciones binomiales condicionales:

$$N_{1+} \sim Bin(N_{++}, p_{1+}), \quad N_{+1} \sim Bin(N_{++}, p_{+1}), \quad N_{11} \sim Bin(N_{++}, p_{11})$$

Una vez que los datos hayan sido recolectados y clasificados bajo este esquema, los estimadores para las probabilidades de interés toman la siguiente forma:

$$\tilde{p}_{11} = \frac{N_{11}}{N_{++}}, \quad \tilde{p}_{1+} = \frac{N_{1+}}{N_{++}}, \quad \tilde{p}_{+1} = \frac{N_{+1}}{N_{++}}$$

Al asumir independencia entre la captura en el censo y la captura en la encuesta, entonces se tiene que $\tilde{p}_{11} = \tilde{p}_{1+} \cdot \tilde{p}_{+1}$. Por consiguiente:

$$\frac{N_{11}}{N_{++}} = \frac{N_{1+}}{N_{++}} \cdot \frac{N_{+1}}{N_{++}}$$

Finalmente, al despejar convenientemente, se encuentra que el estimador del sistema dual para el total poblacional N_{++} está dado por

$$\widetilde{N}_{++} = \frac{N_{1+} \cdot N_{+1}}{N_{11}}$$

A partir de este resultado, podemos reemplazar en las expresiones $\tilde{p}_{11}$, $\tilde{p}_{1+}$ y $\tilde{p}_{+1}$ para obtener estimadores de máxima verosimilitud para todas las restantes probabilidades de interés:

$$\tilde{p}_{11} = \frac{N_{11}}{\widetilde{N}_{++}} = \frac{N_{11}^2}{N_{1+} \cdot N_{+1}}$$

$$\tilde{p}_{1+} = \frac{N_{1+}}{\widetilde{N}_{++}} = \frac{N_{11}}{N_{+1}}$$

$$\tilde{p}_{+1} = \frac{N_{+1}}{\widetilde{N}_{++}} = \frac{N_{11}}{N_{1+}}$$

Wolter (1986, sección 2.4) plantea un esquema conjunto que induce estos mismos estimadores a partir de la función de verosimilitud asociada al modelo, la cual está dada por la siguiente expresión:

$$L(N, p_{i+}, p_{+i}) = \binom{N}{x_{11}, x_{12}, x_{21}} \cdot p_{1+}^{x_{1+}} \cdot (1 - p_{1+})^{N - x_{1+}} \cdot p_{+1}^{x_{+1}} \cdot (1 - p_{+1})^{N - x_{+1}}.$$

Los estimadores de máxima verosimilitud de los parámetros de interés se encuentran maximizando la anterior expresión sujeta a las restricciones pertinentes sobre las sumas de las probabilidades. Para demostrar que el estimador $\widetilde{N}_{++}$ es insesgado, se debe verificar que $E\left[\widetilde{N}_{++}\right] = N_{++}$. En primer lugar, por la propiedad de la esperanza en distribuciones binomiales, se tiene que:

$$E[N_{1+}] = N_{++} \cdot p_{1+}, \quad E[N_{+1}] = N_{++} \cdot p_{+1}, \quad E[N_{11}] = N_{++} \cdot p_{11}$$

Ahora, la esperanza del estimador toma la siguiente forma:

$$E\left[\widetilde{N}_{++}\right] = E\left[\frac{N_{1+} \cdot N_{+1}}{N_{11}}\right]$$

En primera instancia, como N_{1+} y N_{+1} son variables aleatorias, es necesario apelar a las propiedades de la esperanza condicional, de la siguiente manera:

$$E\left[\widetilde{N}_{++}\right] = E\left[E\left(\frac{N_{1+} \cdot N_{+1}}{N_{11}} \mid N_{1+}, N_{+1}\right)\right]$$

Además, como N_{11} también es una variable aleatoria, entonces bajo condiciones de regularidad que permitan utilizar la expansión de Taylor, es posible aproximar la esperanza de este cociente al cociente de las esperanzas (Casella y Berger 2002). De esta forma, se tiene que:

$$E\left(\frac{N_{1+} \cdot N_{+1}}{N_{11}} \mid N_{1+}, N_{+1}\right) = \frac{E(N_{1+} \cdot N_{+1} \mid N_{1+}, N_{+1})}{E(N_{11} \mid N_{1+}, N_{+1})}$$

Dado que N_{1+} y N_{+1} son independientes, entonces, por las propiedades de la esperanza, se tiene que $E[N_{1+} \cdot N_{+1}] = E[N_{1+}] \cdot E[N_{+1}]$. Por ende, reemplazando convenientemente:

$$E\left[\widetilde{N}_{++}\right] = \frac{N_{++}^2 \cdot p_{1+} \cdot p_{+1}}{N_{++} \cdot p_{1+} \cdot p_{+1}} = N_{++} = N$$

Es decir, el estimador $\widetilde{N}_{++}$ es insesgado para el total poblacional N_{++}. Por último, Wolter (1986) también afirma que la varianza del estimador puede ser estimada mediante la siguiente expresión:

$$\tilde{V}\left[\widetilde{N}_{++}\right] = \frac{N_{1+} \cdot N_{+1} \cdot N_{12} \cdot N_{21}}{N_{11}^3}$$

El estimador $\widetilde{N}_{++}$ ha sido ampliamente utilizado en estudios de captura y recaptura para estimar el tamaño de una población. Este método fue desarrollado por el biólogo danés Carl Georg Johannes Petersen (Petersen 1896) y más tarde popularizado por Sekar y Deming (1949) para estimar tasas de nacimientos y defunciones, así como la cobertura de los registros vitales.

Es importante resaltar que todos los indicadores utilizados para evaluar la cobertura censal (tanto las medidas que comparan la población real con la censada, como aquellas que evalúan la calidad del emparejamiento o la magnitud de los errores de inclusión y omisión) dependen directamente del tamaño total de la población (N_{++}). Por lo tanto, el presente documento se enfocará en desarrollar una inferencia correcta sobre $\widetilde{N}_{++}$, dado que su estimación exacta y precisa constituye la base para el cálculo y la interpretación de todos los indicadores de cobertura y omisión censal. A continuación, se presenta una lista de indicadores relevantes en una EPC:

- El error neto de cobertura (ENC), que mide la diferencia entre la población real y la censada:

$$ENC = N_{++} - N_{1+}$$

- A nivel relativo, se define la tasa del error neto de cobertura ($TENC$) como

$$TENCR = \frac{N_{++} - N_{1+}}{N_{++}}$$

- Una omisión es un evento que denota que una persona no fue correctamente contada o registrada durante el proceso censal. El total de omisiones (O) corresponde al número total de personas que no fueron contadas (ENC) más las que el censo enumeró erróneamente (N_{EE}):

$$O = N_{++} - N_{1+} + N_{EE}$$

- La tasa de omisión (TO), que se define como la proporción de la población total que no fue correctamente censada:

$$TO = \frac{O}{N_{++}} = \frac{N_{++} - N_{1+} + N_{EE}}{N_{++}}$$

- La tasa de emparejamiento (TE), que refleja la proporción de registros del censo que fueron correctamente vinculados con la EPC:

$$TE = \frac{N_{11}}{N_{1+}}$$

- La tasa de enumeraciones erróneas (TEE), que indica la proporción de personas contadas en el censo que no pertenecen realmente a la población objetivo:

$$TEE = \frac{N_{EE}}{N_{1+}}$$

En síntesis, el sistema de estimación dual ofrece un marco estadístico sólido para estimar el tamaño total de una población a partir de la combinación de los datos del censo y de una EPC. Su fundamento radica en modelar la probabilidad de clasificación de cada individuo como un proceso aleatorio. En la práctica, este método proporciona una estimación insesgada del total poblacional siempre que se cumplan los supuestos básicos que han sido mencionados anteriormente. De esta manera, la aplicación del sistema de estimación dual en un censo real permite cuantificar la magnitud de la omisión censal para ajustar las cifras poblacionales de forma coherente.

II. Planeación: cuestionario, muestreo y operativo

La planificación de una Encuesta de Posenumeración Censal (EPC) constituye una etapa decisiva que influye de manera determinante en la calidad final de la medición de la cobertura del censo, considerado el operativo estadístico y logístico más complejo que ejecutan los Institutos y Oficinas Nacionales de Estadística. Esta fase no solo implica el desarrollo de un diseño técnico sólido, que contemple una asignación eficiente de recursos humanos y materiales, un plan de muestreo riguroso y procedimientos operativos bien estructurados, sino también la formulación previa de objetivos claramente definidos y sin ambigüedades, que orienten la organización y permitan la implementación coherente de todas las actividades involucradas. Asimismo, es fundamental que la planificación de la EPC se conciba como un componente integral del proceso censal, lo que implica su coordinación estrecha con la planificación del propio censo. Idealmente, su planeación debe avanzar de forma paralela desde etapas tempranas, garantizando la disponibilidad oportuna de los recursos financieros, técnicos y humanos necesarios para su desarrollo exitoso (UN 2010; Hogan 2003).

Como se ha señalado previamente, el propósito fundamental de la Encuesta de Posenumeración Censal (EPC) es obtener estimaciones precisas de los errores de cobertura censal, a través de un diseño de muestreo independiente y del uso de metodologías rigurosas, entre ellas la estimación por sistema dual. Sin embargo, su utilidad va más allá de la medición de cobertura: la EPC también constituye una herramienta clave para la evaluación integral de la calidad de los resultados censales, al permitir identificar fuentes de error y sesgos en distintas etapas del operativo. Además, este ejercicio ofrece una oportunidad de aprendizaje institucional que fortalece las capacidades técnicas y operativas de las oficinas nacionales de estadística, aportando evidencia valiosa para optimizar los procedimientos de recolección, control de calidad y procesamiento de datos en futuros censos.

En este contexto, la EPC persigue varios propósitos complementarios orientados a fortalecer la calidad y la confiabilidad de los resultados censales. Entre sus principales objetivos se encuentra la cuantificación de la subcobertura y la sobrecobertura de personas, hogares y viviendas, lo que permite calcular no solo el conteo neto poblacional, sino también descomponer los componentes del error de cobertura, tales como las enumeraciones erróneas, las duplicaciones y las omisiones. Asimismo, busca que estas estimaciones se pueden desagregar a nivel subnacional, en diferentes niveles territoriales o

incluso para grupos poblacionales específicos, utilizando herramientas estadísticas como la posestratificación o los modelos de regresión logística, que posibilitan un análisis más detallado y desagregado de los errores censales. Por ejemplo, la EPC podría proporcionar indicadores de cobertura por grupos demográficos, los cuales resultan esenciales para las tareas de conciliación, validación y ajuste de los resultados del censo. A lo anterior, se suma su función de evaluar la coherencia y consistencia de las variables de contenido, tales como sexo, edad, estado civil o relación de parentesco con la persona de referencia o jefe del hogar, lo que permite analizar el grado de concordancia entre las respuestas y, en consecuencia, valorar la calidad global de la información recolectada.

Sin embargo, la EPC no solo aporta una medida del error de cobertura, sino que también constituye una herramienta clave para fortalecer el sistema estadístico nacional, apoyando a las Oficinas Nacionales de Estadística (ONE) en la actualización de marcos muestrales y en la retroalimentación metodológica y operativa. También es posible evaluar la idoneidad de las Unidades Primarias de Muestreo (UPM) como marcos de referencia para futuras encuestas de hogares; e identificar prácticas operativas o metodológicas que requieren mejoras en censos posteriores, proporcionando insumos para optimizar la planificación y ejecución de futuros operativos censales.

En consecuencia, y considerando su relevancia estratégica, la EPC demanda una asignación adecuada y suficiente de recursos financieros y humanos, incluyendo tanto sensibilizadores, enumeradores, supervisores y coordinadores calificados como personal profesional capacitado para realizar el emparejamiento de información; analistas con formación estadística; así como un sistema eficiente de control operativo y aseguramiento de la calidad que acompañe todas las etapas del proceso de recolección de datos, garantizando la fiabilidad y precisión de los resultados.

Al comenzar la planificación técnica de la EPC, resulta fundamental definir con claridad los objetivos específicos que se pretenden alcanzar. Estos objetivos deben traducirse en planes detallados, asignados a subgrupos técnicos de planificación, cada uno responsable de un componente particular de la operación. Entre estos se incluyen: el grupo temático, encargado del diseño del cuestionario; el grupo de diseño y ejecución de la muestra; el grupo logístico, responsable de coordinar los aspectos operativos del proceso; y el grupo analítico, encargado del emparejamiento de información y la estimación por sistema dual. Cada grupo debe elaborar su propio plan, siempre dentro de un marco de independencia metodológica respecto a los procesos del censo, coordinando aspectos comunes como la cartografía.

En este capítulo nos centraremos, de manera no exhaustiva, en tres aspectos fundamentales para la planificación de la EPC: el diseño del cuestionario, el diseño de la muestra y el operativo de recolección. Aunque los elementos presentados aquí no abarcan todos los aspectos posibles en cada tema, sí representan una guía para orientar la planificación de la encuesta. Tener en cuenta estos aspectos permite anticipar posibles dificultades operativas y metodológicas, asegurar que los objetivos de recolección de datos se cumplan de manera coherente, y sentar las bases para un procesamiento y análisis de información confiables.

A. Diseño del cuestionario

El cuestionario de la encuesta constituye el vínculo entre la información proporcionada por los hogares tanto en la EPC como en el censo. Debe diseñarse tomando como referencia el cuestionario censal definitivo y adaptarse al procedimiento de emparejamiento que se aplicará para la reconstrucción de los hogares (capítulo III). Su elaboración es esencial, ya que convierte las necesidades de información en preguntas operativas y proporciona la base para el procesamiento y análisis de los datos (UN 2010; Baffour, King, y Valente 2013). El cuestionario debe ser estructurado y claro, acompañado de instrucciones precisas para los enumeradores. Algunas características deseables son:

- Pertinencia: asegurar la recopilación de datos que respondan efectivamente a las necesidades de los usuarios y a los objetivos del estudio.
- Eficiencia: facilitar la labor de recolección, procesamiento y tabulación, evitando la inclusión de información redundante o innecesaria.
- Claridad: formular preguntas de lectura sencilla y comprensión inmediata, sustentadas en definiciones operativas precisas.
- Calidad: propiciar la obtención de datos consistentes y la generación de estimaciones confiables, que respalden la validez de los resultados.

En este sentido, la prueba piloto del cuestionario es esencial para su validación final, ya que permite comprobar su claridad, coherencia y adecuación operativa antes de la recolección definitiva. Durante la aplicación de la EPC, el encuestador debe elaborar un listado de individuos independiente al censo, que incluya tanto a las personas que residen actualmente en el hogar como aquellas personas que vivían allí el día del censo, aunque ya no formen parte del hogar. Para garantizar un emparejamiento más preciso y expedito, la información básica que se debe registrar en la encuesta debería comprender el nombre, los apellidos, el sexo, la fecha de nacimiento, la edad, el parentesco con el jefe del hogar y, de ser posible, el número de identificación personal.

Las preguntas en el cuestionario deben orientarse a establecer otros posibles lugares de residencia o permanencia de los individuos en el día del censo, con el fin de identificar si la persona se trasladó, estaba temporalmente en otra vivienda o reside de manera alternada entre diferentes direcciones. En este último caso, si el censo es de derecho, podría ser necesario indagar cuánto tiempo permanece en cada lugar para determinar su residencia habitual, es decir, el sitio donde debería haber sido contabilizada en el censo. Asimismo, cuando se confirme que una persona se trasladó de domicilio, el encuestador debe procurar obtener su nueva dirección. En este sentido, el cuestionario debería ser redactado con el fin de:

- Verificar si las personas que residen en los hogares de la muestra fueron efectivamente enumeradas en el censo.
- Identificar posibles errores de enumeración (duplicaciones, omisiones, entre otros).
- Determinar si las personas se han trasladado de la vivienda desde la fecha del censo y, en caso afirmativo, recopilar la nueva dirección de residencia.

Con el fin de actualizar la información y facilitar el emparejamiento adecuado entre los registros censales y los de la EPC, las direcciones recopiladas durante la entrevista pueden clasificarse en direcciones de traslado y direcciones alternas. Las primeras corresponden a casos en los que una persona que actualmente reside en la vivienda de la EPC habitaba en otro lugar el día del censo, mientras que las segundas se refieren a situaciones en las que una persona reside temporalmente en otra vivienda por motivos laborales, de estudio, servicio militar u otras razones, manteniendo, no obstante, su vínculo con la vivienda de origen en el censo. Así que, la encuesta debe diseñarse para recopilar la mayor cantidad posible de direcciones alternas con el fin de detectar todos los lugares donde una persona pudo haber sido contada en el censo y determinar si fue contabilizada más de una vez. Es posible que el encuestado no pueda proporcionar la dirección completa, en esos casos se debe intentar obtener cualquier información de referencia cercana o identificativa, lo que facilitará el emparejamiento final entre los registros de la EPC y los del censo.

Aunque pueda parecer obvio, es fundamental recalcar que a cada persona se le debe asignar correctamente su departamento y municipio de residencia, tanto en la EPC como en la fecha del censo. Esta información es esencial para determinar el lugar donde la persona debió haber sido contabilizada el día del operativo censal y para verificar posibles desplazamientos o cambios de residencia. Por ello,

se recomienda incorporar una codificación específica que identifique si la persona se ha trasladado desde entonces. Asimismo, es conveniente incluir una casilla de observaciones donde el enumerador pueda registrar cualquier información adicional o referencia relevante, que facilite el trabajo de los revisores clericales durante el proceso de emparejamiento, el cual debe estar diseñado para determinar si las personas de la EPC fueron enumeradas en el censo, tuvieron errores de enumeración u omisiones. En ese sentido, el cuestionario debería permitir clasificar correctamente a las personas según su condición de residencia en una de las siguientes categorías:

- Residente permanente (*non-mover*): persona que residía en un hogar particular en la fecha del censo y que aún reside allí en la fecha del EPC.
- Residente de salida (*out-mover*): persona que residía en el hogar en la fecha del censo pero que ya no reside en el hogar en la fecha del EPC.
- Residente de entrada (*in-mover*): persona que reside en el hogar en la fecha del EPC pero que no residía allí en la fecha del censo.
- Fuera del universo censal (*out-of-scope*): persona que no pertenece a la población objetivo en la fecha del censo. Por ejemplo, personas nacidas después del censo, fallecidas antes del censo, o que residían fuera del país.

El cuadro 3 resume de manera no exhaustiva algunos casos comunes de movimientos y su respectiva clasificación según su situación de residencia en el censo y en la EPC.

Cuadro 3
Categorización de personas para el emparejamiento entre el Censo y la EPC

Persona que …	Situación en el censo	Situación en la EPC	Clasificación
seguía viviendo en el hogar	Estaba presente	Sigue presente	*Non-mover*
trabaja o estudia, pero mantiene la misma residencia habitual	Estaba presente	Sigue presente	*Non-mover*
falleció después del censo	Estaba presente	Ya no reside en el hogar	*Out-mover*
se mudó a otro hogar	Estaba presente	Ya no reside en el hogar	*Out-mover*
se fue al extranjero después del censo	Estaba presente	Ya no reside en el hogar	*Out-mover*
fue internado en una institución (hospital, cárcel, hogar de ancianos)	Estaba presente	Ya no reside en el hogar	*Out-mover*
se mudó al hogar después del censo	No estaba en el hogar	Está presente en el hogar	*In-mover*
que residía en un alojamiento temporal (cuarto alquilado, residencia laboral)	No estaba en el hogar	Está presente en el hogar	*In-mover*
regresó de un viaje largo o del extranjero después del censo	No estaba en el hogar	Está presente en el hogar	*In-mover*
vivía temporalmente en otro lugar por estudios, trabajo o servicio militar	No estaba en el hogar	Está presente en el hogar	*In-mover*
nació después de la fecha del censo	No existía	Está presente en el hogar	*Out-of-scope*
falleció antes de la fecha del censo	Ya no existía	Ya no reside en el hogar	*Out-of-scope*
residía en el extranjero en la fecha del censo	No estaba en el hogar	Está presente en el hogar	*Out-of-scope*
estaba de visita temporal durante el censo	No era residente del hogar	No reside en el hogar	*Out-of-scope*

Fuente: Elaboración propia.

Como la EPC tiene como propósito fundamental evaluar la cobertura del censo, resulta relevante registrar no solo el estado de la persona y su ubicación geográfica, sino también un conjunto más amplio de variables que permitan verificar con precisión si la persona registrada en la EPC corresponde efectivamente a la misma persona del censo. En este sentido, el cuestionario debe incluir variables sociodemográficas tomadas directamente del cuestionario del censo, tales como nombres y apellidos, sexo, fecha de nacimiento y edad exacta, parentesco con el jefe del hogar, estado civil y nivel educativo, con el fin de identificar posibles errores de contenido. También se deben identificar personas adicionales que residían en el hogar durante el censo, pero no fueron mencionadas inicialmente, así como la pertenencia étnica de los integrantes, dado que su inclusión en la EPC puede contribuir a corregir dificultades en la captura de ciertas comunidades y mejorar los ajustes de cobertura.

Además, resulta útil incorporar preguntas de sondeo que permitan detectar a personas que pertenecen a la población objetivo en la fecha del censo, pero podrían haber sido omitidas involuntariamente del listado del hogar, como bebés que ya existían en esa fecha, personas mayores o residentes habituales temporalmente ausentes (hospitalización u otras razones), contribuyendo así a reducir las omisiones y mejorar la calidad del registro. Para evitar estas omisiones, se recomienda incluir preguntas de sondeo. Por ejemplo:

- *"Por favor, dígame los nombres de todas las personas que pasaron la noche del (fecha del EPC) en este hogar."*
- *"De las personas que actualmente residen en este hogar, ¿hay alguien que no estuvo presente la noche del censo, pero debería haberse contado en este hogar según su residencia habitual? Por favor, indique sus nombres."*
- *"¿Hubo alguna persona que sí estuvo en el hogar la noche del censo pero que actualmente no reside aquí? Por favor, indique sus nombres y, si es posible, su nueva dirección o lugar de residencia habitual."*
- *"Para cada persona mencionada, ¿podría confirmar su parentesco con el jefe(a) del hogar, su sexo, fecha de nacimiento y edad?"*

B. El diseño de muestreo

El diseño muestral de una encuesta de posenumeración censal debe ser siempre probabilístico; por lo general estratificado y por conglomerados en una sola etapa (aunque en la práctica pueden existir variantes que se alejen ligeramente de esta generalización debido a las particularidades de cada país). En términos generales, la estratificación se realiza por zona, considerando factores como urbano/rural, región, departamento o estado, con el objetivo de asegurar la representatividad de diferentes subpoblaciones. De forma independiente, las unidades primarias de muestreo (UPM) se definen a partir de las áreas de enumeración (segmentos o sectores cartográficos del censo) y su selección sigue un diseño de muestreo proporcional al número de viviendas, hogares o personas del área. Dentro de cada UPM, se procede a levantar el recuento del número de viviendas y a visitarlas a todas de manera exhaustiva, recolectando la información de todos los hogares y personas que allí residen, garantizando así la cobertura completa del área muestral seleccionada.

Para que el diseño sea verdaderamente probabilístico, es esencial que todas las personas tengan una probabilidad conocida y mayor que cero de ser incluidas en la muestra. La estratificación, además de favorecer la representatividad, permite reducir la varianza de los estimadores y optimizar el tamaño de muestra para cada subpoblación de interés. Por su parte, las UPM, que suelen corresponder a manzanas, sectores censales u otros conglomerados geográficos pequeños, se seleccionan mediante probabilidad proporcional al número de viviendas, hogares o personas y constituyen la base del

muestreo por conglomerados (CEPAL, 2023). En cada UPM seleccionada se registra la información de todos los hogares y personas; en algunos casos puede aplicarse un submuestreo de viviendas si la logística lo requiere. Este diseño combina de manera integrada los procesos de estratificación y selección de conglomerados, introduciendo un efecto de diseño que debe tenerse en cuenta al calcular los errores estándar de los estimadores, debido a la correlación intraclase dentro de los conglomerados.

En este sentido, es importante señalar que, incluso en situaciones con restricciones presupuestales importantes, se desaconseja emular los diseños de muestreo de las encuestas de hogares tradicionales, los cuales se definen en dos etapas de muestreo: seleccionar primero los segmentos o conglomerados y, en una segunda etapa, elegir únicamente las viviendas que estén ocupadas al momento de la recolección de datos. Esta práctica no es recomendable, ya que podría introducir sesgos importantes y comprometer la representatividad de la encuesta. La selección de viviendas basada en la ocupación observada durante el trabajo de campo puede generar errores sistemáticos y afectar la validez de los resultados. Entre las principales limitaciones del muestreo en dos etapas se encuentran:

- Sesgo de cobertura: al seleccionar únicamente viviendas ocupadas durante la recolección, se pueden excluir hogares que estaban ausentes temporalmente, como personas de viaje, hospitalizadas o en actividades laborales fuera del hogar, generando subregistro.
- Incremento del error de estimación: la reducción de la cobertura completa y la selección basada en la ocupación observada puede aumentar los errores estándar y sesgar los resultados de la EPC.
- Dificultad para estimar probabilidades de selección: la probabilidad de inclusión deja de ser conocida y uniforme para todos los miembros de la población objetivo, lo que compromete los principios de un diseño probabilístico.
- Pérdida de representatividad: la muestra final puede no reflejar adecuadamente la estructura real de la población en el censo, afectando la precisión de las estimaciones por estrato o subpoblación.
- Problemas en la conciliación con los registros censales: la comparación entre la EPC y los datos censales se vuelve más compleja, dificultando la identificación precisa de omisiones, duplicaciones o errores de contenido.

Nótese que, si transcurre un intervalo prolongado entre la realización del censo y la EPC, se pueden violar supuestos clave del diseño muestral. Por ejemplo, una vivienda que estaba ocupada durante el censo podría estar deshabitada al momento de la encuesta, debido a migración, desastres naturales u otras circunstancias, lo que le otorga una probabilidad nula de ser seleccionada. De manera similar, las viviendas construidas después del censo podrían albergar hogares que no fueron incluidos en el censo, afectando la cobertura y la representatividad de la EPC.

Por otro lado, debido a que puede existir resistencia a responder debido a que el censo se realizó hace poco tiempo, es importante considerar registros para el control de la cobertura, para ello se recomienda codificar las novedades o incidencias de acuerdo con los códigos de disposición (AAPOR, 2016) en la cual las unidades de observación se codifican como elegibles respondientes, elegibles no respondientes, con elegibilidad desconocida y no elegibles. A continuación, se abordan algunos elementos transversales en la definición de los diseños de muestreo en una EPC.

Marco de muestreo

Como en todo procedimiento de muestreo probabilístico, es imprescindible contar con un marco de muestreo que permita identificar y ubicar todas las unidades que conforman la población objetivo. Este marco constituye la base sobre la cual se seleccionan las UPM y asegura que cada elemento de la

población tenga una probabilidad conocida de ser incluido en la encuesta. Su existencia y correcta definición son fundamentales para garantizar la representatividad de la muestra y la validez de los resultados de la EPC.

Por lo general, el marco de muestreo de una EPC se basa está limitado a las áreas de enumeración que contienen viviendas particulares, excluyendo unidades colectivas como cárceles, hospitales o residencias estudiantiles. Esta exclusión se justifica porque los hogares en alojamientos colectivos no representan la estructura típica de los hogares en la población general, lo que podría sesgar los resultados si se incluyeran. Aunque los censos incluyen tanto viviendas particulares como colectivas, las encuestas posteriores, como las EPC, se basan en el marco de viviendas particulares para garantizar la representatividad de los datos obtenidos, pero además porque la unidad de emparejamiento principal es el hogar.

Dado que la metodología de estimación en estas encuestas (sistema dual) se fundamenta en el supuesto de independencia entre la EPC y el censo, es esencial que la encuesta de cobertura se realice sin utilizar información auxiliar ni resultados provenientes del censo, de manera que los procesos sean completamente independientes. Por ello, el marco de áreas utilizado para seleccionar las UPM debe coincidir con el planeado para la logística del censo, empleando los mismos segmentos o sectores censales como referencia, pero asegurando que el personal encargado de la recolección sea distinto al que participó en el censo. Esta separación garantiza que la EPC funcione como una operación independiente y confiable para evaluar la cobertura censal.

Construcción de las UPM

Este proceso debe derivarse directamente de la definición establecida en el censo. Es fundamental que las áreas cartográficas (segmentos o sectores) utilizados como UPM en la encuesta de posenumeración correspondan a los mismos bloques cartográficos definidos durante el censo, ya que esto asegura la coherencia espacial y administrativa entre ambos ejercicios. Tomar como referencia los mismos segmentos permite mantener la comparabilidad y facilita el control de cobertura de manera precisa.

Como se expondrá en el capítulo IV, el uso de los mismos segmentos cartográficos es esencial para la indexación y el emparejamiento de los registros durante el proceso de análisis de la EPC. Al contar con una correspondencia directa entre los segmentos del censo y los de la EPC, se pueden identificar de manera eficiente los hogares y personas que fueron enumerados en ambas operaciones, garantizando así la validez de las estimaciones y la detección de posibles omisiones o duplicaciones en la información recolectada. Usar este enfoque contribuye a la calidad y confiabilidad de los resultados finales de la encuesta de cobertura.

La exclusión de alojamientos colectivos influye directamente en la definición de las UPM. Al limitar el universo objetivo a las viviendas particulares, las UPM se seleccionan únicamente dentro de áreas geográficas donde residen hogares particulares, evitando incluir instituciones colectivas (cárceles, residencias estudiantiles u hospitales). Esta delimitación asegura que las unidades de muestreo representen de manera adecuada la población de interés, pero también facilita la logística del operativo y permite diseñar procedimientos de selección muestral coherentes con el universo realmente accesible y relevante para la EPC.

Estratificación geográfica

La estratificación en la encuesta se construye con base en dos objetivos principales. El primero está relacionado con la eficiencia del diseño muestral, buscando asegurar una mayor precisión de las estimaciones, mientras que el segundo se vincula con la necesidad de obtener información desagregada para distintos dominios geográficos.

Para cumplir con estos objetivos, los estratos deben considerar áreas geográficas que puedan presentar diferentes niveles de cobertura. Por ejemplo, se pueden diferenciar áreas urbanas, centros poblados, zonas rurales dispersas o regiones con población étnica de difícil acceso, donde es más probable que las omisiones o errores de cobertura sean distintos respecto a otras áreas. Esta diferenciación permite diseñar la muestra de manera que se reduzca la incertidumbre de las estimaciones dentro de cada estrato, optimizando así la eficiencia del muestreo. En estos casos, se recomienda que dentro de cada estrato se aplique una subestratificación adicional, siguiendo los lineamientos previamente definidos, para garantizar que la muestra represente todos los subdominios de interés.

Selección de la muestra

Habiendo definido tanto el marco de muestreo como la correspondiente estratificación, el siguiente paso consiste en realizar el proceso de selección de las UPM. Este procedimiento debe garantizar estimadores insesgados para el sistema dual, así como una alta precisión y eficiencia, generando intervalos de confianza estrechos. En otras palabras, la inclusión de las unidades en la muestra debe basarse en un esquema probabilístico libre de sesgos, capaz además de minimizar la dispersión en el proceso inferencial posterior.

El muestreo probabilístico asigna a cada posible muestra una probabilidad de selección conocida, lo que permite sostener la validez estadística del estudio. Según Särndal, Swensson y Wretman (2003), el diseño de muestreo es el mecanismo mediante el cual el investigador define estas probabilidades. Aunque su asignación es teórica, el equipo técnico debe decidir cuál es la mejor forma de selección y, sobre esa base, elegir el algoritmo de muestreo más adecuado. Posteriormente, se selecciona una única muestra mediante un proceso aleatorio que respete estrictamente la configuración estocástica definida por el diseño. Es importante que todas las probabilidades de selección sean mayores que cero, ya que de lo contrario se compromete la inferencia insesgada, al excluir segmentos del marco muestral. Estas probabilidades son también fundamentales para el cálculo de los factores de expansión, que sustentan todo el proceso de estimación puntual, así como de sus correspondientes errores de muestreo.

Es necesario diferenciar claramente entre el diseño de muestreo y el algoritmo de muestreo. El diseño de muestreo establece las probabilidades de selección tendrán las posibles muestras, mientras que el algoritmo de muestreo se refiere al proceso de selección de una única muestra, respetando las probabilidades establecidas por el diseño. En el caso de la EPC, es fundamental definir ambos componentes de manera previa. Para ello, el equipo técnico debe documentar exhaustivamente cada etapa del muestreo y explicar claramente qué algoritmos de selección se aplicarán, garantizando así transparencia en la selección de las unidades.

Existen muchas formas de seleccionar una muestra y cada una de ellas induce una medida de probabilidad sobre los elementos que conforman la población de interés. En general, asociado a cada esquema particular de muestreo se define una única función que asocia a cada segmento k con una probabilidad de inclusión en la muestra s, definida de la siguiente manera:

$$\pi_k = Pr(k \in s)$$

Si el diseño de muestreo es de tamaño fijo, incluyendo a n individuos, estas probabilidades de inclusión de los hogares cumplirán con las siguientes propiedades

$$\pi_k > 0$$

$$\sum_U \pi_k = n$$

La primera propiedad garantiza que ningún segmento del marco muestral será excluido de la selección inicial. Aunque no todos los segmentos serán finalmente incluidos en la muestra final s, cada

uno tendrá una probabilidad positiva de ser seleccionado mediante el mecanismo aleatorio definido por el diseño de muestreo. Además, el tamaño final de la muestra de segmentos está directamente relacionado con la magnitud de estas probabilidades de inclusión. Por ello, en encuestas con muestras más grandes, cada segmento y hogar recibirá una probabilidad de inclusión más alta, mientras que en encuestas de menor tamaño de muestra estas probabilidades serán proporcionalmente menores.

Cálculo del tamaño de muestra

El tamaño de muestra depende de si se buscan únicamente estimaciones nacionales o también desagregaciones subnacionales. En el primer caso, una muestra más pequeña puede ser suficiente; sin embargo, si se desean resultados por múltiples dominios (área urbana/rural, regiones, provincias u otras unidades subregionales), se requiere un tamaño de muestra más grande, lo que inevitablemente incrementa los costos de la operación estadística (CEPAL 2023).

El tamaño de la muestra se debe calcular para lograr un nivel de precisión requerido con un nivel de confianza. De manera que, es necesario definir los diferentes tipos de error muestral En principio, se define un intervalo de confianza para el parámetro θ, que representa la proporción de personas omitidas en el censo, utilizando su estimador insesgado $\hat{\theta}$. Se supone que $\hat{\theta}$ sigue una distribución normal con media θ y varianza $Var(\hat{\theta})$, de manera que el intervalo de confianza puede expresarse comoo

$$IC(1-\alpha) = \left[\hat{\theta} - z_{1-\frac{\alpha}{2}}\sqrt{Var(\hat{\theta})}, \hat{\theta} + z_{1-\frac{\alpha}{2}}\sqrt{Var(\hat{\theta})}\right]$$

En donde $z_{1-\alpha/2}$ se refiere al cuantil $(1-\alpha/2)$ de una variable aleatoria con distribución normal estándar. Cuando el diseño de muestreo es complejo, es necesario reemplazar el percentil de la distribución normal estándar por el percentil de una distribución t de Student con $N_I - H$ grados de libertad, suponiendo que hay N_I unidades primarias de muestreo y H estratos. Desde la expresión del intervalo de confianza, se define el margen de error, como aquella cantidad que se suma y se resta al estimador insesgado. En este caso, se define como

$$ME = z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}$$

Para determinar el tamaño de la muestra se deben considerar los efectos de la estratificación y la aglomeración de las unidades de muestreo. Una forma sencilla de incorporar este efecto en las expresiones clásicas del muestreo aleatorio simple es la relación de las varianzas en el efecto de diseño:

$$DEFF(\hat{\theta}) = \frac{Var_p(\hat{\theta})}{Var_{MAS}(\hat{\theta})}$$

Esta cifra da cuenta del efecto de aglomeración causado por la utilización de un diseño de muestreo cualquiera (p), frente a un diseño de muestreo aleatorio simple (MAS) en la inferencia de un parámetro de la población finita. Por lo anterior, es posible escribir la varianza del estimador bajo el diseño de muestreo complejo como

$$\begin{aligned} Var_p(\hat{\theta}) &= DEFF(\hat{\theta})\, Var_{MAS}(\hat{\theta}) \\ &= DEFF(\hat{\theta}) \cdot \frac{N^2}{n} \cdot \left(1 - \frac{n}{N}\right) \cdot S^2_{y_U} \end{aligned}$$

Por lo tanto, si al implementar un muestreo aleatorio simple, el tamaño de muestra n_0 es suficiente para conseguir la precisión deseada, entonces el valor del tamaño de muestra que tendrá en cuenta el efecto de aglomeración para un diseño complejo estará cercano a $n \approx n_0 \times DEFF$. Por ejemplo, un efecto de diseño DEFF = 2 implicaría que se deberían seleccionar casi el doble de unidades para lograr la misma precisión que la producida por una muestra aleatoria simple.

Se evidencia que valores grandes del efecto de diseño inducirán un mayor tamaño de muestra. Claramente el incremento no es lineal, más aún, el tamaño de muestre se ve más afectado en la medida en que el $DEFF$ sea más grande. En el caso de la EPC, el interés se centra en tener una muestra suficiente de hogares que permita establecer con precisión la proporción de personas que fueron omitidas en el censo. Para ello, es necesario establecer los siguientes elementos:

i) El número promedio de hogares (tamaño de los segmentos) en las áreas cartográficas o UPM está dado por $\bar{n}_{II}$.

ii) Siendo b el número de habitantes promedio de un hogar en el país, entonces el número promedio de personas en las UPM está dado por $\bar{n} = b \cdot \bar{n}_{II}$

iii) A partir de estudios anteriores, es necesario calcular el valor de la correlación intraclase ρ_y de la variable de interés (omisión censal) con las UPM. Luego de esto, se debe calcular el efecto de diseño $DEFF$ como función de ρ_y y de n_{II}; según Gutiérrez (2016), está cantidad está dada por:

$$DEFF \approx 1 + (\bar{n} - 1)\rho_y$$

iv) Partiendo de las expresiones de tamaño de muestra generales para muestreos complejos (CEPAL, 2023) y asumiendo que en la población de interés hay N individuos (que pueden ser tomados de las proyecciones demográficas actuales), entonces el tamaño de muestra necesario para alcanzar un margen de error relativo máximo es de

$$n \geq \frac{\theta \ \cdot (1-\theta) \ \cdot DEFF}{\frac{ME^2}{z^2_{1-\alpha/2}} + \frac{\theta \ \cdot (1-\theta) \ \cdot DEFF}{N}}$$

v) Dado que el número de habitantes promedio del hogar es b, entonces la muestra contendrá, en promedio, la siguiente cantidad de hogares:

$$n_{II} = \frac{n}{b}$$

vi) Así mismo, el número de UPM que deben ser seleccionadas en la muestra, se calcula a partir de la siguiente relación

$$n_I = \frac{n_{II}}{\bar{n}_{II}}$$

En particular, para el caso de una proporción, la calidad del estimador se puede medir en términos de la amplitud del intervalo de confianza de al menos $(1-\alpha) \times 100\%$; esto es, la distancia entre el estimador y el parámetro no debería superar un margen de error previamente establecido (ME). Así:

$$1 - \alpha \geq Pr\big(\left|\hat{\theta} - \theta\right| < ME \ \big)$$

Por ejemplo, suponga que se desea estimar la proporción de personas omitidas en el censo, que se presupone que está alrededor de $\theta = 0.04$ (4% de omisión censal esperada) con un margen de error máximo de $ME = 0.005$, y un nivel de confianza del 95% ($z_{1-\alpha/2} = 1.96$). Bajo estas condiciones, el intervalo de confianza esperado está alrededor de:

$$IC(1-\alpha) = \left[\hat{\theta} - ME, \hat{\theta} + ME \right] = [\ 0.04 - 0.005, 0.04 + 0.005\ ] = [\ 0.035, 0.045\ ]$$

Ahora, asumiendo que el número promedio de hogares por UPM es de $\bar{n}_{II} = 25$, que el número promedio de habitantes por hogar es $b = 4$, que la correlación intraclase de la omisión censal dentro de las UPM es de $\rho_y = 0.12$ y que el tamaño de la población está proyectado en $N = 1\ 000\ 000$. Entonces, el efecto de diseño será:

$$DEFF \approx 1 + (\bar{n} - 1)\rho_y = 1 + (100 - 1) \cdot 0.12 = 1 + 11.88 = 12.88$$

Por lo tanto, el tamaño mínimo que debe tener la muestra está dado por:

$$n \geq \frac{(0.04)(0.96)(12.88)}{\frac{0.005^2}{1.96^2} + \frac{(0.04)(0.96)(12.88)}{1\,000\,000}} \approx 70\,815$$

Por lo tanto, si se necesitan aproximadamente $n = 70\,815$ individuos en la muestra para alcanzar el margen de error deseado, entonces se deben seleccionar $n_{II} = 17\,704$ hogares en $n_I = 708$ unidades primarias de muestreo; así:

$$n_{II} = \frac{70\,815}{4} \approx 17\,704, \qquad n_I = \frac{17\,704}{25} \approx 708$$

C. Operativo de recolección

Habiendo seleccionado las UPM, el operativo de recolección debe garantizar un empadronamiento exhaustivo de todas las estructuras habitacionales contenidas en cada una de ellas. Esto implica que ningún tipo de vivienda, sea ocupada, desocupada o de uso temporal, puede quedar fuera del recuento. El objetivo es asegurar que todos los hogares y las personas residentes dentro del área muestral sean correctamente incluidos en la encuesta, de modo que la información recolectada refleje fielmente la cobertura y calidad del censo. Para lograrlo, el trabajo de campo debe planificarse cuidadosamente, asignando a cada equipo un área claramente delimitada, materiales cartográficos actualizados y procedimientos precisos para el recorrido y verificación de viviendas. Asimismo, el personal de campo debe recibir capacitación específica para identificar correctamente las unidades habitacionales, distinguir entre viviendas ocupadas y desocupadas, y aplicar el cuestionario completo a todos los hogares encontrados.

Un requisito fundamental en la planeación del operativo es tener en cuenta los supuestos del sistema de estimación dual, puesto que solo de esta manera se garantizará la validez de los resultados obtenidos. Por ello, esta sección considera algunos aspectos relevantes. En primer lugar, la independencia entre la Encuesta de Posenumeración Censal (EPC) y el censo es un requisito esencial para la aplicación del sistema de estimación dual, pues la validez de las estimaciones dependerá directamente de que esta suposición de independencia sea verificable, por lo que se deben realizar todos los esfuerzos posibles para mantener esta separación operativa.

El modelo basado en el SED requiere que las probabilidades de captura en los dos sistemas sean independientes para todos los individuos (Wolter 1986). Este supuesto implica dos tipos de independencia: la independencia causal y la independencia heterogénea. La primera establece que la inclusión en el censo es independiente de la inclusión en la EPC. La segunda plantea que la covarianza entre la probabilidad de ser incluido en el censo y la probabilidad de ser incluido en la EPC es igual a cero. Una condición suficiente para lograr la independencia heterogénea es que las probabilidades de inclusión en el censo o en la EPC sean iguales para todas las personas (Mulry y Spencer 1991).

La falla de alguno de estos supuestos produce un sesgo de correlación, generalmente a la baja, ya que las personas omitidas en el censo suelen tener mayor probabilidad de ser también omitidas en la EPC (Griffin 2000). Los supuestos pueden fallar por dependencia causal o por heterogeneidad en las probabilidades de captura. La dependencia causal ocurre cuando el hecho de que un individuo sea incluido o excluido de un sistema afecta su probabilidad de inclusión en el otro.

Para mitigar la dependencia causal, es necesario garantizar la independencia operativa entre la EPC y el censo. Esto implica que las operaciones de recolección de datos de ambos sistemas sean independientes, lo cual se puede lograr mediante acciones como:

- Asignar al personal de la EPC a áreas en las que no trabajaron durante el censo.
- Realizar las entrevistas de la EPC una vez finalizadas las operaciones censales en un área.
- Restringir el acceso del personal del censo a la información sobre la muestra de la EPC.
- Restringir el acceso del personal de la EPC a los resultados del censo durante la recolección de datos.

Asimismo, se recomienda que la EPC cuente con una unidad técnica independiente, dirigida por una persona responsable que dedique toda su atención a las actividades de la encuesta, sin asumir responsabilidades relacionadas con el censo. De igual forma, el personal asignado a esta unidad debe concentrarse exclusivamente en la EPC, sin funciones operativas vinculadas al censo. Por ende, la planificación operativa debe incluir desde las primeras etapas la capacitación del personal, garantizando la preparación oportuna y adecuada de quienes participarán en la operación.

Por otro lado, el cronograma de la investigación es un aspecto muy importante que debe detallar todas las actividades que se deben realizar con tiempos realistas para cada fase del proceso, como capacitación, prueba piloto y fechas de inicio y finalización del trabajo de campo, entre muchas otras. Es aconsejable contemplar alternativas al diseño inicial, de modo que, si surge algún inconveniente con el procedimiento planificado, se pueda recurrir a otro método sin afectar la operación. También es posible realizar una prueba piloto de todos los procedimientos relacionados con la EPC. Esta prueba debe funcionar como un ensayo general, evaluando desde la capacitación y recolección hasta el emparejamiento de registros. La muestra del estudio piloto no necesita ser probabilística y puede realizarse en áreas seleccionadas, con el objetivo de evaluar la adecuación del plan general y la organización de la EPC.

Como el proceso de emparejamiento de registros entre la EPC y el censo constituye uno de los elementos centrales para evaluar la cobertura y es una de las tareas más complejas de la operación, los resultados de la prueba piloto serán insumos fundamentales para planificar adecuadamente estas operaciones, permitiendo establecer las reglas de emparejamiento, los procedimientos de reconciliación y el flujo de trabajo de documentos entre la EPC y el censo, asegurando así la validez y utilidad de los resultados finales.

III. Procedimientos y enumeraciones

En este capítulo se presentan en detalle los principios fundamentales para definir y clasificar las inclusiones erradas y los errores de enumeración. El objetivo es establecer un marco claro y sistemático que facilite la identificación de estas inclusiones, de manera que puedan incorporarse correctamente en los estimadores del sistema dual. Asimismo, se abordarán los fundamentos metodológicos necesarios para la reconstrucción de los hogares, un paso clave que permite realizar un emparejamiento preciso entre los registros de la encuesta y los del censo, asegurando la validez y consistencia de los resultados.

A. La muestra P y la muestra E

El diseño de muestreo de una encuesta de posenumeración censal (EPC) constituye la base sobre la cual se estimará la cobertura del censo y se identificarán posibles errores de enumeración. Este diseño no se limita a la selección de una sola muestra; por el contrario, se estructura como un proceso de doble enfoque. En primer lugar, se selecciona una muestra de áreas geográficas que serán empadronadas de manera independiente, asegurando una captura directa de los hogares y las personas en la EPC. Pero, a partir de esta selección surge una segunda muestra, esta vez construida a partir de los registros del censo correspondientes a esas mismas áreas, que permite realizar la revisión y comparación necesarias para detectar omisiones, duplicaciones o inconsistencias, de la siguiente manera:

- La primera muestra se denomina **muestra de posenumeración** o **muestra P**, y consiste en un subconjunto de áreas cartográficas o UPM que serán empadronadas durante el operativo de la EPC, una vez concluido el censo.
- La segunda muestra se denomina **muestra de la enumeración** o **muestra E**, y está compuesta por todos y cada uno de los registros del censo correspondientes a las mismas áreas cartográficas o UPM seleccionadas en la muestra P.

La muestra P y la muestra E desempeñan roles complementarios en la estimación de la cobertura poblacional y en la corrección de errores en los conteos del censo. Inicialmente, ambas muestras provienen de las mismas áreas geográficas, lo que garantiza una base común para la comparación y el análisis de los datos. Sin embargo, como se verá en los capítulos de emparejamiento (capítulos IV y V), es posible que la muestra E necesite ser ampliada para incluir UPM adicionales, con el fin de localizar a personas que se hayan trasladado de domicilio.

El objetivo de la muestra P es doble. Por un lado, permite estimar directamente los valores de N_{11} y N_{+1} de la matriz de conteos del sistema de estimación dual (Cuadro 2). Por otro lado, al emparejarse con la información provista por la muestra E, se puede estimar indirectamente el valor de N_{1+}. Al combinar la muestra E con la muestra P, se pueden comparar los registros del censo y de la encuesta para obtener estimaciones directas del número de personas contadas correctamente y estimaciones indirectas del número de personas no contadas pero que deberían haber sido incluidas en el censo.

La muestra E, por su parte, permite corregir la presencia de eventos espurios, garantizando la validez del supuesto central del sistema de estimación dual. En particular, posibilita estimar el número de personas que fueron contadas en el censo pero que no deberían haber sido parte de la enumeración, como duplicados, personas nacidas después del censo, fallecidos antes del censo, migrantes o registros ficticios. A partir de esta muestra se calcula la proporción de inclusiones erróneas y se ajusta el conteo del censo eliminando estas imprecisiones.

B. Procedimientos para la reconstrucción de los hogares

En esta sección se presentan tres procedimientos para la reconstrucción de los hogares en el marco de la EPC. Estos procedimientos están basados en United Nations (capítulo 6, 2010) y tienen como objetivo asegurar que cada hogar y persona sea correctamente identificado y ubicado, facilitando la evaluación de la cobertura y calidad del censo. La aplicación de estos métodos se realiza a partir de la muestra P haciendo posible verificar la unicidad, completitud y corrección geográfica de los registros censales, y con ello detectar posibles omisiones o duplicaciones en el conteo de hogares y personas.

Procedimiento A

Este método tiene como objetivo reconstruir los hogares tal como existían durante el día del censo. Mediante entrevistas retrospectivas con un miembro del hogar, generalmente el jefe de familia, se recopila información sobre todas las personas que residían en la vivienda en la fecha censal, incluyendo quienes ya se han trasladado a otra ubicación. La información recolectada se contrasta con los registros del censo para identificar personas que pudieron haber sido omitidas o registradas incorrectamente. El emparejamiento se realiza tomando en cuenta datos demográficos clave, como nombre, edad y sexo, y considerando la ubicación geográfica de cada hogar.

Este enfoque es especialmente efectivo en áreas donde la mayoría de los habitantes permanece en su vivienda habitual, ya que facilita el emparejamiento al reducir la necesidad de buscar registros en distintas localizaciones. No obstante, en contextos con alta movilidad, depender de informantes para rastrear a quienes se han mudado puede ocasionar información incompleta o inexacta, lo que podría llevar a subestimar las omisiones en zonas urbanas o entre poblaciones migrantes.

Procedimiento B

Este procedimiento se centra en todas las personas que viven en el hogar al momento de la EPC. Durante la entrevista, se solicita al informante que indique la dirección en la que residía cada persona en la fecha del censo. Este enfoque permite identificar tanto a quienes se mudaron al hogar después del censo (*in-movers*) como a quienes ya habían dejado la vivienda (*out-movers*). Los datos recopilados se comparan con los registros del censo para verificar si estas personas fueron correctamente incluidas en su ubicación original.

Para ejecutar este procedimiento, es necesario rastrear a las personas que se mudaron en las áreas donde fueron censadas originalmente. Dado que estas áreas pueden no estar incluidas en la muestra E inicialmente, el emparejamiento debe extenderse a otras zonas según corresponda. Este

método resulta especialmente útil en contextos de alta movilidad, porque permite capturar mejor a los individuos que se trasladaron y ofrecer una visión más completa de posibles errores de cobertura. Sin embargo, su efectividad puede verse limitada por la dificultad de validar direcciones, particularmente en zonas rurales o donde la información disponible es imprecisa.

Procedimiento C

Este método integra elementos de los procedimientos A y B, con el fin de identificar tanto a las personas que actualmente residen en el hogar como a aquellos que vivían allí en la fecha del censo, incluyendo a quienes se han trasladado (*in-movers* y *out-movers*). Durante la entrevista, se recaba información sobre los habitantes actuales y se solicitan detalles de quienes estaban presentes en el hogar en la fecha de referencia del censo, lo que permite reconstruir la composición del hogar en ambos momentos. No obstante, únicamente se emparejan con los registros censales aquellos que estaban presentes en la fecha del censo; es decir, los residentes permanentes (*non-movers*), así como los residentes de salida (*out-movers*). Este enfoque ofrece una visión más completa de los errores de cobertura, al capturar tanto a los residentes actuales como a los previos, siendo especialmente útil en zonas con elevada movilidad.

Para ejemplificar cada procedimiento, y poder distinguir sus diferencias, considere el Cuadro 4 conformado por cinco personas.

Cuadro 4
Ejemplo de los miembros de un hogar según su residencia en la fecha del censo y de la EPC.

Persona	¿Vivía en el hogar en la fecha del censo?	¿Vivía en el hogar en la fecha de la EPC?	Clasificación	Procedimiento A	Procedimiento B	Procedimiento C
Ana	Sí	Sí	*Non-mover*	1	1	1
Juan	Sí	No	*Out-mover*	1	0	1
Marta	No	Sí	*In-mover*	0	1	1
Luis	Sí	Sí	*Non-mover*	1	1	1
Carlos	No	Sí	*In-mover*	0	1	1
Lassie	No	No	*Out-of-scope*	0	0	0

Fuente: Elaboración propia.

Con respecto al Cuadro 4, en las tres últimas columnas se muestra cómo se clasifican las personas según su presencia en el hogar en la fecha del censo y en la fecha de la EPC, y cómo esto determina su consideración como miembro del hogar bajo los tres procedimientos. Nótese que un uno indica que la persona se considera miembro del hogar y un cero que no lo es. Por lo anterior, se tiene que:

- Ana y Luis son *non-movers*, es decir, estuvieron presentes tanto en el censo como en la EPC, por lo que se consideran miembros en los tres procedimientos.
- Juan es un *out-mover*, estuvo presente en el censo, pero no en la EPC; por lo tanto, se considera miembro en el procedimiento A, que solo evalúa la fecha del censo, y también en el procedimiento C, que incluye todos los miembros de la fecha del censo o de la EPC. No así en el procedimiento B, ya que este solo considera miembros que pueden emparejarse en la EPC.
- Marta y Carlos son *in-movers*, ausentes en el censo, pero presentes en la EPC; por ello, no se consideran miembros en A, pero sí en B y C, que incluyen la reconstrucción del hogar con base en la EPC.

- Finalmente, Lassie es *out-of-scope*, no pertenece al universo censal y no estuvo presente en ninguna de las fechas, por lo que se clasifica como no miembro en todos los procedimientos (o en A, B y C).

¿Qué procedimiento utilizar para la reconstrucción de hogares? La respuesta está estrechamente vinculada al tipo de censo que implementó el país. En los censos de hecho, donde las personas se enumeran en el lugar donde pasan la noche del día del censo, la estructura de los hogares es más sensible a la movilidad temporal. Esto puede generar discrepancias entre la composición del hogar observada por el censo y la registrada posteriormente en la EPC, especialmente en el caso de miembros ausentes temporalmente o de personas que pernoctaron en una vivienda distinta a su residencia habitual. En tales contextos, los procedimientos de reconstrucción deben centrarse en identificar con claridad quiénes estaban presentes físicamente y bajo qué condiciones, lo que tiende a favorecer el uso de procedimientos más estrictos como el Procedimiento A.

Por el contrario, en los censos de derecho, donde las personas son enumeradas en su residencia habitual, la reconstrucción de los hogares en la EPC es conceptualmente más coherente, pues ambos operativos se basan en el mismo criterio de pertenencia al hogar. Esto reduce la probabilidad de inconsistencias en la composición del hogar y facilita la aplicación de procedimientos más flexibles como los Procedimientos B o C, que permiten considerar ausentes habituales y otras situaciones de residencia más complejas.

C. Clasificación de las enumeraciones

Los registros censales en la muestra E deben ser revisados cuidadosamente con el fin de clasificar cada caso en función de su correspondencia con la información de la EPC. Este proceso de revisión implica verificar, para cada persona o vivienda, si su registro en la encuesta coincide con uno existente en el censo y si la información recolectada cumple con los criterios de elegibilidad y residencia establecidos. La clasificación de los registros se organiza en cuatro categorías principales:

- Enumeraciones correctas, que corresponden a personas que debían ser censadas y fueron efectivamente incluidas en el censo.
- Enumeraciones incorrectas, que se refieren a personas incluidas en el censo, aunque no debían haber sido contadas.
- Omisiones, que representan a personas que debían haber sido censadas, pero no fueron incluidas en el censo.

Este proceso de clasificación es fundamental para medir los diferentes tipos de errores de cobertura y para estimar la exactitud global del censo a partir de la comparación entre la encuesta y los datos censales.

Enumeraciones correctas

Una enumeración correcta es aquella en la que una persona que debía ser incluida en el censo, conforme a las reglas de residencia establecidas, fue efectivamente registrada en los datos censales. En otras palabras, la enumeración es correcta cuando la persona fue efectivamente incluida en las bases de datos censales y además cumple con los criterios de elegibilidad del censo (por ejemplo, residir habitualmente en el país o en la vivienda al momento de referencia, entre otros).

Para poder estimar los parámetros necesarios en el sistema de estimación dual de forma apropiada es fundamental definir las condiciones bajo las cuales un individuo se considera correctamente enumerado en el censo. Este proceso implica identificar y eliminar errores, tales como duplicaciones, casos

inexistentes o personas fuera del alcance del censo. Según Hogan (2003), una enumeración correcta debe cumplir con las siguientes dimensiones:

- Correspondencia: una persona debe ser incluida en el censo únicamente si forma parte de la población objetivo. Por ello, es necesario excluir a quienes fallecieron antes del día del censo o nacieron después de esa fecha, ya que no pertenecen al grupo poblacional que se busca medir. Según Bureau (2022), también se deben excluir registros correspondientes a individuos fuera del universo censal (*out of scope*), como turistas, animales o personas ficticias. De manera similar, las personas que deberían haber sido contadas en alojamientos colectivos (como residencias estudiantiles o cárceles) no se consideran parte del universo objetivo de la EPC, por lo que sus registros se clasifican igualmente como fuera de alcance.
- Unicidad: el objetivo es contar a cada persona una sola vez. Si un individuo aparece en más de un registro censal, se considera una duplicación y es esencial eliminar estos duplicados, ya que distorsionarían el conteo poblacional. La unicidad asegura que el número de registros coincida con el número real de personas. Es posible realizar una búsqueda exhaustiva en las bases de datos censales para identificar posibles duplicados.
- Completitud: un registro censal debe contener información suficiente para identificar de manera única a una persona. Si faltan datos clave, como nombre, edad o dirección, no será posible verificar si la persona fue incluida correctamente en el censo o si también aparece en la encuesta. Solo aquellos individuos que cumplan con un nivel mínimo de completitud y tengan información completa en las variables esenciales podrán considerarse correctamente enumerados.
- Corrección geográfica: las personas deben estar enumeradas en la ubicación correcta según las reglas de residencia del censo. Para evaluar esto a partir de la EPC, se debe utilizar una definición específica para determinar la corrección geográfica. Por ejemplo, es posible considerar que una persona está correctamente enumerada si fue contada en una vivienda dentro del segmento censal, o si fue incluida en una vivienda que está ubicada en un segmento adyacente. Esta definición amplía el área de búsqueda para incluir no solo la ubicación exacta, sino también las áreas circundantes, lo que permite corregir errores menores en la asignación geográfica. Sin embargo, un área de búsqueda más grande aumenta la complejidad del emparejamiento y el riesgo de coincidencias incorrectas entre personas diferentes.

En términos operativos, una enumeración se considera correcta cuando la persona que debía ser censada fue incluida en la misma vivienda donde residía en la fecha del censo. También se acepta como correcta si fue registrada en otra vivienda dentro del mismo sector censal, siempre que se trate de la misma persona que debía ser contada en ese lugar. Más aún, la enumeración sigue siendo correcta a nivel nacional si la persona fue incluida en el censo en cualquier otro lugar del país, incluso si no corresponde a su dirección o vivienda real. Lo fundamental es que la persona haya sido contada, sin importar el lugar exacto donde fue registrada.

Por ejemplo, si alguien que debía ser censado en una vivienda específica aparece incluido en otra unidad de vivienda en una zona diferente, la enumeración sigue siendo válida, aunque esté ubicada en el lugar equivocado. Sin embargo, a nivel subnacional esta flexibilidad no se aplica. Para fines de análisis regional, departamental o municipal, una enumeración solo se considera correcta si la persona fue registrada en la vivienda o unidad correspondiente dentro de la misma área geográfica. Esta distinción es fundamental para garantizar la precisión en la estimación de la población por zonas específicas y para evaluar correctamente la cobertura del censo en ámbitos subnacionales

Enumeraciones erróneas

Desde la EPC también se debe estimar el número de enumeraciones erróneas, que pueden producirse por diversas razones. En términos generales, estas se clasifican en dos grandes categorías: las que ocurren por duplicación y las que se producen por otras razones. Una enumeración errónea por duplicación corresponde a un registro que replica a una persona que ya fue contada correctamente en una unidad de vivienda o en el universo de viviendas colectivas del censo. Asimismo, las enumeraciones erróneas por otras razones Las enumeraciones erróneas por otras razones incluyen registros ficticios, personas registradas en el lugar incorrecto, nacimientos o fallecimientos fuera de la fecha del censo, así como personas ausentes temporalmente o visitantes extranjeros presentes durante el censo. El cuadro 5 presenta un resumen de algunas enumeraciones erróneas que se pueden encontrar en una EPC.

Debido a que la muestra E es probabilística, es posible utilizarla para estimar de manera insesgada el número total de enumeraciones erróneas en todo el censo. Cada registro de la muestra representa un subconjunto de la población, y mediante la aplicación de los factores de expansión correspondientes, se puede extrapolar la cantidad de errores detectados en la muestra a todo el universo censal. Esto permite obtener estimaciones confiables de la sobreenumeración y de otros tipos de errores, sin necesidad de revisar exhaustivamente cada uno de los registros del censo. Este proceso no solo permite refinar las cifras finales de población, sino que también proporciona información valiosa para mejorar los procedimientos de recolección y registro en censos futuros, asegurando que se reduzcan los errores y que los datos reflejen con mayor precisión la realidad poblacional.

Cuadro 5
Tipos de enumeraciones erróneas

Tipo	Descripción
Erróneo por duplicación	- La enumeración es un duplicado de una persona que fue contada correctamente en una unidad de vivienda o en el universo de viviendas colectivas en el censo.
Erróneo por otras razones	- El registro es ficticio y no corresponde a una persona real. - La persona fue enumerada en el universo de unidades de vivienda, pero debería haber sido enumerada en el universo de viviendas colectivas o estaba en situación de calle el día del censo. - La persona nació después del día del censo. - La persona falleció antes del día del censo. - La persona estaba trabajando, estudiando o viviendo fuera del país el día del censo. - La persona es un visitante con residencia habitual en el extranjero que estaba temporalmente en el país el día del censo.

Fuente: Adaptación de Zamora, J. (2022).

Omisiones

Las omisiones corresponden a personas que debieron ser incluidas en el censo, pero que no aparecen en ningún registro censal emparejado con la encuesta de posenumeración. Estas omisiones pueden producirse por diversas razones, como errores de cobertura, dificultades en la enumeración de poblaciones móviles, rechazo a participar en el censo, o problemas operacionales durante la recolección de datos. La identificación de omisiones es uno de los objetivos más importantes de la EPC, ya que permite medir con precisión qué tan completa fue la cobertura del censo y estimar la magnitud de la población que no fue contabilizada.

Para determinar que un individuo fue efectivamente omitido, es necesario que cumpla los siguientes criterios: debe residir en el país al momento del censo, debe estar correctamente vinculado a una unidad censal en la encuesta, y no debe haber sido emparejado con ningún registro censal, incluso después del emparejamiento ampliado y de la revisión clerical. A continuación, se presentan algunos casos comunes de omisión en los censos:

- Áreas rurales remotas: personas que viven en lugares de difícil acceso y no fueron contactadas durante el censo. Esto incluye habitantes de comunidades alejadas, islas

pequeñas o zonas selváticas donde no llegaron los censistas, así como familias dispersas en grandes extensiones de terreno que no fueron localizadas debido a la falta de caminos transitables o transporte.

- Poblaciones móviles: personas que estaban temporalmente fuera de sus hogares habituales durante la fecha del censo. Entre ellas se encuentran trabajadores agrícolas migrantes que se encontraban en otra región por la temporada de cosecha, estudiantes universitarios residiendo temporalmente en otra ciudad o país, y tripulantes de barcos, camiones o aviones que estaban lejos de sus domicilios al momento de la enumeración.
- Rechazo a participar: personas que se negaron a responder el censo o no quisieron ser registradas. Esto puede incluir residentes urbanos que temen brindar información por desconfianza hacia las autoridades, personas que no comprendieron el propósito del censo y decidieron no responder, u hogares donde ningún miembro quiso atender al censista a pesar de múltiples intentos.
- Problemas operacionales: casos en que los errores logísticos o administrativos impidieron la enumeración. Por ejemplo, hogares no visitados debido a fallas en la planificación de rutas censales o cambios en la ubicación de la vivienda, así como registros perdidos o mal clasificados por errores en la captura de datos durante la recolección.
- Situaciones especiales: personas que no encajan en las categorías tradicionales de vivienda y fueron omitidas. Esto incluye personas en situación de calle, residentes de viviendas colectivas no previstas como refugios temporales, albergues o campamentos, y personas alojadas temporalmente en hospitales, prisiones o instituciones donde no se realizó la enumeración completa.

De igual manera, las omisiones se cuantifican utilizando los factores de expansión de la EPC, para extrapolar los errores detectados en la muestra al universo censal completo. Estos ajustes se incorporan luego en el cálculo del error neto de cobertura proporcionando una medida precisa de la diferencia entre la población realmente censada y la población que debería haber sido incluida. El impacto de las omisiones se realiza por grupos demográficos, como edad, sexo, nivel educativo o ubicación geográfica, lo que permite identificar patrones de cobertura diferencial y evaluar si ciertos segmentos de la población fueron subrepresentados en el censo. Esta información es clave para comprender la calidad y de los datos censales, y diseñar estrategias de mejora en futuros censos (Wolter, 1986; Bureau, 2022).

IV. Emparejamiento estadístico

El emparejamiento estadístico constituye una etapa fundamental en el análisis de la cobertura del censo a través de la Encuesta de Posenumeración Censal (EPC), ya que permite vincular la información de la encuesta con la recopilada durante el censo, pues su propósito principal es identificar, para cada persona registrada en la muestra E, si existe una correspondencia en la muestra P y viceversa. A partir de este proceso se determinan las coincidencias y discrepancias que servirán de base para estimar las omisiones y los errores de enumeración. Una vez logrado este vínculo, el emparejamiento de personas permite determinar quiénes de los residentes censales fueron efectivamente encontrados en la encuesta y quiénes no, así como identificar a las personas que fueron incluidas erróneamente o que se incorporaron al hogar después del censo.

En la práctica, el emparejamiento estadístico combina procedimientos automáticos y revisiones manuales. En la etapa automática, se utilizan variables comunes entre ambas fuentes (como nombre, sexo, edad y parentesco con el jefe del hogar) y se aplican algoritmos de comparación que asignan un puntaje de similitud a cada posible coincidencia. Posteriormente, los casos ambiguos o de baja coincidencia se someten a una revisión manual, en la cual se analizan los registros para confirmar o descartar la correspondencia. El resultado final del proceso es un conjunto de pares emparejados entre las muestras P y E. Si al final del proceso existen registros que no se han logrado emparejar, entonces la muestra E puede ampliarse a otras áreas para identificar si la persona encontrada en la muestra P sí fue censada, pero en un segmento diferente.

La calidad del emparejamiento entre el censo y la EPC mejora notablemente cuando se solicita el número de identificación nacional durante ambas entrevistas. Contar con este dato permite verificar la identidad de cada persona con mayor precisión, reduce los errores de clasificación entre individuos con nombres similares y disminuye las omisiones y duplicaciones asociadas a fallas de reconocimiento o recordación. Su uso sistemático fortalece la vinculación de registros y contribuye a estimaciones más confiables del error de cobertura. No obstante, en muchos censos este número no se solicita de manera uniforme, en parte por consideraciones de confidencialidad y seguridad que podrían generar desconfianza entre los informantes ante el temor de un uso indebido de su información personal. Aun así, disponer del número de identificación tanto en el censo como en la EPC constituiría un insumo

valioso para identificar coincidencias exactas y realizar un emparejamiento determinístico, permitiendo vincular registros con alto grado de certeza antes de recurrir a métodos probabilísticos cuando la información es incompleta o presenta inconsistencias.

Por consiguiente, el emparejamiento determinístico no siempre es viable, ya que la información registrada en la muestra P y la muestra E puede presentar diferencias en la escritura de nombres, errores de tipográficos o de digitación, diferencias de formato, entre otros. Estas inconsistencias impiden obtener coincidencias perfectas, por lo que resulta necesario recurrir a técnicas de emparejamiento probabilístico que permitan estimar el grado de similitud entre registros y determinar las correspondencias más probables. Por ejemplo, los registros "*Nohora Rodriguez, nacida el 8/10/1960*" y "*Nora Rodrigues, nacida el 19601008*" podrían corresponder a la misma persona, aunque un algoritmo determinístico no los vincularía. En cambio, el emparejamiento probabilístico sí permitiría identificar estas coincidencias aproximadas mediante criterios estadísticos.

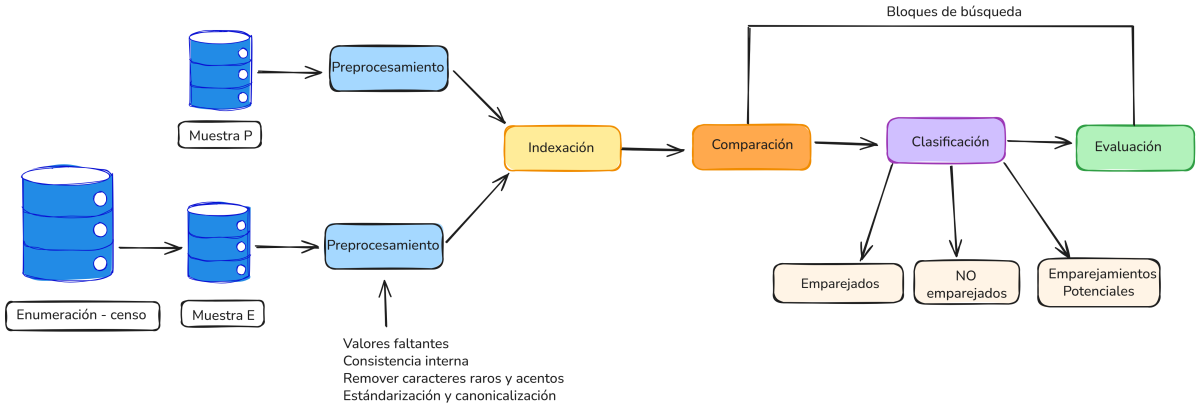
El emparejamiento probabilístico de registros (*record linkage)* es una técnica esencial en el ámbito de los censos y las encuestas de cobertura, cuyo propósito es identificar registros que se refieren a la misma persona en diferentes fuentes de datos, incluso en ausencia de un identificador único o cuando existen errores, inconsistencias o diferencias en los formatos. El término fue introducido formalmente en 1946 por Dunn, en el contexto de la construcción de un gran registro que integrara eventos vitales. En las décadas de 1950 y 1960, se desarrolló el enfoque probabilístico de vinculación de registros, impulsado por los trabajos de Newcombe y Kennedy (1962), quienes propusieron asignar probabilidades de correspondencia a partir de las coincidencias en variables clave. Este enfoque fue formalizado por Fellegi y Sunter (1969), quienes demostraron que es posible establecer una regla óptima de decisión para determinar si dos registros pertenecen a la misma entidad. En la actualidad, estas metodologías constituyen la base del emparejamiento entre las muestras E y P.

El emparejamiento probabilístico se aplica específicamente a los registros que no pueden vincularse mediante coincidencias determinísticas. Por lo general esta búsqueda probabilística se realiza dentro de áreas de enumeración o bloques de búsqueda definidos para limitar las comparaciones a segmentos censales y zonas adyacentes. Un aspecto clave es la definición del umbral de coincidencia: por ejemplo, es posible definir que los pares con una probabilidad superior al 99% se consideran emparejados, los que se ubican entre el 90% y el 99% se clasifican como emparejamientos potenciales, y los inferiores al 90% como no emparejados. Los casos clasificados como emparejamientos potenciales deben ser revisados y validados manualmente por personal especializado. En el caso de los registros no emparejados, es posible ampliar el área de búsqueda a segmentos o zonas adyacentes; sin embargo, esta ampliación incrementa el riesgo de emparejamientos erróneos. Por ello, se recomienda realizar también una revisión clerical de los registros emparejados tras dicha expansión, incluso cuando su probabilidad de coincidencia sea alta. Si definitivamente no hay coincidencia tras ampliar el área de búsqueda, el registro se debería clasificar como una omisión; es decir, personas que no fueron enumeradas en el censo.

En primer lugar, se realiza el preprocesamiento de datos, cuyo objetivo es garantizar que la información proveniente del censo y de la EPC, a través de las muestras E y P, esté en un formato uniforme y comparable. A continuación, en la etapa de indexación, se busca reducir la complejidad del proceso mediante estructuras de datos que permiten generar de forma eficiente pares de registros candidatos que probablemente correspondan a la misma persona. En el siguiente paso, se lleva a cabo la comparación de registros, utilizando diversas variables para evaluar la similitud entre los pares identificados. Posteriormente, durante la clasificación, cada par se asigna a una de tres categorías: emparejado, no emparejado o emparejamiento potencial; estos últimos requieren una revisión clerical para confirmar su correspondencia final. Finalmente, se realiza una evaluación de la calidad y completitud de los resultados del emparejamiento, con el fin de garantizar la validez del proceso.

El diagrama 1 muestra un esquema de las principales fases del proceso de emparejamiento, el cual comprende varias etapas interrelacionadas.

Diagrama 1
Flujo general del proceso de emparejamiento



Fuente: Elaboración propia.

A. Preprocesamiento

La fase de preprocesamiento constituye un primer paso fundamental para garantizar la calidad y consistencia de las bases de datos antes del emparejamiento. Este paso se aplica tanto a la base de datos derivada de la muestra P como al subconjunto de datos censal inducido por la muestra P, asegurando que la información con la que se realizará el emparejamiento esté estandarizada, validada y lista para vincularse.

El preprocesamiento incluye la geocodificación, para verificar que las direcciones correspondan a los segmentos apropiados de la muestra; la consistencia lógica, que valida relaciones de parentesco, edades y estructura de los hogares; la estandarización de formatos en variables clave como fechas, sexo, edad y categorías, unificando formatos y corrigiendo errores de codificación; y la normalización de nombres, que aplica reglas para asegurar la validez de nombres y apellidos, eliminando caracteres especiales, espacios innecesarios y uniformando el formato de escritura. Este conjunto de procedimientos garantiza que los registros estén en condiciones óptimas para el emparejamiento, reduciendo errores y mejorando la precisión del proceso.

Geolocalización

A partir de la muestra P, el primer paso consiste en codificar geográficamente las direcciones proporcionadas por los encuestados y verificar que estas coincidan con las UPM (segmentos cartográficos) efectivamente seleccionadas en la EPC. En caso de que algunas direcciones no tengan la precisión deseada en el segmento cartográfico, será necesaria una revisión clerical para verificar las direcciones proporcionadas por los encuestados.

El ejemplo del cuadro 6 muestra un conjunto de datos con cinco direcciones en distintos municipios de un país. Cada dirección se combina con el nombre del municipio y del país para formar una dirección completa. Luego, se obtiene la georreferenciación, es decir, las coordenadas de latitud y longitud correspondientes a cada ubicación.

En el caso de que algunos de los puntos de longitud y latitud no estén ubicados dentro de los segmentos de la muestra P, los revisores clericales deben verificar las direcciones y establecer si se describieron algunos puntos de referencia que no se usaron durante el procesamiento automatizado y que hubieran afectado la precisión del proceso automático. Los resultados de la geocodificación se utilizan durante el proceso de emparejamiento para identificar áreas de búsqueda alrededor de la dirección proporcionada por el encuestado.

Cuadro 6
Direcciones y localización geográfica asociada

Dirección	Municipio	Latitud	Longitud
Av. Jaime Mendoza 123	Sucre	-19,03	-65,26
Calle Bolívar 456	Monteagudo	-19,77	-63,41
Plaza 25 de mayo 789	Camargo	-20,05	-64,55
Av. del Maestro 321	Villa Serrano	-20,43	-64,55
Calle Potosí 654	Zudáñez	-19,98	-64,85

Fuente: Elaboración propia.

Durante el proceso de geocodificación manual, los revisores asignan una coordenada que permita una mayor precisión. Si no es posible lograr una precisión que apunte a una UPM específica de la muestra P, entonces la misma podrá asociarse a más de una UPM para crear áreas de búsqueda que abarquen dicha dirección. Asimismo, es recomendable que se asigne un código que refleje el nivel de confianza que el revisor manual considera que existe en que la dirección se encuentra dentro del área de búsqueda.

Como se verá más adelante, es recomendable que el emparejamiento automático de personas incluya los geocódigos asignados a las direcciones proporcionadas por los encuestados, así como los nombres, apellidos, la edad, el sexo, y el día y mes de nacimiento. Otra información que puede ser usada en el proceso son los números de teléfono de los encuestados del hogar y datos geográficos como el departamento, municipio o código del segmento.

Consistencia de formatos

Las tablas de datos utilizadas en el proceso de emparejamiento suelen variar en formato, estructura y contenido. Dado que este proceso se basa frecuentemente en información personal, como nombres, sexo, direcciones y fechas de nacimiento, es fundamental garantizar que los datos provenientes de distintas bases sean estandarizados adecuadamente. El objetivo de esta etapa es asegurar que los atributos utilizados para el emparejamiento tengan estructuras homogéneas y formatos consistentes. Esta etapa generalmente involucra, al menos, cinco pasos esenciales:

i) Eliminar caracteres y palabras irrelevantes: esta limpieza inicial consiste en remover caracteres como comas, dos puntos, puntos y comas, puntos, numerales y comillas. En algunas aplicaciones, también se eliminan palabras que no aportan información relevante para el emparejamiento, conocidas como *stop words* o palabras vacías.

ii) Expandir abreviaturas y corregir errores ortográficos: este paso es crucial para mejorar la calidad de los datos. Usualmente se emplean tablas de búsqueda que incluyen variaciones de nombres, apodos y errores ortográficos comunes, junto con sus formas correctas o expandidas. La estandarización resultante reduce significativamente las variaciones en los atributos que contienen nombres.

iii) Codificación fonética: los errores de ortografía o la escritura diferente de nombres pueden dificultar el emparejamiento automático. Por ejemplo, "Catalina Benavides" podría aparecer como "Katalina Venavidez", y un algoritmo simple no identificaría esta coincidencia. La codificación fonética ayuda a superar este tipo de variaciones.

iv) Segmentación: consiste en dividir atributos que contienen múltiples piezas de información en nuevos atributos individuales, cada uno con una información bien definida (Herzog, Scheuren y Winkler, 2007). Por ejemplo, un nombre completo como "*María José Pérez López*" se puede separar en dos campos: Nombres: "*María José*", Apellidos "*Pérez López*".

v) Verificación: Cuando existen fuentes externas confiables, se puede validar la información recolectada. Por ejemplo, una base de datos con direcciones válidas en un país o región permite corroborar rangos de números de calles y combinaciones de nombres de calles, asegurando la precisión de la información del censo y de la EPC.

Limpieza y normalización

Es posible implementar una rutina de limpieza que prepare y estandarice un texto, facilitando su análisis y posterior procesamiento automático de datos. Por ejemplo:

- Hay que asegurar que el texto esté en una codificación estándar, lo que evita la aparición de símbolos extraños; por ejemplo, un nombre como "*JosÃ©*" se convierte correctamente en "*José*". A continuación, todas las letras se transforman a minúsculas para uniformar el estilo, de modo que nombres como "*Andrés*", "*ANDRÉS*" y "*andrés*" queden representados de manera consistente como "*andrés*".

- El siguiente paso consiste en eliminar acentos y signos diacríticos, de manera que "*José*" se transforme en "*jose*" y "*María*" en "*maria*". Luego, se reemplazan todos los signos de puntuación por espacios, de modo que un nombre escrito como "*Juan-Camilo!*" se convierta en "*juan camilo*". Posteriormente, se reducen los espacios múltiples a un solo espacio y se eliminan los espacios sobrantes al inicio y al final del texto, garantizando que cadenas como " *Luis Fernando* " queden uniformizadas como "*luis fernando*".

- Además, el investigador puede definir un conjunto de palabras vacías o irrelevantes para el análisis, que se eliminarán de las cadenas de texto. Estas palabras suelen ser artículos, preposiciones o conjunciones, como "*de*", "*del*", "*la*", "*los*", "*las*", "*el*" o "*y*", que no aportan información significativa. Una vez definido este conjunto, se recorren las cadenas de texto para eliminar todas sus ocurrencias, cuidando de no dejar espacios múltiples o sobrantes.

Como resultado de estos pasos, se obtendrá un texto limpio, homogéneo y estandarizado, listo para un análisis más preciso y eficiente. Este preprocesamiento asegura que los datos textuales estén en un formato uniforme, facilitando su comparación, emparejamiento y análisis automatizado.

Codificación fonética

Existen diversas funciones diseñadas para codificar fonéticamente los valores de ciertos atributos antes de utilizarlos en procesos de emparejamiento. Su propósito es mitigar los errores derivados de variaciones en la escritura o errores ortográficos, especialmente en variables como nombres, apellidos u otras susceptibles a inconsistencias tipográficas. Estas funciones buscan agrupar cadenas de texto que suenan de forma similar al ser pronunciadas, aunque estén escritas de manera distinta. Como se verá más adelante, la codificación fonética también puede combinarse con medidas de similitud (como la distancia de Levenshtein, Smith-Waterman o el coeficiente de Jaccard), para comparar cadenas de texto que suenan de forma similar (Navarro 2001; Nauman y Herschel 2022).

El objetivo principal de este proceso es transformar un texto en un código fonético que refleje su pronunciación. No obstante, muchas de las técnicas clásicas de codificación fonética fueron desarrolladas originalmente para el idioma inglés, lo que limita su aplicación directa en los contextos de América Latina y el Caribe, donde predominan otros idiomas como el español, el portugués, el francés y diversas lenguas indígenas. A pesar de estas limitaciones, algunos métodos pueden resultar útiles en este contexto. En esta sección se describirán algunos de los algoritmos más conocidos.

Algoritmo Soundex

Este algoritmo es uno de los métodos más antiguos y ampliamente conocidos para la codificación fonética de cadenas de texto. Fue desarrollado originalmente por Odell y Russell (1918) y ha sido utilizado tradicionalmente en tareas como la consolidación de listas de nombres y la indexación de registros. En el ámbito del emparejamiento de registros entre censos y encuestas de cobertura en América Latina y el Caribe, Soundex puede servir como una herramienta complementaria para enfrentar errores de escritura, diferencias dialectales, y variaciones ortográficas en nombres y apellidos.

Es importante recalcar que este algoritmo fue diseñado originalmente para nombres en inglés estadounidense, por lo que puede presentar limitaciones en su aplicación directa a nombres hispanos, portugueses o de otras lenguas de la región. Sin embargo, su simplicidad y bajo costo computacional lo convierten en un buen punto de partida para ilustrar los principios básicos de codificación fonética. La racionalidad del algoritmo descansa en que las palabras que suenan de forma similar, aun cuando se escriban distinto, compartan el mismo código, facilitando su comparación y emparejamiento. Por ejemplo, las letras *b, f, p, v* se codifican como 1, ya que comparten una pronunciación similar; mientras que *c, g, j, k, q, s, x, z* se codifican como 2. Las vocales (*a, e, i, o, u*) y las letras *h, w, y* no se consideran en la codificación y se eliminan del resultado. EL cuadro 7 muestra la asignación de los códigos.

Cuadro 7
Codificación de caracteres en el algoritmo Soundex

Letras	Código
b, f, p, v	1
c, g, j, k, q, s, x, z	2
d, t	3
l	4
m, n	5
r	6
a, e, i, o, u, h, w, y	0 - Se elimina

Fuente: Elaboración propia.

El proceso estándar del algoritmo Soundex consiste en conservar la primera letra de la palabra, sustituir las demás letras por los códigos numéricos según el anterior cuadro, eliminar los ceros y los números repetidos consecutivos, y ajustar el resultado a una longitud de cuatro caracteres, rellenando con ceros si es necesario. Por ejemplo, la palabra "*Gómez*" y "*Gomes*" se codifican con el mismo código: G520. En el caso de "*Yenny*", el algoritmo Soundex conserva la primera letra *Y*, codifica las siguientes según la tabla, así: *e* → 0, *n* → 5, *n* → 5, *y* → 0, obteniendo la secuencia Y0550. Luego elimina los ceros y los números repetidos consecutivos, quedando Y5, y finalmente agrega ceros hasta completar cuatro caracteres, resultando en Y500.

Algoritmo Metaphone

El algoritmo Metaphone es una técnica de codificación fonética desarrollada por Philips (1990), diseñada para mejorar la coincidencia de palabras con escritura diferente pero pronunciación similar. El algoritmo Metaphone no asigna códigos numéricos sino representaciones fonéticas alfabéticas, lo que permite una mayor precisión, especialmente para consonantes. No obstante, como fue diseñado originalmente para el inglés, su aplicación en nombres de origen hispano o indígena puede ser limitada. Para superar estas limitaciones, se desarrollaron algoritmos posteriores como Double Metaphone, que permite hasta dos codificaciones por palabra para capturar variaciones fonéticas adicionales (P. Christen 2012).

El algoritmo genérico del Metaphone consiste en convertir primero todas las letras de la palabra a mayúsculas y eliminar los caracteres no alfabéticos. Se conserva la primera letra si es una consonante sonora inicial y luego se recorre cada letra o grupo de letras aplicando reglas fonéticas que reemplazan combinaciones por consonantes según su sonido, eliminando vocales internas, letras silenciosas y reduciendo consonantes dobles a una sola. Después se eliminan letras consecutivas redundantes que representen el mismo sonido, y finalmente se concatenan las letras resultantes para formar la cadena fonética, que puede limitarse a una longitud máxima según la implementación. Esta cadena representa la pronunciación de la palabra para compararla con otras.

Aplicando este algoritmo a las palabras "*Gómez*" y "*Gomes*" se codifican de manera similar porque se conserva la inicial G y las consonantes siguientes se transforman según su sonido, eliminando las vocales internas y reduciendo consonantes dobles; así, ambas generan el código GM, reflejando que su pronunciación es prácticamente igual pese a la diferencia ortográfica. De manera análoga, "*Yenny*" y "*Yeni*" se codifican como YN, ya que el algoritmo elimina las vocales internas y reduce las consonantes repetidas, indicando que fonéticamente suenan de manera equivalente, aunque se escriban distinto.

Algoritmo de Lynch y Arends (L&A)

Este algoritmo fonético fue desarrollado por Statistics Canada (Lynch, Arends, et al. 1977) y es una alternativa simple y eficiente para la codificación fonética de nombres, ampliamente utilizada en censos y procesos de vinculación de registros administrativos en Canadá. Este método es útil cuando se requiere una solución rápida, pero con capacidad de captura de errores comunes de transcripción y ortografía. Es especialmente relevante en contextos de censos de población y encuestas de gran escala en países de América Latina y el Caribe, donde los nombres pueden tener múltiples variantes fonéticas y ortográficas debido a la diversidad cultural y proporciona una forma simplificada de agrupación fonética que no depende del idioma, a diferencia de algoritmos como Soundex o Metaphone.

Su enfoque se centra en simplificar las cadenas de texto mediante la eliminación de las vocales, preservando únicamente la estructura consonántica del nombre, que suele contener la información fonética más relevante para la identificación. Además, el algoritmo reduce los sonidos duplicados o consecutivos, unificando repeticiones que frecuentemente se originan por diferencias en la forma de escribir un mismo nombre. Asimismo, este algoritmo no recodifica letras individuales ni aplica reglas complejas de sustitución, lo que disminuye la carga computacional y permite una implementación más rápida y eficiente. Gracias a estas características, se considera especialmente útil en aplicaciones de registro civil, censos y procesos de emparejamiento de datos donde se requiere alta velocidad de procesamiento y una razonable tolerancia a errores de escritura.

Según el algoritmo, "*Gómez*" y "*Gomes*" se codifican de la misma forma, ya que el método elimina las vocales y conserva solo las consonantes, además de no distinguir entre acentos ni mayúsculas/minúsculas. Aplicando las reglas paso a paso: "*Gómez*" → GMZ, "*Gomes*" → GMS. Sin embargo, el algoritmo también reduce duplicaciones o sonidos equivalentes que puedan deberse a variaciones ortográficas. En este caso, "*z*" y "*s*" representan un sonido similar al final del nombre, por lo

que el algoritmo los considera equivalentes o intercambiables. Por tanto, ambos apellidos quedarían codificados igual → GMS, reflejando su similitud fonética. En el caso de "*Yenny"* y "*Yeni"* se codifican como YN.

Algoritmo LatAm adaptado

En América Latina los nombres presentan una gran diversidad fonética y ortográfica influenciada por lenguas indígenas, castellano, portugués y otras tradiciones europeas. Por ello, es posible desarrollar un algoritmo que tenga en cuenta las transformaciones fonéticas y ortográficas más comunes en la región. Para este documento, los autores diseñaron un algoritmo especial que captura las variantes más frecuentes en los nombres latinoamericanos, con el objetivo de reducir las variaciones ortográficas y conservar únicamente los elementos sonoros más distintivos. Las reglas del algoritmo se enuncian a continuación:

i) En primer lugar, se unifican las duplicidades y sílabas características; es decir, ser convierte *ll* → *y*, *qu* → *k* y *ch* → *x*.

ii) Posteriormente, se estandarizan combinaciones ortográficas que producen el mismo sonido, convirtiendo *ce* → *se*, *ci* → *si*, y sustituyendo *gue* → *ge* y *gui* → *gi*.

iii) A continuación, se incorpora reglas específicas para reflejar particularidades fonéticas de nombres de origen quechua o aimara, aplicando transformaciones como *j* → *y*, *hua* → *wa* y *hu* → *w* cuando aparecen al inicio de la palabra.

iv) Luego, se normalizan los caracteres eliminando acentos y diacríticos, y reemplazando la letra *ñ* → *n*; de tal forma que todos los nombres se representen con un conjunto uniforme de caracteres.

v) Después, se eliminan las vocales y letras mudas, con el propósito de conservar únicamente la estructura consonántica que define el sonido esencial del nombre.

vi) Conversión de *v* → *b*, y de *z* → *s*, fonéticamente indistinguibles en la mayoría de los dialectos del español latino.

El orden en que se aplican las transformaciones también juega un rol especial, el usuario puede ampliar las reglas si así lo desea, incorporando nuevas líneas. Para el caso de "*Gómez*" y "*Gomes*", ambos apellidos se codificarían de igual manera "*Gómez*" → GMS, "*Gomes*" → GMS. De manera similar, "*Yenny"* y "*Yeni"* se codifican como YN.

Ejemplo

En esta sección se presenta la aplicación de distintos algoritmos fonéticos a nombres y apellidos, con el propósito de evaluar su capacidad para identificar coincidencias entre registros que pueden tener variaciones ortográficas. El Cuadro 8 muestra los resultados obtenidos al aplicar dichos algoritmos sobre un conjunto de nombres seleccionados. Se observa que el algoritmo LatAm adaptado ofrece un desempeño más consistente y preciso que los demás métodos considerados.

Cuadro 8
Codificación de nombres con algunos algoritmos de codificación fonética

Nombre	Soundex	Metaphone	L&A	LatAm
Wilmer	W456	WLMR	WLMR	WLMR
Guilmer	G456	KLMR	GLMR	GLMR
Wilmar	W456	WLMR	WLMR	WLMR
Yohana	Y500	YHN	YHN	YN

Nombre	Soundex	Metaphone	L&A	LatAm
Johanna	J500	JHN	JHN	YN
Bryan	B650	BRYN	BRN	BRYN
Brayan	B650	BRYN	BRN	BRYN
Marleni	M645	MRLN	MRLN	MRLN
Marleny	M645	MRLN	MRLN	MRLN
Marlenni	M645	MRLN	MRLN	MRLN
Nely	N400	NL	NL	NL
Neli	N400	NL	NL	NL
Nelly	N400	NL	NL	NL
Ximena	X550	SMN	XMN	YMN
Jimena	J550	JMN	JMN	YMN

Fuente: Elaboración propia.

De la misma manera, el cuadro 9 presenta la aplicación de los algoritmos fonéticos a algunos apellidos. Es importante tener en cuenta las particularidades culturales de cada región, ya que pueden influir significativamente en la forma en que son escritos o pronunciados. Estas variaciones hacen que ningún algoritmo de codificación fonética sea completamente robusto por sí solo, por lo que es recomendable adaptar o complementar los métodos según el contexto local.

Cuadro 9
Codificación de apellidos con algunos algoritmos de codificación fonética

Apellido	Soundex	Metaphone	L&A	LatAm
Huanca	H520	HNK	HNC	WNK
Wuanca	W520	WNK	WNC	WNK
Guanca	G520	KNK	GNC	GNK
Kuispe	K210	KSP	KSP	KSP
Quispe	Q210	KSP	QSP	KSP
Kispe	K210	KSP	KSP	KSP
Qhispe	Q210	KHSP	QHSP	KSP
Rodriguez	R362	RTRKS	RDRG	RDRGS
Rodrigues	R362	RTRKS	RDRG	RDRGS
Rodriwues	R362	RTRWS	RDRW	RDRWS
Ñahui	N000	NH	NH	NW
Nahui	N000	NH	NH	NW
Nahuy	N000	NH	NH	NW
Ñawi	N000	NW	NW	NW
Ñahui	N000	NH	NH	NW

Fuente: Elaboración propia.

B. Indexación

Una vez que las tablas de datos provenientes de las muestras E y P se encuentren limpias y estandarizadas, se procede a emparejadas. Inicialmente, cada registro de la tabla del censo necesita compararse con todos los registros de la tabla de la encuesta. Esto conduce a un número total de comparaciones de pares de registros que es cuadrático respecto al tamaño de las tablas de datos a emparejar. Por ejemplo, en el caso de Colombia, su censo del año 2018 tuvo una enumeración de más de 44 millones de personas y usó una EPC de 283 mil personas, lo que originaría más de 12 billones de comparaciones de pares de registros. Incluso si se pudieran realizar 100 mil comparaciones por segundo, el proceso de comparación tomaría más de 33 mil horas, más de mil días, que equivale a casi 4 años. Para reducir este costo se utilizan técnicas de indexación (*blocking)*, que limitan las comparaciones a subconjuntos plausibles de registros. Por lo anterior es necesario realizar una optimización del proceso combinado con un proceso de procesamiento en paralelo y de ser posible sistemas de computación distribuidos (como Apache Spark).

Como se mencionó anteriormente, en las muestras de cobertura, las UPM usan segmentos cartográficos equivalentes a los del censo, es decir, el código del segmento se refiere a la misma área geográfica, y en consecuencia es más probable que una persona que vive, por ejemplo, en el segmento A001 de la muestra de la EPC, también se encuentre en el segmento A001 del censo; así que comparar los pares de registros dentro del mismo segmento será la primera alternativa. Sin embargo, cuando ha pasado mucho tiempo entre la recolección de los datos del censo y la EPC, empieza a ser mayor la probabilidad de que las personas se encuentren en un segmento diferente, esto debido a que las personas se pueden trasladar de domicilio. En ese caso el enfoque de bloqueo pierde oportunidad porque las personas pueden encontrarse en segmentos diferentes; algo similar ocurre las personas que el día del censo no están en su lugar de residencia habitual. Otros ejemplos más complejos pueden darse cuando una mujer se ha casado y cambia su apellido y dirección, y por lo tanto no es detectada por los criterios de bloqueo y tampoco se detectaría en la comparación completa.

En este sentido, si n denota el tamaño de la muestra de la EPC, N_{1+} la cantidad de personas enumeradas en el censo y n_E la cantidad de personas enumeradas en la muestra E. Los pasos de la indexación son:

i) Realizar el emparejamiento entre la muestra E y la muestra P. Suponga que $C^{(1)}$ es el conjunto de personas emparejadas en este paso, y que $n_1 < n_0$ representa la cantidad de personas emparejadas, entonces $P^{(1)}$ es el conjunto de personas de la muestra P que no fueron emparejadas y $m_1 = n - n_1$ es la cantidad de personas que no fueron emparejadas en este paso.

ii) Sea $M^{(2)}$ la muestra de segmentos en un área más grande alrededor de cada segmento de la muestra P, esto para generar los nuevos bloques de indexación, es decir, si el segmento de la muestra P es una manzana cartográfica entonces el bloque podría ampliarse a una sección cartográfica o barrio para generar una búsqueda en un área mayor pero sin que se desborde la cantidad de comparaciones.

iii) Sea $E_2 = M^{(2)} - C^{(1)}$ la muestra de enumeración en un área más grande luego de retirar los elementos que ya fueron emparejados.

iv) Realizar el emparejamiento entre la muestra E_2 y la muestra $P^{(1)}$. Suponga que $C^{(2)}$ es el conjunto de personas emparejadas en este paso, donde $n_2 < m_1$ es la cantidad de personas emparejadas, entonces $P^{(2)}$ es el conjunto de personas de la muestra $P^{(1)}$ que no fueron emparejadas y $m_2 = m_1 - n_2$ es la cantidad de personas que no fueron emparejadas en este paso.

v) Sea $M^{(3)}$ la muestra de segmentos en un área más grande alrededor de cada bloque usado en $M^{(2)}$, es decir, si en el paso anterior el bloque se amplió a una sección cartográfica entonces ahora se puede ampliar a un sector censal o si era el barrio entonces ampliarlo a una zona catastral más grande, y así generar una búsqueda en un área mayor pero sin que se desborde la cantidad de comparaciones.

vi) Sea $E_3 = M^{(3)} - \bigcup_{i=1}^{2} C^{(i)}$ la muestra de enumeración en un área más grande luego de retirar los elementos que ya fueron emparejados.

vii) Realizar el emparejamiento entre la muestra E_3 y la muestra $P^{(2)}$. Ahora $C^{(3)}$ es el conjunto de personas emparejadas en este paso, donde $n_3 < m_2$ es la cantidad de personas emparejadas, entonces $P^{(3)}$ es el conjunto de personas de la muestra $P^{(2)}$ que no fueron emparejadas y $m_3 = m_2 - n_3$ es la cantidad de personas que no fueron emparejadas en este paso.

viii) Continuar el procedimiento hasta que $M^{(j)}$ sea igual al censo o hasta que $m_j = 0$, es decir, que no hay elementos sin emparejar.

C. Comparación

El paso de comparación constituye una etapa central en el proceso de emparejamiento de registros, ya que permite cuantificar el grado de similitud entre los pares potenciales identificados durante la indexación. Su objetivo principal es determinar, a partir de distintas variables, cuán probable es que dos registros se refieran a la misma entidad, aun cuando existan errores tipográficos, diferencias en la escritura o variaciones en la representación de la información.

Para ello, se utilizan diversas medidas de similitud o distancia según el tipo de variable que se compare. En el caso de los textos o cadenas de caracteres, la distancia de Levenshtein calcula el número mínimo de operaciones (inserciones, eliminaciones o sustituciones) necesarias para transformar una cadena en otra, ofreciendo una medida sensible a pequeños errores de digitación. La medida de Jaro-Winkler, por su parte, se orienta a capturar la similitud en nombres o apellidos, dando mayor peso a los prefijos coincidentes y siendo especialmente útil cuando los errores se concentran hacia el final de las palabras. En el caso de variables espaciales o geográficas, la distancia de Haversine permite estimar la separación en metros o kilómetros entre dos puntos definidos por sus coordenadas de latitud y longitud, proporcionando una comparación geográfica directa.

El objetivo final de este paso es generar un conjunto de puntajes de similitud que sirvan de insumo para la etapa de decisión o clasificación, en la cual se determina si cada par de registros debe considerarse un vínculo verdadero, un vínculo falso o un caso dudoso que requiere revisión manual. De esta manera, la comparación no solo cuantifica la cercanía entre registros, sino que sienta las bases para una identificación precisa y eficiente de coincidencias en bases de datos complejas. A continuación, se describen algunas de las métricas que son más utilizadas.

Distancia de Levenshtein

Esta es una métrica que calcula el número mínimo de operaciones de edición (inserciones, eliminaciones y sustituciones) necesarias para transformar una cadena de texto en otra. Siendo s_1 y s_2 dos cadenas de texto diferentes, con longitudes $|s_1|$ y $|s_2|$, la distancia de Levenshtein $dist_{lev}(s_1, s_2)$, la cual puede ser normalizada y convertirse en una medidad de similitud, que se calcula de la siguiente manera:

$$sim_{lev}(s_1, s_2) = 1 - \frac{dist_{lev}(s_1, s_2)}{max(|s_1|, |s_2|)}$$

Por ejemplo, suponga que se tienen las cadenas $s_1 =$ "*Gómez*" y $s_2 =$ "*Gomes*". Para transformar "*Gómez* " en "*Gomes* " se deben sustituir dos letras: *ó* $\rightarrow$ *o* y *z* $\rightarrow$ *s*. Por lo tanto, la distancia de Levenshtein es 2. Considerando que las longitudes de las palabras son $|s_1| = |s_2| = 5$, la similitud se calcula como $sim_{lev}(s_1, s_2) = 1 - \frac{2}{5} = 0.6$. Esto indica que "*Gómez*" y "*Gomes*" son moderadamente similares, con un 60% de coincidencia según esta métrica. Asimismo, Para transformar "*Yenny*" en "*Yeni*" se requieren dos operaciones (una sustitución y una eliminación), lo que da una distancia de 2 y una similitud de 0.6, reflejando también una similitud moderada. Estos ejemplos muestran cómo la distancia de Levenshtein permite medir cuantitativamente la cercanía entre cadenas de texto, útil para emparejamiento de nombres o corrección de datos.

Similitud de Jaro y Winkler

Esta métrica es utilizada para medir la cercanía entre dos cadenas de texto, y es especialmente útil para nombres propios y palabras cortas. A diferencia de la distancia de Levenshtein, que cuenta operaciones necesarias para transformar una cadena en otra, Jaro-Winkler se centra en la coincidencia de caracteres y su orden relativo. Además, considera las transposiciones, es decir, caracteres que coinciden, pero aparecen en posiciones diferentes. La métrica evalúa tanto la cantidad de caracteres coincidentes como cuán alineados están, otorgando una medida de similitud entre cero (sin coincidencia) y uno (coincidencia perfecta). Además, Jaro-Winkler incorpora un ajuste adicional que da más peso a los prefijos coincidentes al inicio de las cadenas, lo que es útil en nombres propios donde las primeras letras suelen ser más significativas. Por ejemplo, dos nombres que comienzan igual recibirán un puntaje más alto que otros que solo coinciden parcialmente al final.

El punto de partida es la similitud de Jaro, especialmente diseñada para nombres y toma en cuenta caracteres comunes y transposiciones (P. Christen 2012), y dada por la siguiente expresión:

$$sim_{jaro}(s_1, s_2) = \frac{1}{3}\left(\frac{c}{|s_1|} + \frac{c}{|s_2|} + \frac{c-t}{c}\right)$$

donde c es el número de caracteres coincidentes y t el número de transposiciones. Luego, la similitud de Jaro-Winkler es un ajuste a la de Jaro con base en un prefijo común:

$$sim_{winkler}(s_1, s_2) = sim_{jaro}(s_1, s_2) + p \cdot \left(1 - sim_{jaro}(s_1, s_2)\right) \cdot 0.1$$

donde p es el número de caracteres idénticos al inicio ($0 \leq p \leq 4$). Por ejemplo: Para las cadenas $s_1 =$ "*Gómez*" y $s_2 =$ "*Gomes*", primero se cuenta el número de caracteres coincidentes y se consideran las transposiciones. Las letras "G", "m" y "e" coinciden y están en posiciones similares, lo que da un total de tres coincidencias. Sustituyendo en la fórmula:

$$sim_{jaro}(Gómez, Gomes) = \frac{1}{3}\left(\frac{3}{5} + \frac{3}{5} + \frac{3-0}{3}\right) \approx 0.733$$

Como el prefijo inicial común es "*G*", con longitud $p = 1$, entonces $sim_{winkler}(Gómez, Gomes) \approx 0.760$. De la misma forma, para las cadenas $s_1 =$ "*Yenny*" y $s_2 =$ "*Yeni*", la similitud de Jaro es 0.783. Como el prefijo inicial común es "*Ye*", con longitud $p = 2$, entonces la similitud de Jaro-Winkler será igual a 0.826.

Comparación de fechas y edades

Las fechas y edades se comparan de manera directa, evaluando la diferencia en días, meses o años. Es fundamental definir previamente los rangos de tolerancia aceptables para considerar una coincidencia, por ejemplo, permitir una diferencia de hasta un año en la edad. Cuando se comparan edad y fecha de nacimiento, esta comparación permite además validar la coherencia temporal entre ambos datos. Como alternativa, las fechas pueden transformarse en edades y calcularse la diferencia en términos

porcentuales, lo que facilita incorporar cierto grado de tolerancia. En este caso, las edades deben calcularse respecto a una fecha de referencia, que puede ser la fecha de cierre de la EPC, la fecha del emparejamiento de las bases de datos o cualquier otra fecha relevante para el contexto del análisis.

Supongamos que d_1 y d_2 representan la edad (en días o años) calculada desde la fecha fija. Entonces, la diferencia porcentual de edad (*dpe*) se calcula como:

$$dpe = \frac{/d_1 - d_2/}{max\,(d_1, d_2)} \cdot 100\ \%.$$

Con base en este valor, se puede calcular la similitud porcentual de edad como:

$$sim_{edad} = \begin{cases} 1 - \dfrac{dpe}{dpe_{max}}, & si\ dpe < dpe_{max} \\ 0, & en\ otro\ caso \end{cases}$$

En donde $dpe_{max} \in (0, 100)$ representa la diferencia porcentual máxima tolerada (P. Christen 2012). Por ejemplo, si se comparan dos registros donde una persona tiene 30 años y otra tiene 31 años, la diferencia porcentual de edad sería $dpe = |\ 30 - 31\ |/31 \times 100 = 3{,}23\%$. Por otro lado, si se establece una diferencia máxima tolerada $dpe_{max} = 10\%$, la similitud porcentual de la edad estaría dada por $sim_{edad} = 1 - 3{,}23/10 = 0{,}677$; es decir, 67,7% de similitud. En cambio, si se comparan edades de 25 y 40 años, la diferencia porcentual sería 37,5%, y dado que este valor supera el umbral del 10%, la similitud porcentual se fija en 0, indicando que las edades son demasiado distintas para considerarse similares.

Comparación geográfica

Para campos geográficos, como coordenadas o nombres de lugares, es posible cuantificar la cercanía mediante distancias euclidianas o geodésicas. Un ejemplo común es la fórmula de Haversine, que permite calcular la distancia entre dos puntos sobre la superficie de una esfera a partir de sus coordenadas de latitud y longitud. Si se dispone de las coordenadas, la distancia se calcula como:

$$d = 2r \cdot arcsin\left(\sqrt{sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + cos(\phi_1) \cdot cos(\phi_2) \cdot sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

donde ϕ es la latitud, λ la longitud y r el radio de la Tierra. Este enfoque permite evaluar distancias precisas entre puntos, independientemente de la escala, y puede aplicarse desde el nivel de país hasta barrios o manzanas. Alternativamente, cuando no se dispone de coordenadas exactas, se pueden utilizar nombres de lugares o códigos administrativos normalizados, lo que facilita la comparación geográfica basada en jerarquías territoriales y niveles administrativo.

Por ejemplo, en Colombia, si se comparan dos viviendas localizadas dentro del mismo municipio (Bogotá), una en el barrio La Soledad (latitud 4.627° N, longitud -74.074° O) y otra en el barrio Teusaquillo (latitud 4.638° N, longitud -74.085° O), la distancia calculada mediante la fórmula de Haversine es aproximadamente $d = 1{,}6$ kilómetros, lo que indica una cercanía espacial alta dentro del mismo municipio. En contraste, si se comparan una vivienda en Bogotá (4.711° N, -74.072° O) y otra en el municipio de Chía (4.861° N, -74.058° O), también en el departamento de Cundinamarca, la distancia sería $d = 17$ kilómetros, reflejando que, aunque están en la misma región administrativa, pertenecen a municipios diferentes y la proximidad geográfica es considerablemente menor.

D. Clasificación

Una vez que los registros de las muestras E (del censo) y de la muestra P (de la encuesta) han sido depurados y estandarizados, es necesario determinar si cada registro de la encuesta puede vincularse o no con un registro censal correspondiente. Este proceso se conoce como emparejamiento de registros (*record linkage*), y su objetivo es identificar qué pares de registros representan a la misma persona, hogar o unidad de análisis. El problema central consiste en decidir, para cada posible par de registros, si ambos corresponden a la misma entidad (*match*) o si pertenecen a personas distintas (*non-match*). Dado el gran número de posibles combinaciones, se utilizan modelos probabilísticos que permiten tomar decisiones de manera sistemática y objetiva.

El método clásico para la clasificación de pares fue propuesto por Fellegi y Sunter (1969). Cada par posible $(a, b) \in A \times B$ puede pertenecer a uno de dos grupos: M, que es el conjunto de pares que representan la misma entidad, y U, que es el conjunto de pares que no corresponden al mismo individuo. Para cada par se define un vector de comparación:

$$\gamma(a, b) = (\gamma_1, \gamma_2, \dots, \gamma_p),$$

En donde p es el número de variables comparadas (por ejemplo, nombre, sexo, edad o fecha de nacimiento). Cada componente γ_j toma el valor 1, si existe coincidencia en el atributo j, y 0 en caso contrario. El modelo estima la probabilidad de observar un patrón de coincidencias γ bajo los dos escenarios posibles, cuando el par es un emparejamiento (M - *match*) o cuando el par resultar un no emparejamiento (U - *non-match*) y, con base en ello, se calcula la razón de verosimilitud, que resume la evidencia de que los dos registros corresponden a la misma entidad. Al aplicar logaritmo, se tiene que:

$$log\, L(\gamma) = log\ L(\gamma \mid M) - log\ L(\gamma \mid U)$$

Estos parámetros pueden estimarse mediante métodos de máxima verosimilitud, como el algoritmo EM o mediante enfoques bayesianos (Winkler, 2000; Larsen y Rubin, 2001). Cuanto mayor sea el valor de $log\, L(\gamma)$, mayor será la probabilidad de que el par (a, b) sea un emparejamiento verdadero. Para clasificar los pares, el modelo utiliza dos umbrales de decisión:

- Si $log\, L(\gamma) \geq T_M$: el par se clasifica como emparejado (*match*).
- Si $log\, L(\gamma) \leq T_U$: el par se clasifica como no emparejado (*non-match*).
- Si $T_U < log\, L(\gamma) < T_M$: el par se clasifica como emparejamiento potencial, el cual se revisa clericalmente para confirmar o descartar la coincidencia.

De esta forma, el modelo proporciona un mecanismo transparente para decidir, con base en la evidencia, si los registros de la EPC y del censo se refieren o no al mismo individuo. El método de Fellegi y Sunter puede complementarse con enfoques modernos de aprendizaje automático, tanto supervisados como no supervisados.

Estos modelos utilizan pares previamente clasificados para entrenar algoritmos que aprenden a distinguir patrones de similitud entre registros, incorporando características fonéticas, textuales o geográficas. Su aplicación permite mejorar la calidad del emparejamiento en contextos donde los nombres presentan errores ortográficos, las fechas pueden estar incompletas o los datos contienen inconsistencias. En estos casos, los pares de registros se representan como vectores de características derivadas de la comparación y se utilizan reglas de clasificación que buscan maximizar las coincidencias reales, para más detalles se recomienda consultar (P. Christen 2012, Capítulo 6).

Para ilustrar el procedimiento, supóngase que se comparan tres registros. Uno proviene de la EPC, con nombre "Yenny Gómez", sexo femenino y año de nacimiento 1988. Los otros dos provienen del censo: uno con nombre "Yeni Gomes", sexo femenino y año de nacimiento 1988, y otro con nombre

"Jenifer Gonzáles", sexo femenino y año de nacimiento 1988. Aplicando el algoritmo Soundex a los nombres y apellidos, se obtiene la representación fonética de cada elemento. A partir de esta codificación y considerando la coincidencia en sexo y año de nacimiento, el modelo de Fellegi-Sunter calcula un puntaje de coincidencia para cada par. El cuadro 10 resume la comparación.

Cuadro 10
Clasificación de registros basados en la codificación Soundex

Registro EPC	Registro Censo	Nombre Soundex	Apellido Soundex	Puntaje aproximado	Decisión
Yenny Gómez	*Yeni Gomes*	Y500	G520	Alto	Emparejado
Yenny Gómez	*Jenifer Gonzales*	Y500	G524	Bajo	No emparejado

Fuente: Elaboración propia.

E. Evaluación

Como Christen (2012) afirma, las técnicas de clasificación para el emparejamiento de datos buscan maximizar la calidad de los resultados. No obstante, evaluar dicha calidad requiere la existencia de un conjunto de referencia, es decir, un conjunto donde se conozca con certeza si cada par de registros corresponde a la misma entidad o no. Esta información debe reflejar fielmente las características de los datos reales bajo análisis. En el contexto de censos y encuestas de posenumeración censal, un emparejamiento correcto implica que un registro del censo y uno de la encuesta representan a la misma persona. De manera análoga, un par no emparejado representa dos entidades distintas.

La disponibilidad de datos de referencia permite calcular métricas similares a las usadas en modelos de aprendizaje automático para problemas de clasificación binaria (Menestrina, Whang, y Garcia-Molina 2010). En la práctica, los conjuntos de referencia necesarios para evaluar el desempeño de los procesos de emparejamiento rara vez están disponibles de manera directa. Por esta razón, es habitual implementar procedimientos de codificación manual, que consisten en seleccionar un subgrupo de registros de la muestra P y realizar una verificación clerical (manual) con los registros de la muestra E, con el fin de determinar la correspondencia real entre ambos.

Este proceso, aunque fundamental para validar los resultados, puede resultar costoso y demandante, especialmente cuando se utilizan esquemas de muestreo estratificado que requieren una cantidad considerable de revisiones manuales. Una vez disponible el conjunto de referencia, los pares de registros se clasifican en cuatro categorías básicas (Christen, 2012): verdaderos positivos (VP), que corresponden a pares correctamente emparejados; falsos positivos (FP), que representan emparejamientos incorrectos; verdaderos negativos (VN), que son pares correctamente identificados como no coincidentes; y, finalmente los falsos negativos (FN), que son pares que deberían haberse emparejado, pero no lo fueron.

Para evaluar la calidad y el desempeño de los procesos de emparejamiento, se emplean métricas que permiten cuantificar la capacidad del algoritmo para identificar correctamente los pares verdaderos y evitar emparejamientos incorrectos. Entre las métricas más utilizadas se encuentran las siguientes:

i) Precisión (P): proporción de emparejamientos correctos entre todos los que fueron clasificados como positivos; es decir, indica cuántos de los pares identificados como coincidencias son realmente verdaderos.

$$P = \frac{VP}{VP + FP}$$

ii) Exhaustividad (R): proporción de emparejamientos reales que fueron efectivamente detectados, mostrando la capacidad y eficacia del método para recuperar todas las coincidencias verdaderas.

$$R = \frac{VP}{VP + FN}$$

iii) Medida-F (F): combina ambas métricas en una única medida de desempeño mediante su media armónica, equilibrando la precisión y la exhaustividad.

$$F = 2 \cdot \frac{P \cdot R}{P + R}$$

En la práctica, estas métricas pueden mostrar comportamientos distintos según las características del conjunto de datos y el umbral de decisión utilizado. Una alta precisión implica que la mayoría de los emparejamientos identificados son correctos, aunque podría perderse una parte importante de los verdaderos pares (baja exhaustividad). En cambio, una alta exhaustividad asegura que la mayoría de los pares verdaderos son detectados, pero a costa de incluir también emparejamientos incorrectos (baja precisión). La medida F permite encontrar un equilibrio entre ambos extremos, proporcionando una evaluación más integral del desempeño del proceso.

Además de las métricas orientadas a medir la calidad del emparejamiento, es fundamental evaluar la eficiencia del proceso, pues el número de registros que se deben comparar es muy grande y el emparejamiento completo entre todas las combinaciones posibles resultaría inviable. En este sentido, las etapas de indexación o bloqueo cumplen un papel crucial, pues permiten reducir el espacio de búsqueda al comparar únicamente los registros que comparten ciertas características comunes. Para analizar el desempeño de estas etapas preliminares, se emplean tres métricas complementarias:

i) La reducción, que mide la proporción de pares descartados durante el bloqueo respecto del total de combinaciones posibles, reflejando la eficiencia del procedimiento para disminuir el número de comparaciones.

ii) La completitud de pares, que indica la proporción de verdaderos emparejamientos que fueron efectivamente retenidos después del bloqueo, mostrando qué tan bien el método logra preservar las coincidencias reales.

iii) La calidad de pares expresa la proporción de los pares retenidos que corresponden a verdaderos emparejamientos, proporcionando una idea de la calidad e integridad del conjunto resultante.

A pesar de lo anterior, en aplicaciones reales de EPC, el emparejamiento automático entre la muestra E y la muestra P suele ser insuficiente. Por esta razón, es común implementar procesos de revisión manual (revisión clerical) que son realizadas por un equipo de expertos, quienes validan los posibles emparejamientos ambiguos o dudosos. La calidad y consistencia de esta revisión dependen de diversos factores. Entre ellos destacan la experiencia y capacitación del personal revisor, que influye directamente en la coherencia de las decisiones; la disponibilidad de herramientas informáticas adecuadas, que faciliten la comparación contextual de los registros (por ejemplo, mostrando coincidencias potenciales o agrupando personas por hogar); y el acceso a fuentes de información complementarias, como historiales de direcciones, variantes de nombres o registros administrativos.

En los casos de emparejamientos potenciales o situaciones sospechosas que requieren verificación adicional, es posible realizar nuevas entrevistas a los hogares para recolectar información complementaria que permita clasificar cada caso de manera definitiva como emparejado o no. Tal como lo señala United Nations (2010), las visitas de reconciliación constituyen una etapa adicional del proceso operativo de la EPC, pues se llevan a cabo después del emparejamiento preliminar y antes de consolidar el emparejamiento definitivo. Su propósito es aclarar los casos no emparejados o con información

insuficiente, garantizando que cada persona y hogar de las muestras P y E reciba un estatus de coincidencia final correcto.

Durante estas visitas, los equipos de campo indagan en los segmentos seleccionados para verificar, por ejemplo, cuando una persona aparece en los registros del censo pero no en la EPC, lo que permite evaluar si se trató de una enumeración correcta o errónea; cuando una persona figura en la EPC pero no en el censo, lo que ayuda a establecer si era residente habitual en la fecha censal, si se omitió en el operativo censal o si corresponde a un recién llegado o nacido con posterioridad; o cuando existen dudas sobre la residencia habitual o la pertenencia a un hogar en la fecha de referencia. Tal como lo señalan Borges y Queiroz (2025), en Brasil esta verificación se usó para mejorar el emparejamiento.

Además de validar el estatus de emparejamiento, estas nuevas visitas también permiten descartar casos con información insuficiente y asegurar la coherencia de los datos utilizados en la estimación final de la cobertura. Es importante distinguir que, mientras el censo realiza la enumeración completa de la población y la EPC genera un recuento independiente para evaluar su cobertura, las visitas de reconciliación constituyeron un componente complementario orientado exclusivamente a mejorar la precisión del emparejamiento entre ambas fuentes y, con ello, fortalecer la calidad de las estimaciones de omisión censal.

F. Ejemplo aplicado

En este ejemplo se presenta un ejercicio ilustrativo del proceso de emparejamiento entre los registros de una Encuesta de Posenumeración Censal (EPC) y los del Censo de población, con el fin de estimar la omisión censal. Los datos, completamente simulados, corresponden a un conjunto reducido de viviendas de un barrio ficticio, en una ciudad latinoamericana. Cada registro contiene información básica sobre el nombre, el sexo, la edad, la fecha de nacimiento, la dirección y las coordenadas geográficas aproximadas de la vivienda. El cuadro 11 muestra los registros de ambas fuentes.

Cuadro 11
Registros simulados entre la EPC y el censo

Fuente	Nombre	Sexo	Edad	Fecha de nacimiento	Dirección	Latitud	Longitud
EPC	*Yeni Gomes*	F	33	3/15/1989	Calle 12 #45-8	-12,057	-77,0850
EPC	*Jennifer Gonzales*	F	28	7/20/1995	Calle 14 #50-19	-12,056	-77,0846
EPC	*Luis A. Ramos*	M	41	10/10/1982	Av. Central 102	-12,057	-77,0840
Censo	*Yenny Gómez Pérez*	F	34	3/12/1989	Calle 12 #45-08	-12,056	-77,0851
Censo	*Jenifer González*	F	28	7/22/1995	Calle 14 #50-21	-12,056	-77,0845
Censo	*Luis Alberto Ramos*	M	41	10/9/1982	Av. Central 102	-12,057	-77,0839
Censo	*Carlos Mejía Torres*	M	52	11/18/1973	Calle 16 #48-10	-12,058	-77,0862

Fuente: Elaboración propia.

El proceso inicia con la geolocalización y verificación espacial de los registros. Se constata que todos pertenecen al área de estudio. Luego, se calculan las distancias geodésicas entre coordenadas de posibles pares. Por ejemplo, entre *Yeni Gomes* (EPC, latitud -12.0565; longitud -77.0850) y *Yenny Gómez Pérez* (Censo, latitud -12.0564; longitud -77.0851) la distancia es de 15 metros, lo que sugiere una localización prácticamente coincidente. El par *Jennifer Gonzales* (EPC, -12.0558; -77.0846) y *Jenifer González* (Censo, -12.0559; -77.0845) muestra una separación de 13 metros, correspondiente a viviendas contiguas. De forma similar, *Luis A. Ramos* (EPC, -12.0569; -77.0840) y *Luis Alberto Ramos* (Censo, -12.0568; -77.0839) se encuentran a 11 metros de distancia. En contraste, *Carlos Mejía Torres* no tiene

ninguna observación de la EPC con coordenadas cercanas (la más próxima se encuentra a más de 120 metros), lo que anticipa la ausencia de emparejamiento.

Posteriormente, se aplica la codificación fonética Metaphone para capturar las similitudes en la pronunciación de los nombres y apellidos, aun cuando existan variaciones ortográficas. Los códigos generados para cada registro se resumen en el cuadro 12. Como puede observarse, los pares *Yeni–Yenny*, *Gomes–Gómez*, *Jennifer–Jenifer* y *Gonzales–González* comparten códigos idénticos o prácticamente idénticos, lo cual respalda su coincidencia potencial. En cambio, *Carlos Mejía Torres* no presenta ninguna equivalencia fonética con otros registros.

Cuadro 12
Codificación Metaphone de nombres y apellidos para los registros simulados

Nombre original	Nombre Metaphone	Apellido Metaphone
Yeni Gomes	YN	GMS
Jennifer Gonzales	JNFR	JNSLS
Luis A. Ramos	LS	RMS
Yenny Gómez Pérez	YN	GMS
Jenifer González	JNFR	JNSLS
Luis Alberto Ramos	LS	RMS
Carlos Mejía Torres	KRLS	MHTRS

Fuente: Elaboración propia.

Con la información ya codificada, se realiza la indexación y *blocking*. En este ejemplo, se definen bloques únicamente por sexo, de modo que los registros femeninos y masculinos se procesan por separado. El bloque femenino, cuenta con 2 registros de la EPC y 2 del Censo, generando cuatro comparaciones posibles, mientras que el bloque masculino, tiene 1 registro de la EPC y 2 del Censo, lo que produce 2 comparaciones, dando un total de 6 comparaciones en todo el proceso. Esta estrategia reduce el número de comparaciones, garantizando que solo se contrasten registros con alta probabilidad de pertenecer a la misma población demográfica, puesto que, si no se aplicara el *blocking* y se comparara cada registro de la EPC con todos los del Censo, el número total de comparaciones sería igual $3 \times 4 = 12$ comparaciones.

Dentro de cada bloque se calculan los índices de similitud. Para las variables textuales (nombres y apellidos) se utiliza la distancia de Jaro–Winkler, que asigna valores entre 0 y 1 según la proximidad entre cadenas. El par (*Yeni Gomes, Yenny Gómez)* alcanza una similitud de 0.93; *(Jennifer Gonzales, Jenifer González)* obtiene 0.96; y *(Luis A. Ramos, Luis Alberto Ramos)*, 0.91. La similitud entre (*Luis A. Ramos, Carlos Mejía Torres)* es baja, aproximadamente 0.48, lo que refleja que comparten pocos caracteres en posiciones similares y, por tanto, es poco probable que se trate del mismo individuo. Un siguiente paso consiste en la evaluación de las fechas de nacimiento y edades. En los tres pares potencialmente coincidentes las diferencias son mínimas (de uno a tres días en la fecha o un año en la edad, atribuible a redondeo), por lo que se asignan similitudes altas (≥0.9). En el caso de *Carlos Mejía Torres*, no existe ninguna coincidencia temporal con los registros de la EPC, por lo que su similitud es prácticamente nula.

En la siguiente fase se efectúa la limpieza y normalización de los textos: se eliminan tildes, se unifican mayúsculas y se corrigen formatos de dirección, reemplazando símbolos o abreviaturas inconsistentes. Así, "*Calle 12 #45-8*" y "*Calle 12 #45-08*" se transforman en "*CALLE 12 45 08*". Este paso mejora la comparabilidad entre las dos fuentes y reduce errores debidos a digitación o diferencias ortográficas. A continuación, se realiza la comparación geográfica, utilizando tanto la información textual de dirección como la distancia entre coordenadas. Para el par (*Yeni Gomes, Yenny Gómez)*, las

direcciones estandarizadas ("*CALLE 12 45 08*") coinciden plenamente, y las coordenadas (-12.0565, -77.0850) vs. (-12.0564, -77.0851) generan una distancia de 15 metros, equivalente a una similitud espacial de 1.0. El par *(Jennifer Gonzales, Jenifer González)* presenta una diferencia mínima en la numeración de la vivienda (19 vs. 21) y distancia de 13 metros, por lo que su similitud espacial es de 0.95. Para el par *(Luis A. Ramos, Luis Alberto Ramos)*, las direcciones son idénticas y la distancia de 11 metros implica una similitud geográfica de 1.0. En contraste, *Carlos Mejía Torres* se ubica a más de 120 metros del registro más cercano, con similitud espacial de 0.2.

Una vez integradas las distintas dimensiones (nombre, fecha, edad y localización), se aplica el modelo probabilístico de Fellegi–Sunter, que clasifica los registros según la evidencia de coincidencia. El resultado es un puntaje global (*S*) para cada par, que permite clasificar los registros según su probabilidad de correspondencia. En este caso se considera un n umbrales operativos *S>3.5* para emparejamientos seguros, y *S<2.0* para no emparejados (*non-match*). El cuadro 13 resume los resultados finales, presentando primero los tres registros de la EPC y sus posibles emparejamientos en el Censo.

Cuadro 13
Clasificación de registros entre la EPC y el censo

Registro EPC	Registro Censo	Puntaje (S)	Clasificación
Yeni Gomes	*Yenny Gómez Pérez*	5,2	Emparejado
Jennifer Gonzales	*Jenifer González*	4,9	Emparejado
Luis A. Ramos	*Luis Alberto Ramos*	5,4	Emparejado
—	*Carlos Mejía Torres*	—	No emparejado
Yeni Gomes	*Yenny Gómez Pérez*	5,2	Emparejado

Fuente: Elaboración propia.

Los resultados indican que los tres registros de la EPC encontraron su correspondiente emparejamiento en la base censal, mientras que Carlos Mejía Torres no fue vinculado con ningún registro de la EPC. En un contexto real, este caso se consideraría un no emparejamiento, representando un posible ejemplo de omisión censal o de falta de cobertura de la encuesta de control.

Una vez concluido el proceso de emparejamiento, es posible evaluar su desempeño mediante indicadores de calidad basados en la comparación entre los resultados del algoritmo y la revisión clerical. En este ejemplo, la precisión, (proporción de emparejamientos identificados correctamente entre todos los declarados como tales), alcanza un valor de $P = 1.0$. La exhaustividad (proporción de emparejamientos verdaderos que fueron efectivamente detectados por el algoritmo) es de $R = 0.75$, al haberse omitido uno de los cuatro casos posibles. A partir de estos dos indicadores se calcula la medida F (media armónica entre precisión y exhaustividad) con un valor de $F = 0.86$. Estos resultados sugieren que el proceso de emparejamiento fue altamente preciso; es decir, que no generó falsos vínculos, aunque con una leve pérdida en la capacidad de recuperación, aspecto que en aplicaciones reales puede optimizarse ajustando los umbrales de decisión del modelo de Fellegi–Sunter o incorporando variables auxiliares adicionales.

Por último, para implementar este tipo de emparejamientos de forma automatizada, existen diversas herramientas en los lenguajes R y Python que facilitan su ejecución. En R, el paquete RecordLinkage (Sariyar & Borg, 2010) permite aplicar el modelo de Fellegi-Sunter y utilizar medidas de similitud como Soundex, Jaro-Winkler o Levenshtein, además de incorporar estrategias de bloqueo y métodos de clasificación supervisados y no supervisados. El paquete fastLink (Enamorado, Fifield & Imai, 2019) extiende el enfoque probabilístico del modelo de Fellegi-Sunter, maneja datos faltantes y permite estimar probabilidades de coincidencia con mayor eficiencia y escalabilidad. Por su parte,

fuzzyjoin (Robinson, Bryan & Elias, 2020) facilita la unión de tablas mediante coincidencias parciales empleando funciones de distancia o expresiones regulares, integrándose plenamente con el ecosistema tidyverse. Asimismo, stringdist (van der Loo, 2014) ofrece una amplia gama de métricas de distancia (incluyendo Levenshtein, Jaccard, Jaro y Hamming) muy útiles para comparar cadenas de texto. En el entorno Python, el paquete recordlinkage (de Bruin, 2019) implementa el modelo de Fellegi-Sunter y algoritmos de aprendizaje automático como Support Vector Machines y Random Forests, junto con herramientas avanzadas para el bloqueo y la evaluación del desempeño. Finalmente, la librería Dedupe (Gregg & Eder, 2022) emplea aprendizaje supervisado y semisupervisado, e integra técnicas de bloqueo adaptativo y agrupamiento para mejorar la eficiencia y precisión del emparejamiento.

V. Estimadores en el Sistema de Estimación Dual

El presente capítulo se centra en los estimadores de muestreo aplicados en el contexto del sistema de estimación dual (SED), combinando información proveniente del censo y de la EPC para estimar la omisión censal. Se revisan los principales métodos clásicos de estimación, comenzando con el enfoque de Petersen, que constituye la base del SED mediante la proporción de coincidencias entre ambas fuentes. A partir de esta referencia, se abordan extensiones y ajustes que mejoran la precisión frente a casos de heterogeneidad en la captura de registros, incluyendo los estimadores de Chao, Chapman, Nour, Webster-Kemp y Zelterman, cada uno con sus propiedades particulares y supuestos específicos sobre la probabilidad de inclusión de individuos.

Asimismo, el capítulo explora la integración de los modelos loglineales, que permiten un tratamiento flexible de la dependencia entre listas y la incorporación de covariables explicativas, ofreciendo una alternativa robusta frente a violaciones de los supuestos clásicos del sistema dual. Se discute cómo estos modelos permiten estimar la población total y la omisión censal de manera más realista, considerando interacciones y heterogeneidad entre subpoblaciones. La combinación de los estimadores tradicionales y los enfoques basados en modelos loglineales proporciona un enfoque complementario para la evaluación de cobertura censal y la interpretación de resultados de la EPC, facilitando la selección del método más adecuado según las características de la población y la calidad de los datos disponibles.

A. Estimadores de muestreo

Al presentar el SED, se partió de un supuesto simplificador: que tanto el censo como la encuesta lograrían captar a todos los miembros de la población. Esta hipótesis, aunque útil para construir el marco teórico inicial, rara vez se cumple en los estudios reales sobre error de cobertura. En la práctica, es poco probable que los individuos de una población sean observados en ambas fuentes de información. En particular, si bien todos tienen una probabilidad positiva de ser incluidos en la EPC, solo un pequeño subconjunto de ellos será efectivamente incluido en la muestra.

Por esta razón, resulta necesario ajustar el enfoque teórico para representar de manera más realista las condiciones bajo las cuales se desarrolla la inferencia.

En un escenario más cercano a la realidad, se asume que todos los miembros de la población están expuestos a ser incluidos en el censo, ya que este busca una cobertura completa del territorio y de la población objetivo. En cambio, solo una muestra de esa población es seleccionada para la muestra de la EPC. Esta diferencia introduce una asimetría esencial en la relación entre ambas fuentes de datos: mientras el censo constituye un esfuerzo exhaustivo de enumeración, la encuesta representa una medición parcial basada en un diseño muestral. Reconocer esta asimetría es clave para comprender las limitaciones del SED y para adaptar su formulación a contextos empíricos más realistas.

Este cambio en los supuestos implica una modificación significativa en los métodos de análisis, ya que se altera la estructura de la información disponible y las cantidades que se consideran conocidas o desconocidas. Anteriormente, se podía asumir que ciertos totales poblacionales eran observables o directamente medibles, pero bajo este nuevo enfoque, solo el total del censo, denotado como N_{1+}, se considera observable; sin embargo, no es directamente conocido, puesto que el censo está expuesto a errores de enumeración y duplicaciones. Esto significa que el número de individuos capturados correctamente por el censo no es una cantidad que se toma como dada y debe ser corregida con la muestra. Por otro lado, el total de la población capturado por la encuesta, representado como N_{+1}, ahora se considera no observable. Además, otras cantidades clave, como N_{11} (el número de individuos capturados por ambas fuentes), N_{12} (individuos capturados por el censo, pero no por la encuesta), y N_{21} (individuos capturados por la encuesta, pero no por el censo), también se consideran desconocidas. Sin embargo, todas estas cantidades pueden estimarse indirectamente a partir de los datos de la encuesta, utilizando los métodos estadísticos adecuados.

La estructura básica del Sistema de Estimación Dual (SED) puede representarse mediante una matriz de contingencia (cuadro 2) que resume la relación entre la cobertura del censo y la de la EPC. En dicha matriz, los individuos de la población se clasifican según si fueron o no enumerados por cada una de las dos fuentes. Sin embargo, no todas estas cantidades requieren ser observadas o estimadas de manera directa. En la práctica, los únicos elementos esenciales para la estimación mediante el SED son N_{11}, N_{12} y N_{21}. Estas tres celdas contienen la información fundamental necesaria para inferir el tamaño total de la población, mientras que las demás pueden deducirse a partir de ellas o no son estrictamente necesarias para el cálculo. De esta forma, la estructura de los datos y estimaciones

El cuadro 14 presenta los estimadores básicos necesarios para la aplicación del SED. Nótese que el superíndice (ˆ) sobre cada símbolo indica que se trata de una cantidad estimada, es decir, un valor inferido a partir de los datos del censo y de la EPC, y no de un conteo directo de la población. Cada elemento de esta matriz representa una categoría específica de individuos según su inclusión o no en cada fuente. Así, $\widehat{N}_{11}$ corresponde a la cantidad estimada de personas observadas tanto en el censo como en la encuesta (coincidencias), $\widehat{N}_{12}$ es una estimación de las personas que aparecen únicamente en el censo, y $\widehat{N}_{21}$ corresponde a la estimación del número total de personas que solo fueron captadas por la encuesta. A partir de estas tres celdas se estiman los totales marginales $\widehat{N}_{1+}$ (total estimado de personas enumeradas correctamente en el censo) y $\widehat{N}_{+1}$ (total estimado de personas en la EPC), mientras que $\widehat{N}$ representa[2] la estimación global de la población N_{++}.

Como la encuesta representa una muestra de la población que viene de una medida de probabilidad, y a su vez, existe un modelo multinomial, entonces se introduce una complejidad metodológica clave: la necesidad de establecer las bases inferenciales para incluir dos fuentes de incertidumbre (Binder 2011): el modelo y el muestreo. Wolter (1986) afirma que este cambio de enfoque implica que la estimación del error de cobertura debe considerar dos fuentes principales de

[2] Para simplificar la notación, se usará el símbolo $\widehat{\mathrm{N}}$ en lugar de $\widehat{\mathrm{N}}_{++}$ para representar el estimador del total poblacional real.

incertidumbre: (1) la variabilidad debida a la selección muestral de la encuesta, y (2) la variabilidad del modelo asociada con el modelo de error de cobertura.

Cuadro 14
Matriz inicial de estimadores necesarios para el SED

	En la encuesta	Fuera de la encuesta	Total
En el censo	$\widehat{N}_{11}$	$\widehat{N}_{12}$	$\widehat{N}_{1+}$
Fuera del censo	$\widehat{N}_{21}$		
Total	$\widehat{N}_{+1}$		$\widehat{N}_{++}$

Fuente: Elaboración propia.

La variabilidad inducida por la selección de la muestra de la encuesta implica que las estimaciones derivadas de ella (como $\widehat{N}_{+1}$ o $\widehat{N}_{11}$) están afectadas por la aleatoriedad inherente a la selección de unidades en la muestra. Si la encuesta utiliza un diseño complejo (con estratificación y conglomerados), la variabilidad aumenta debido a estos efectos de diseño. Este tipo de variabilidad se mide con los métodos clásicos de inferencia estadística en encuestas de hogares. En segundo lugar, está la variabilidad derivada del modelo multinomial. En esta instancia, la novedad radica en integrar estas incertidumbres por medio de una inferencia doble, usando los resultados bien conocidos de las esperanzas y varianzas condicionales.

Si denotamos por π_k la probabilidad de inclusión del elemento k en la muestra s_P, la cual está determinada por su selección probabilística, entonces el peso de muestreo del elemento k-ésimo en la muestra P se define como $w_k = \pi_k^{-1}$. Este peso refleja la inversa de la probabilidad de inclusión y se utiliza para ajustar las estimaciones en función del diseño de muestreo. De manera similar, los pesos de muestreo se definirán para la muestra s_E. Para simplificar la notación, vincularemos la muestra correspondiente a través de los subíndices en las sumas. Por ejemplo, al referirnos a la muestra s_P, utilizaremos el subíndice P en las sumas, y para la muestra s_E, emplearemos el subíndice E.

Asumiendo que $x_{k,11}$ representa una variable aleatoria dicotómica que toma el valor de uno si el individuo k fue encontrado tanto en la muestra como en el censo y, cero, en otro caso, entonces los estimadores de muestreo de N_{+1} y N_{11}, serán respectivamente:

$$\widehat{N}_{+1} = \sum_{k \in s_P} w_k$$

$$\widehat{N}_{11} = \sum_{k \in s_P} w_k \; x_{k,11}$$

Asimismo, si z_k representa una variable aleatoria dicotómica que toma el valor de uno si el individuo k fue correctamente enumerado en el censo y, cero, en otro caso, entonces el estimador de muestreo de N_{1+} será:

$$\widehat{N}_{1+} = N_{1+}^0 - \widehat{N}_{EE} = N_{1+}^0 - \sum_{k \in s_E} w_k \, (1 - \; z_k)$$

En donde N_{1+}^0 denota el número bruto de registros censales, el cual difiere del conteo de personas en la población, y representa el conteo no corregido de personas en el censo. Esta cifra debe basarse exclusivamente en los datos recopilados durante el operativo censal, sin incluir imputaciones, proyecciones ni ningún otro tipo de ajustes estadísticos. Esto garantiza que los resultados reflejen fielmente la información obtenida en el campo. Además, $\widehat{N}_{EE} = \sum_{k \in s_E} w_k \, (1 - \; z_k)$, representa el

número estimado de enumeraciones erróneas en el conteo censal bruto. Para los anteriores estimadores, es claro que $x_{k,11}$ es una variable aletaoria que se define en la muestra s_P, mientras que z_k es una variable aleatoria que se define en la muestra s_E. Por otro lado, Bureau (2022) propone un estimador directo alternativo para N_{1+}, que se define a partir de la muestra E, y que corresponde a un conteo ponderado de enumeraciones correctas. Este estimador toma la siguiente forma:

$$\widehat{N}_{1+} = \sum_{k \in s_E} w_k \; z_k$$

Estimador de Petersen

Recordando la teoría del Capítulo I, el estimador del modelo para N_{++} es $\widetilde{N}$, entonces, su estimador insesgado bajo el diseño de muestreo se encuentra reemplazando N_{1+}, N_{+1} y N_{11} por sus respectivos estimadores insesgados en la muestra. En consecuencia, un estimador de muestreo insesgado para el tamaño poblacional N tomará la siguiente forma:

$$\widehat{N}_{Pet} = \frac{\widehat{N}_{1+} \cdot \widehat{N}_{+1}}{\widehat{N}_{11}}$$

De la misma manera, los estimadores de muestreo insesgados para N_{12} y N_{21} toman la siguiente forma, respectivamente:

$$\begin{aligned} \widehat{N}_{12} &= \widehat{N}_{1+} - \widehat{N}_{11} \\ \widehat{N}_{21} &= \widehat{N}_{+1} - \widehat{N}_{11} \end{aligned}$$

El estimador $\widehat{N}_{Pet}$ se conoce como estimador de Petersen y también como el estimador de Lincoln-Petersen. Este fue originalmente desarrollado para estudios de fauna (Petersen, 1896), pero su uso se ha extendido a otros campos. Como se vio en el capítulo 1, asume que la población es cerrada (sin nacimientos, muertes, inmigración o emigración), que todos los individuos tienen la misma probabilidad de captura, y que la marcación no afecta la probabilidad de recaptura; es decir, que las fuentes de identificación son independientes. El estimador de Petersen es el más conocido de los estimadores de tamaño poblacional en el sistema dual y puede demostrarse que es un estimador de máxima verosimilitud condicional para el modelo log-lineal de independencia con dos variables (Fienberg 1972; Bishop, Fienberg, y Holland 2007).

La existencia de individuos que no fueron capturados en ninguno de los dos listados representa un desafío significativo, ya que su número solo puede ser estimado indirectamente a partir de la superposición observada entre la encuesta y el censo. Por otro lado, Wolter (1986) establece las condiciones sobre las cuales estos estimadores son insesgados y además propone el siguiente estimador aproximadamente insesgado de su varianza:

$$\hat{V}\left(\widehat{N}_{Pet}\right) = \hat{V}_m\left(\widehat{N}_{Pet}\right) + \hat{V}_p\left(\widehat{N}_{Pet}\right)$$

En donde $\hat{V}_m\left(\widehat{N}_{Pet}\right)$ es el estimador de la varianza bajo el modelo multinomial, que usa las contrapartes muestrales en lugar de las poblacionales, de la siguiente forma:

$$\hat{V}_m\left(\widehat{N}_{Pet}\right) = \frac{\widehat{N}_{1+} \cdot \widehat{N}_{+1} \cdot \left(\widehat{N}_{1+} - \widehat{N}_{11}\right) \cdot \left(\widehat{N}_{+1} - \widehat{N}_{11}\right)}{\widehat{N}_{11}^3}$$

Asimismo, $\hat{V}_p\left(\widehat{N}_{Pet}\right)$ corresponde a un estimador tradicional de la varianza de muestreo para estimadores de encuestas complejas (CEPAL 2023). De esta forma, suponiendo que el diseño de muestreo seleccionó n_I unidades primarias de muestreo (UPM), de un total de N_I, con una fracción de muestreo $f = n_I/N_I$, Wolter (1986, sección 3.1.) afirma que

$$\hat{V}_p\left(\hat{N}_{Pet}\right) \approx \frac{{N_I}^2}{n_I}(1-f)\,S_d^2$$

Definiendo a $\hat{N}_{i,+1} = \sum_{k\in i} w_k$ como la estimación del tamaño de la i-ésima UPM (es decir; la suma de los factores de expansión en esa UPM), a partir de la muestra s_P se tiene que $S_d^2 = \frac{1}{n-1}\sum_{i=1}^{n} d_i^2$ y además $d_i = \frac{\hat{N}_{+1}}{\hat{N}_{11}}\left(\hat{N}_{i,+1} - \frac{\hat{N}_{+1}}{\hat{N}_{11}}x_{i,11}\right)$.

Ejemplo

En esta sección presentaremos un ejemplo ilustrativo que permitirá comprender, paso a paso, cómo se aplican los ajustes y estimaciones dentro del SED. Suponga que el conteo bruto censal, correspondiente al número de registros censales en la base de datos sin depurar (antes de aplicar cualquier corrección) es $N_{1+}^0 = 1\,500\,000$. A partir de la EPC, tras expandir los resultados muestrales con los factores de expansión, se estimó que existen aproximadamente $\hat{N}_{EE} = 20\,000$ enumeraciones incorrectas en el universo. Este ajuste permite depurar el total censal efectivo que participará en el emparejamiento, resultando en un total corregido de $\hat{N}_{1+} = 1\,480\,000$. De la misma manera, asuma que el total poblacional estimado desde la EPC fue $\hat{N}_{+1} = 1\,300\,000$, y que el número expandido de personas encontradas en ambas fuentes (emparejadas) fue $\hat{N}_{11} = 1\,200\,000$.

Con base en esta información, definimos los componentes observados que serán utilizados en los estimadores del SED. A partir de estas relaciones, los valores de las celdas restantes se calculan como diferencias entre totales y coincidencias:

$$\hat{N}_{12} = \hat{N}_{1+} - \hat{N}_{11} = 280\,000$$

$$\hat{N}_{21} = \hat{N}_{+1} - \hat{N}_{11} = 100\,000$$

A partir de las anteriores estimaciones, es posible estimar el tamaño total de la población usando el estimador de Petersen, definido como:

$$\hat{N}_{Pet} = \frac{\hat{N}_{1+}\cdot\hat{N}_{+1}}{\hat{N}_{11}} = \frac{1\,480\,000 \times 1\,300\,000}{1\,200\,000} \approx 1\,603\,333$$

Este resultado indica que, al combinar la información del censo y de la EPC, y considerando los emparejamientos observados, se estima que la población total es aproximadamente 1 603 333 personas. Nótese que este valor es mayor que los totales marginales individuales, lo que refleja que existen individuos no observados en ambas fuentes y que el estimador de Petersen corrige justamente por esta omisión. Como la celda correspondiente a los individuos que no fueron capturados ni por el censo ni por la EPC ($\hat{N}_{22}$), utilizaremos el estimador de Petersen para inferirla a partir de las celdas observadas ($\hat{N}_{11}, \hat{N}_{12}, \hat{N}_{21}$) y los totales marginales. Con este valor, podremos completar la tabla y obtener un panorama completo de las relaciones entre las dos fuentes de información. El cuadro 15 presenta la matriz de contingencia estimada con el método de Petersen.

Cuadro 15
Matriz de estimadores de Petersen para el SED

	En la encuesta	Fuera de la encuesta	Total
En el censo	$\hat{N}_{11} = 1\,200\,000$	$\hat{N}_{12} = 280\,000$	$\hat{N}_{1+} = 1\,480\,000$
Fuera del censo	$\hat{N}_{21} = 100\,000$	$\hat{N}_{12} = 23\,333$	$\hat{N}_{12} = 123\,333$
Total	$\hat{N}_{+1} = 1\,300\,000$	$\hat{N}_{12} = 303\,333$	$\hat{N}_{++} = 1\,603\,333$

Fuente: Elaboración propia.

Este resultado indica que, al combinar la información del censo y de la EPC, y considerando los emparejamientos observados, se estima que la población total es aproximadamente 1 603 333 personas. Nótese que este valor es mayor que los totales marginales individuales, lo que refleja que existen individuos no observados en ambas fuentes y que el estimador de Petersen corrige justamente por esta omisión. El cuadro 14 presenta la matriz de contingencia estimada con el método de Petersen que resume la relación entre la información del censo y de la EPC.

Para los valores del ejemplo la varianza basada únicamente en el modelo toma el valor numérico $\hat{V}_m(\widehat{N}_{Pet}) \approx 31\,176$, y su error estándar asociado es $\widehat{SE} = 176.57$. Por ende, un intervalo de confianza aproximado al 95%, usando la aproximación normal, estará dado por:

$$\widehat{N}_{Pet} \pm 1.96 \cdot \widehat{SE} \approx (1\,602\,987,\ 1\,603\,679)$$

Con base en los datos disponibles, se estima un error neto de cobertura (ENC) de 123 333 personas, lo que equivale a una tasa del error neto de cobertura (TENC) de 7,7 %, indicando que el censo subestimó aproximadamente ese porcentaje de la población. Al considerar además las enumeraciones erróneas, el total de omisiones (O) asciende a 143 333 personas, lo que implica una tasa de omisión (TO) del 8,9 % respecto a la población real. Por su parte, la tasa de emparejamiento (TE) alcanza el 81,1 %, reflejando un buen nivel de vinculación entre el censo y la EPC, mientras que la tasa de enumeraciones erróneas (TEE) es del 1,4 %, lo que sugiere que las inclusiones incorrectas en el censo fueron relativamente bajas.

B. Otros estimadores del SED

En el ámbito de la evaluación censal, el estimador de Petersen constituye la base conceptual del SED. El principio es análogo al del experimento de captura y recaptura: el censo representa la primera "captura" y la encuesta la segunda, siendo las coincidencias entre ambas fuentes el elemento clave para estimar la fracción de la población no observada en ninguna de las dos. A partir de este esquema general se han propuesto diversos estimadores para el tamaño total de la población, que varían según los supuestos que imponen y los ajustes que incorporan frente a sesgos potenciales, como la heterogeneidad en las probabilidades de inclusión, la dependencia entre las fuentes o el tamaño reducido de las muestras. En esta sección se presenta una revisión no exhaustiva de estos estimadores, destacando sus formulaciones, supuestos y ámbitos de aplicación, así como las ventajas y limitaciones de cada uno en el contexto de las encuestas de posenumeración y la evaluación de la cobertura censal.

Estimador de razón

Es posible combinar los diferentes estimadores de Petersen en las muestras E y P, junto con la información de los registros censales para crear otro tipo de estimadores. Siendo $\widehat{N}_{1+}^{0} = \sum_{k \in s_E} w_k$ un estimador de muestreo del número de enumeraciones en el censo (correctas o erroneas), es posible ajustar el número de enumeraciones en el censo con su contraparte muestral, y definir el siguiente estimador de razón:

$$\widehat{N}_{Raz} = \frac{N_{1+}^{0}}{\widehat{N}_{1+}^{0}} \frac{\widehat{N}_{1+} \cdot \widehat{N}_{+1}}{\widehat{N}_{11}}$$

Para aplicar el estimador de razón, es necesario introducir el total censal bruto estimado desde la EPC, el cual se denotó como $\widehat{N}_{1+}^{0}$. Este valor representa la cantidad de personas que, según la EPC y sus factores de expansión, se habría registrado en el censo sin aplicar aún las correcciones por enumeraciones incorrectas. En nuestro ejemplo, asumimos que $\widehat{N}_{1+}^{0} = 1\,490\,000$. Este total sirve como referencia para ajustar la estimación final del tamaño poblacional y permite reflejar posibles

discrepancias entre los registros censales y la proyección derivada de la EPC. Por ende, sustituyendo los valores de nuestro ejemplo, el estimador de razón da como resultado:

$$\widehat{N}_{Raz} = \frac{N^0_{1+}}{\widehat{N}^0_{1+}} \cdot \frac{\widehat{N}_{1+} \cdot \widehat{N}_{+1}}{\widehat{N}_{11}} = \frac{1\,500\,000}{1\,490\,000} \cdot \frac{1\,480\,000 \times 1\,300\,000}{1\,200\,000} = 1\,614\,000$$

Este resultado indica que, al incorporar la diferencia entre el total crudo del censo y su estimación a partir de la EPC, el tamaño total de la población se ajusta ligeramente hacia arriba, reflejando individuos posiblemente no observados en ambas fuentes.

Estimador de posestratificación

En el mismo espíritu, también se puede refinar el estimador de Petersen utilizando técnicas de posestratificación (Gutiérrez, 2016), consistentes en la partición de la población en subgrupos lo más homogéneos posible de acuerdo con variables que influyen en la probabilidad de cobertura. Este procedimiento busca reducir el sesgo de correlación, el cual se produce cuando las probabilidades de omisión en el censo y en la encuesta no son independientes entre sí. En otras palabras, el sesgo de correlación surge cuando los individuos que no fueron enumerados en el censo tienden a tener también una mayor probabilidad de no ser incluidos en la encuesta de posenumeración. Este fenómeno puede originarse por diversas razones: factores geográficos (por ejemplo, viviendas en áreas de difícil acceso), socioeconómicos (población con mayor movilidad o en condiciones precarias de residencia), o incluso comportamentales (personas renuentes a responder o difíciles de localizar). En tales casos, uno de los supuestos del SED (que las omisiones en ambas fuentes son independientes) dejará de cumplirse, provocando una subestimación sistemática del número de individuos omitidos.

La posestratificación contribuye a mitigar este problema al crear estratos en los que la independencia entre las fuentes es más plausible, agrupando a las personas según características relacionadas con la cobertura (como edad, sexo, zona de residencia o condición socioeconómica). De este modo, las estimaciones del SED se calculan separadamente dentro de cada subgrupo y luego se combinan (con ponderaciones apropiadas) para obtener una estimación global insesgada. Suponiendo que existen G postestratos (cada una de las particiones inducidas por el cruce de estas variables), entonces el estimador de razón postestratificada toma la siguiente forma:

$$\widehat{N}_{Pos} = \sum_{g=1}^{G} \left[\frac{N^0_{g1+}}{\widehat{N}^0_{g1+}} \cdot \frac{\widehat{N}_{g1+} \cdot \widehat{N}_{g+1}}{\widehat{N}_{g11}} \right] = \sum_{g=1}^{G} \left[N^0_{g1+} \cdot \frac{\hat{p}_{g1+}}{\hat{p}_{g11}} \right]$$

En donde $\hat{p}_{g1+} = \frac{\widehat{N}_{g1+}}{\widehat{N}^0_{g1+}}$ y $\hat{p}_{g11} = \frac{\widehat{N}_{g11}}{\widehat{N}_{g+1}}$ son respectivamente estimadores directos de la proporción de individuos correctamente enumerados y de la proporción de emparejamiento en el postestrato g. Esta última expresión resultará muy valiosa para desarrollar modelos de estimación en áreas pequeñas, permitiendo calcular con mayor precisión la omisión censal.

Siguiendo con el ejemplo, suponga que nuestra población se divide en dos postestratos ($G = 2$) por sexo (hombres y mujeres) y que los totales crudos censales son $N^0_{m,1+} = 800\,000$, para las mujeres, mientras que, para los hombres, $N^0_{h,1+} = 700\,000$. Además, los totales censales corregidos por postestrato a partir de la EPC son $\widehat{N}^0_{m,1+} = 790\,000$ y $\widehat{N}^0_{h,1+} = 690\,000$. Por otro lado, los totales marginales estimados en la EPC son $\widehat{N}_{m,+1} = 700\,000$ y $\widehat{N}_{h,+1} = 600\,000$, mientras que, los emparejamientos expandidos en cada postestrato son $\widehat{N}_{m,11} = 650\,000$ y $\widehat{N}_{h,11} = 580\,000$.

Con esta información, para el postestrato de las mujeres se define la siguiente cantidad:

$$\frac{N^0_{m,1+}}{\widehat{N}^0_{m,1+}} \cdot \frac{\widehat{N}_{m,1+} \cdot \widehat{N}_{m,+1}}{\widehat{N}_{m,11}} = \frac{800\,000}{790\,000} \cdot \frac{790\,000 \cdot 700\,000}{650\,000} = 861\,538$$

Mientras que, para el postestrato de los hombres, se tiene que:

$$\frac{N^0_{h,1+}}{\widehat{N}^0_{h,1+}} \cdot \frac{\widehat{N}_{h,1+} \cdot \widehat{N}_{h,+1}}{\widehat{N}_{h,11}} = \frac{700\,000}{690\,000} \cdot \frac{690\,000 \cdot 600\,000}{580\,000} = 724\,138$$

Al sumar ambos postestratos, la estimación resultante es:

$$\widehat{N}_{Pos} = \sum_{g=m,h} \frac{N^0_{g,1+}}{\widehat{N}^0_{g,1+}} \cdot \frac{\widehat{N}_{g,1+} \cdot \widehat{N}_{g,+1}}{\widehat{N}_{g,11}} = 861\,538 + 724\,138 = 1\,585\,676$$

Estimador de Chapman

El Estimador de Chapman surge como una corrección al estimador de Petersen y es especialmente útil cuando el número de recapturas N_{11} es pequeño (ya que el estimador de Petersen tiende a ser sesgado en esos casos), lo cual es frecuente en estudios con poblaciones grandes o tasas de captura bajas. Chapman (1951) propuso una alternativa manteniendo los supuestos de población cerrada, independencia y probabilidades homogéneas de captura. En este caso se sugiere estimar el tamaño de la población usando la siguiente corrección:

$$\widehat{N}_{Cha} = \frac{\left(\widehat{N}_{1+} + 1\right) \cdot \left(\widehat{N}_{+1} + 1\right)}{\widehat{N}_{11} + 1} - 1$$

Este estimador, basado en la distribución hipergeométrica, garantiza momentos finitos debido a que el denominador no puede ser cero. La expresión para la estimación de la varianza se puede obtener usando expansión de Taylor bajo un modelo hipergeométrico (Seber 1982), donde se obtiene:

$$\hat{V}\left(\widehat{N}_{Cha}\right) = \frac{\left(\widehat{N}_{1+} + 1\right)\left(\widehat{N}_{+1} + 1\right)\left(\widehat{N}_{1+} - \widehat{N}_{11}\right)\left(\widehat{N}_{+1} - \widehat{N}_{11}\right)}{\left(\widehat{N}_{11} + 1\right)^2\left(\widehat{N}_{11} + 2\right)}$$

Posteriormente, Sadinle (2009) propuso un método para calcular intervalos de confianza robustos para las estimaciones que provienen del estimador de Chapman. Usando los datos del ejemplo anterior, el estimador de Chapman toma la siguiente forma:

$$\widehat{N}_{Cha} = \frac{\left(\widehat{N}_{1+} + 1\right) \cdot \left(\widehat{N}_{+1} + 1\right)}{\widehat{N}_{11} + 1} - 1 = \frac{(1\,480\,000 + 1) \cdot (1\,300\,000 + 1)}{1\,200\,000 + 1} - 1 = 1\,603\,332$$

Este resultado refleja una estimación del tamaño total de la población que incorpora, de manera implícita, a los individuos no capturados por ninguna de las dos fuentes, representando así una corrección más precisa frente a las limitaciones del estimador de Petersen.

Estimador de Nour

En algunos casos, el fracaso de capturar a un individuo en ambas listas puede deberse a causas comunes, lo que conduce a una asociación positiva entre las dos fuentes. Esto es habitual en las encuestas de cobertura, donde los individuos pueden estar menos dispuestos a ser registrados generando altas tasas de rechazo, lo que resulta en una asociación negativa entre las listas. Estos fenómenos se conocen como variación de respuesta conductual (Wolter 1986). El estimador de la cota inferior del tamaño poblacional propuesto por Nour (1982), se basa en una estimación para N_{22} (correspondiente a individuos no capturados por ninguna de las dos listas), bajo los siguientes tres supuestos:

i) Existe correlación positiva entre las listas, es decir: $N_{11} \cdot N_{22} > N_{12} \cdot N_{21}$.

ii) La probabilidad de que una unidad sea registrada en alguna de las dos fuentes es mayor que $\frac{1}{2}$; es decir: $\frac{N_{1+}}{N} > \frac{1}{2}\ y\ \frac{N_{+1}}{N} > \frac{1}{2}$. Este supuesto asegura que $N_{11} > N_{22}$. Conjuntamente, se garantiza que $N_{11}^2 > N_{11} \cdot N_{22} > N_{12} \cdot N_{21}$.

iii) Se cumple que $N_{12} \cdot N_{21} - N_{22}^2 > 0$, y por tanto $N_{12} \cdot N_{21} < \left(\frac{N_{12} \cdot N_{21}}{N_{22}}\right)^2$.

Bajo los anteriores supuestos, Nour (1982) mostró que la parte no observada de la población N_{22} se encuentra en el intervalo:

$$\left[\frac{2 \cdot N_{12} \cdot N_{21} \cdot N_{11}}{N_{12} \cdot N_{21} + N_{11}^2},\ \sqrt{N_{12} \cdot N_{21}}\right]$$

Basado en lo anterior, propuso el siguiente estimador puntual, con la justificación de que es más robusto frente a la posible violación del tercer supuesto (el cual en la práctica es difícil de verificar):

$$\widehat{N}_{22} = \frac{2 \cdot \widehat{N}_{12} \cdot \widehat{N}_{21} \cdot \widehat{N}_{11}}{\widehat{N}_{12} \cdot \widehat{N}_{21} + \widehat{N}_{11}^2}$$

En caso de que este supuesto no se vea violado, también se puede considerar otros estimadores para N_{22}, tales como:

$$\widehat{N}_{22} = \sqrt{\widehat{N}_{12} \cdot \widehat{N}_{21}}$$

o incluso cualquier punto dentro del intervalo:

$$\left[\frac{2 \cdot \widehat{N}_{12} \cdot \widehat{N}_{21} \cdot \widehat{N}_{11}}{\widehat{N}_{12} \cdot \widehat{N}_{21} + \widehat{N}_{11}^2},\ \sqrt{\widehat{N}_{12} \cdot \widehat{N}_{21}}\right]$$

Dependiendo de la selección del estimador para N_{22}, se obtienen dos estimadores del tamaño total de la población bajo dependencia positiva, denotados como: cota inferior $\widehat{N}_L$ y cota superior $\widehat{N}_U$, las cvuales están dadas respectivamente por:

$$\widehat{N}_L = \widehat{N}_{11} + \widehat{N}_{12} + \widehat{N}_{21} + \frac{2 \cdot \widehat{N}_{12} \cdot \widehat{N}_{21} \cdot \widehat{N}_{11}}{\widehat{N}_{12} \cdot \widehat{N}_{21} + \widehat{N}_{11}^2}$$

$$\widehat{N}_U = \widehat{N}_{11} + \widehat{N}_{12} + \widehat{N}_{21} + \sqrt{\widehat{N}_{12} \cdot \widehat{N}_{21}}$$

Estos estimadores no cuentan con una expresión analítica simple que permita calcular directamente su varianza. Para superar esta limitación, es necesario recurrir a métodos basados en réplicas (CEPAL, 2023), que permiten aproximar la variabilidad del estimador mediante la generación de múltiples versiones de este a partir de la muestra original. Usando los datos del ejemplo, podemos estimar que el número estimado de individuos no observados en ninguna de las dos fuentes se encuentra contenido en el siguiente intervalo:

$$\left[\frac{2 \cdot \widehat{N}_{12} \cdot \widehat{N}_{21} \cdot \widehat{N}_{11}}{\widehat{N}_{12} \cdot \widehat{N}_{21} + \widehat{N}_{11}^2},\ \sqrt{\widehat{N}_{12} \cdot \widehat{N}_{21}}\right] \approx [45\,800, 167\,332]$$

Por ende, se estima que el total de la población se encuentra contenido en el siguiente intervalo:

$$\left[\widehat{N}_L,\ \widehat{N}_U\right] = [1\,580\,000 + 45\,800,\ 1\,580\,000 + 167\,332] = [1\,625\,800, 1\,747\,332]$$

Estimador de Chao

Chao (1987, 1989) propuso un estimador del tamaño poblacional que relaja el supuesto de independencia entre las fuentes —censo y encuesta de posenumeración— y permite considerar heterogeneidad no observada en las probabilidades de captura de los individuos. Este estimador fue diseñado para corregir la subestimación del tamaño poblacional que ocurre cuando las probabilidades de inclusión varían entre los individuos o cuando el tamaño de la muestra es reducido. La idea central es que esta variabilidad genera que ciertos individuos tengan muy bajas probabilidades de ser observados en ambas fuentes, lo que provoca que los estimadores lineales tradicionales subestimen sistemáticamente la población total.

Este estimador se enfoca en el número mínimo de individuos no observados N_{22} que puede explicarse a partir de los datos observados y se fundamenta en la idea de que los individuos capturados solo una vez contienen información sobre la cantidad de personas no observadas en ninguna fuente. En su formulación general, el estimador se expresa como:

$$N = N_2 + N_1 + N_0$$

En donde N_2 corresponde al total de individuos capturados por dos veces (por el censo y por la EPC), N_1 corresponde al total de individuos capturados exactamente una vez (por el censo o por la EPC) y N_0 el número de individuos que no fueron capturados nunca (ni por el censo ni por la EPC). Esta expresión proporciona un ajuste que corrige la subestimación que surge cuando algunos individuos tienen probabilidades muy bajas de ser capturados en ambas fuentes. En el contexto del SED, los términos se pueden identificar directamente a partir de la matriz de emparejamiento. La racionalidad del estimador descansa en un modelo binomial mixto para dos fuentes de captura, donde cada individuo tiene una probabilidad de ser capturado p distribuida como $f(p)$. En este contexto, el número esperado de individuos presentes en *j* fuentes se expresa como:

$$E(N_j) = N \cdot \int_0^1 \binom{2}{j} p^j \cdot (1-p)^{2-j} \cdot f(p) \cdot dp, \quad j = 0,1,2,$$

En donde N_j es el número de individuos presentes en *j* fuentes y $f(p)$ representa la distribución de las probabilidades de captura entre los individuos. Es decir, donde N_0 representa los individuos no observados en ninguna fuente; $N_1 = N_{12} + N_{21}$, los individuos capturados solo en una fuente y $N_2 = N_{11}$, los individuos capturados en ambas fuentes. Para derivar un límite inferior para N_0, es posible acudir a la desigualdad de Cauchy-Schwarz

$$[E(XY)]^2 \leq E(X^2) \cdot E(Y^2),$$

Sustituyendo $X = p$y $Y = 1 - p$, se tiene que:

$$\left(\int_0^1 p \cdot (1-p) \cdot f(p) \cdot dp\right)^2 \leq \int_0^1 (1-p)^2 \cdot f(p) \cdot dp \int_0^1 p^2 \cdot f(p) \cdot dp,$$

que puede reescribirse como:

$$\left(\frac{1}{2} \cdot E(N_1)\right)^2 \leq E(N_0) \cdot E(N_2),$$

de donde se deduce que:

$$E(N_0) \geq \frac{[E(N_1)]^2}{4E(N_2)}.$$

Sustituyendo las frecuencias esperadas por las observadas en la matriz de coincidencias, se obtiene el límite inferior empírico para estimar N_0:

$$\widehat{N}_0 = \frac{N_1^2}{4N_2} = \frac{(N_{21} + N_{12})^2}{4N_{11}}.$$

Finalmente, el estimador de Chao para el tamaño poblacional total se construye sumando los individuos observados al menos una vez y el límite inferior de los no observados:

$$\begin{aligned}\widehat{N}_{Chao} &= \widehat{N}_{11} + \widehat{N}_{12} + \widehat{N}_{21} + \frac{\widehat{N}_1^2}{4\widehat{N}_2} \\ &= \widehat{N}_{11} + \widehat{N}_{12} + \widehat{N}_{21} + \frac{\left(\widehat{N}_{21} + \widehat{N}_{12}\right)^2}{4\widehat{N}_{11}}\end{aligned}$$

La aproximación de la varianza de estimador es:

$$\hat{V}\left(\widehat{N}\right) \approx \frac{\left(\widehat{N}_{21} + \widehat{N}_{12}\right)^2}{4\widehat{N}_{11}^3}\left(\frac{\left(\widehat{N}_{21} + \widehat{N}_{12}\right)^2}{4\widehat{N}_{11}} + \widehat{N}_{21} + \widehat{N}_{12} + \widehat{N}_{11}\right)$$

Con los datos del ejemplo, $\widehat{N}_{11} = 1\,200\,000$, $\widehat{N}_{12} = 280\,000$, $\widehat{N}_{21} = 100\,000$, el estimador de Chao se calcula así:

$$\widehat{N}_{Chao} = \widehat{N}_{11} + \widehat{N}_{12} + \widehat{N}_{21} + \frac{\left(\widehat{N}_{21} + \widehat{N}_{12}\right)^2}{4 \cdot \widehat{N}_{11}} = 1\,580\,000 + \frac{(280\,000 + 100\,000)^2}{4 \cdot 1\,200\,000} = 1\,610\,083$$

Estimador de Webster-Kemp

El enfoque bayesiano propuesto por Webster y Kemp (2013) surge como una alternativa robusta a los estimadores clásicos como Petersen o Chapman, especialmente en escenarios donde el número de coincidencias entre las fuentes es pequeño o existe dependencia entre ellas. Mientras los métodos tradicionales se basan en aproximaciones frecuentistas que asumen independencia y homogeneidad en las probabilidades de captura, el método bayesiano permite incorporar de manera natural la incertidumbre y las probabilidades de ocurrencia de cada patrón de observación mediante una distribución posterior.

La principal ventaja de este enfoque es que ofrece una estimación coherente del número de individuos no observados, incluso cuando los datos son escasos o presentan sesgos de heterogeneidad. Además, la formulación bayesiana permite introducir distribuciones previas no informativas, de manera que la estimación se base principalmente en la evidencia observada. Otro aspecto clave es que este enfoque proporciona, de manera natural, una estimación de la varianza para el total de individuos no observados, lo que facilita la construcción de intervalos de confianza y la evaluación de la incertidumbre asociada a la estimación total de la población. En escenarios con muestreo complejo, donde los conteos observados son en realidad estimaciones ponderadas por los factores de expansión, el enfoque bayesiano se combina de manera eficiente con métodos basados en réplicas, permitiendo incorporar tanto la variabilidad debida al diseño muestral como la incertidumbre inherente al número de individuos no observados.

La idea central es que la probabilidad de observar un número específico de coincidencias depende del tamaño poblacional N, lo que permite aplicar la regla de Bayes para estimar su distribución posterior:

$$Pr(N|N_{+1}, N_{1+}, N_{11}) = \frac{Pr(N_{1+}, N_{+1}, N_{11}|N) \cdot Pr(N)}{Pr(N_{1+}, N_{+1}, N_{11})}$$

Bajo este enfoque, suponiendo que $\widehat{N}_{11} > 2$, el estimador bayesiano del número de individuos no observados (N_{22}) se define como:

$$\widehat{N}_{22} = \frac{\left(\widehat{N}_{12}+1\right)\cdot\left(\widehat{N}_{21}+1\right)}{\widehat{N}_{11}+2}$$

En consecuencia, el estimador del tamaño total de la población se escribe como:

$$\widehat{N}_{WK} = \widehat{N}_{11} + \widehat{N}_{12} + \widehat{N}_{21} + \frac{\left(\widehat{N}_{12}+1\right)\cdot\left(\widehat{N}_{21}+1\right)}{\widehat{N}_{11}+2}$$

Además, si $\widehat{N}_{11} > 3$, el estimador de la varianza para los no observados ($\widehat{N}_{22}$) puede expresarse, bajo la misma aproximación bayesiana, como:

$$\widehat{V}\left(\widehat{N}_{22}\right) = \frac{\left(\widehat{N}_{12}+1\right)\cdot\left(\widehat{N}_{21}+1\right)\cdot\left(\widehat{N}_{+1}-1\right)\cdot\left(\widehat{N}_{1+}-1\right)}{\left(\widehat{N}_{11}-2\right)^2\cdot\left(\widehat{N}_{11}-3\right)}$$

En algunos casos se asume que los valores de N_{11}, N_{12} y N_{21} son observados de manera exacta, por lo que $\widehat{V}\left(\widehat{N}_{WK}\right) = \widehat{V}\left(\widehat{N}_{22}\right)$. Sin embargo, en un muestreo complejo como el de las EPC, los $\widehat{N}_{ij}$ son estimaciones obtenidas con estimadores como el de Horvitz-Thompson o de calibración, y por tanto se debe incorporar la incertidumbre proveniente del diseño, por lo que es necesario incorporar la incertidumbre proveniente del diseño muestral. Para ello, se recomienda el uso de métodos basados en réplicas, que permiten estimar la varianza de manera consistente y reflejar correctamente la incertidumbre asociada al diseño.

Usando los datos del ejemplo, $\widehat{N}_{11} = 1\,200\,000$, $\widehat{N}_{12} = 280\,000$, $\widehat{N}_{21} = 100\,000$, la metodología de Webster-Kemp asume que:

$$\widehat{N}_{22} = \frac{\left(\widehat{N}_{12}+1\right)\left(\widehat{N}_{21}+1\right)}{\widehat{N}_{11}+2} = \frac{(280\,000+1)(100\,000+1)}{1\,200\,000+2} = 23\,333$$

Y, por tanto, el tamaño poblacional se estima de la siguiente manera:

$$\widehat{N}_{WK} = \widehat{N}_{11} + \widehat{N}_{12} + \widehat{N}_{21} + \widehat{N}_{22} = 1\,603\,333$$

Estimador de Zelterman

Zelterman (1988) propuso un enfoque innovador para estimar el tamaño poblacional en estudios de captura y recaptura, basado en el estimador de Horvitz–Thompson y en la consideración explícita de la heterogeneidad de las probabilidades de captura entre individuos. Este enfoque se fundamenta en que los conteos observados pueden aproximarse mediante una distribución de Poisson truncada en cero, ya que solo se incluyen en los datos los individuos capturados al menos una vez, mientras que los que no aparecen en ninguna fuente permanecen invisibles. El objetivo principal del estimador es calcular la probabilidad de no observación, denotada como p_0, utilizando los individuos observados una sola vez ($N_1 = N_{12} + N_{21}$) y los capturados ambas veces ($N_2 = N_{11}$), a partir de la distribución Poisson.

En la práctica, los individuos que no aparecen en ninguna fuente permanecen ocultos y, por tanto, los datos disponibles corresponden a la distribución condicionada a que el conteo sea mayor que cero. De ese modo, la información que se tiene procede exclusivamente de quienes fueron capturados una vez y de quienes fueron capturados dos veces. Para estimar el promedio de captura en la distribución Poisson (λ) a partir de esos datos observados, λ se estima con base en la información empírica, en términos simples, se aproxima a un múltiplo de la razón entre el número de individuos capturados dos veces y el número de individuos capturados una vez. Esta estimación ofrece una forma directa y parsimoniosa es el paso clave que permite, a continuación, estimar la probabilidad de no ser observado y corregir la población observada para estimar el total. Por ende, la aproximación de p_0 se expresa como:

$$p_0 \approx exp(-\lambda) = exp\left(-\frac{2N_{11}}{N_{12}+N_{21}}\right)$$

Este término representa la probabilidad de que un individuo no haya sido capturado en ninguna de las dos fuentes. Por lo tanto, $1 - p_0$ será la probabilidad de haber sido observado al menos una vez. Ubicar este término en el denominador permite expandir la población observada y corregir la subestimación debida a los individuos no observados. De manera intuitiva, si solo observamos una fracción de la población, dividir por esta fracción nos da una estimación del tamaño total. Por lo tanto, es estimador queda escrito como:

$$\widehat{N}_{Zel} = \frac{\widehat{N}_{11} + \widehat{N}_{12} + \widehat{N}_{21}}{1 - \hat{p}_0} = \frac{\widehat{N}_{11} + \widehat{N}_{12} + \widehat{N}_{21}}{1 - exp\left(-\frac{2\widehat{N}_{11}}{\widehat{N}_{12} + \widehat{N}_{21}}\right)}$$

La varianza aproximada del estimador se puede derivar a partir del modelo de Poisson truncado, al asumir que $\hat{\lambda} = 2\widehat{N}_{11}/(\widehat{N}_{12} + \widehat{N}_{21})$. Por tanto:

$$\hat{V}(\widehat{N}_{Zel}) = \left(\frac{\left(\widehat{N}_{11} + \widehat{N}_{12} + \widehat{N}_{21}\right) \cdot exp\left(-\hat{\lambda}\right)}{1 - ex\,p\left(-\hat{\lambda}\right)}\right)^2 \cdot \left(\frac{4\widehat{N}_{11}^2}{(\widehat{N}_{12} + \widehat{N}_{21})^4} + \frac{2}{(\widehat{N}_{12} + \widehat{N}_{21})^2}\right)$$

Brittain y Böhning (2009) presentaron una adaptación del estimador de Zelterman, aplicando la misma lógica sobre la verosimilitud binomial mixta de parámetro 2, truncada en cero. En este contexto, se consideran explícitamente los individuos que no aparecen en ninguna fuente, y el estimador se puede expresar como:

$$\widehat{N}_{BB} = \frac{\widehat{N}_{11} + \widehat{N}_{12} + \widehat{N}_{21}}{1 - \left(\frac{\widehat{N}_{12} + \widehat{N}_{21}}{\widehat{N}_{12} + \widehat{N}_{21} + 2\widehat{N}_{11}}\right)^2}$$

Usando los datos del ejemplo, $\widehat{N}_{11} = 1\,200\,000$, $\widehat{N}_{12} = 280\,000$, $\widehat{N}_{21} = 100\,000$, el estimador de Zelterman se calcula como:

$$\widehat{N}_{Zel} = \frac{1\,580\,000}{1 - exp\left(-\frac{2 \cdot 1\,200\,000}{280\,000 + 100\,000}\right)} = \frac{1\,580\,000}{1 - exp\,(-6\,316)} \approx 1\,580\,000$$

En este caso, debido a que la proporción de emparejamiento es muy alta en comparación con los individuos que aparecen solo en una fuente, el ajuste por la probabilidad de no ser capturado (p_0) resulta prácticamente nulo. Esto indica que, bajo esta situación, la heterogeneidad en las probabilidades de captura tiene un efecto mínimo. Por otro lado, el estimador de Brittain-Böhning se calcula como:

$$\widehat{N}_{BB} = \frac{1\,580\,000}{1 - \left(\frac{380\,000}{2\,580\,000}\right)^2} = 1\,615\,000$$

El resultado es ligeramente mayor que el estimador de Zelterman, reflejando que, aunque la cobertura es alta, es probable que existan algunas unidades no capturadas. Este incremento moderado da cuenta de la incertidumbre asociada a los elementos ausentes y proporciona una estimación más conservadora del tamaño total de la población.

C. Estimadores basados en modelos log-lineales

A diferencia de los anteriores estimadores, los modelos log-lineales proporcionan un enfoque alternativo y generalizado para estimar la población total en este tipo de estudios de captura y recaptura, introduciendo un enfoque probabilístico más flexible. Mientras los estimadores anteriores se basan en relaciones algebraicas entre los conteos observados en el censo y en la encuesta, los modelos

loglineales tratan los conteos N_{ij} como variables aleatorias que siguen una distribución Poisson, con medias θ_{ij} determinadas por una estructura logarítmica. Este enfoque permite modelar explícitamente los efectos asociados a cada fuente de captura, separando el efecto de estar en ambas, sin requerir supuestos estrictos de independencia o proporciones fijas de captura (Fienberg 1972; Cormack 1989). Para el caso del SED, el modelo log-lineal, puede representarse como un modelo lineal generalizado de Poisson, aplicado sobre los tres conteos observados, N_{11}, N_{12}, y N_{21}, así para $i, j \in (1, 2)$, se tiene que:

$$N_{ij} \sim Poisson\left(\theta_{ij}\right)$$

$$log\left(\theta_{ij}\right) = \lambda + \lambda_1^{(i)} + \lambda_2^{(j)}$$

En este modelo, λ representa el intercepto general, mientras que los términos $\lambda_1^{(i)}$ y $\lambda_2^{(j)}$ capturan los efectos de estar o no en cada lista: $\lambda_1^{(1)}$ y $\lambda_1^{(2)}$ se asocian a la inclusión o exclusión en el censo, respectivamente; mientras que $\lambda_2^{(1)}$ y $\lambda_2^{(2)}$ a la inclusión o exclusión en la encuesta de cobertura, respectivamente. Este modelo no es identificable (hay más parámetros que información independiente en los datos) y se debe imponer las siguientes restricciones:

$$\sum\nolimits_i \lambda_1^{(i)} = 0, \qquad \sum\nolimits_j \lambda_2^{(j)} = 0$$

Por lo anterior, los efectos se interpretan de forma relativa; es decir, un valor positivo de $\lambda_1^{(1)}$ o $\lambda_2^{(1)}$ indica una mayor probabilidad de captura en la fuente correspondiente. A partir de los parámetros estimados por máxima verosimilitud, el número esperado de individuos no observados en ninguna de las dos fuentes se obtiene como

$$\widehat{N}_{22} = exp\left(\hat{\lambda} + \hat{\lambda}_1^{(2)} + \hat{\lambda}_2^{(2)}\right)$$

En donde $\hat{\lambda}$ es el estimador de máxima verosimilitud del parámetro λ. En consecuencia, el tamaño total de la población se estima mediante

$$\widehat{N}_{log} = \widehat{N}_{11} + \widehat{N}_{12} + \widehat{N}_{21} + \widehat{N}_{22}$$

Este modelo básico supone independencia entre el censo y la EPC; es decir, que la probabilidad de que una persona sea incluida en una de las fuentes no depende de si fue incluida en la otra. Sin embargo, en la práctica esta suposición puede no cumplirse, ya que en ciertos contextos puede existir algún grado de dependencia entre ambas fuentes (por ejemplo, si las características de las personas o de las viviendas influyen de manera similar en la probabilidad de ser registradas). Para capturar esta posible dependencia, el modelo loglineal se amplía incorporando un término de interacción:

$$log\left(\theta_{ij}\right) = \hat{\lambda} + \hat{\lambda}_1^{(i)} + \hat{\lambda}_2^{(j)} + \hat{\lambda}_{12}^{(ij)}$$

En esta formulación, los términos de interacción $\lambda_{12}^{(ij)}$ permiten modelar la dependencia entre las fuentes. Existen cuatro posibles términos de interacción correspondientes a cada celda de la tabla de cruce. Por ejemplo, $\lambda_{12}^{(11)}$ refleja la interacción para los individuos que aparecen en ambas fuentes, indicando si hay más o menos coincidencias de lo esperado bajo independencia; $\lambda_{12}^{(12)}$ y $\lambda_{12}^{(21)}$ ajustan la dependencia para los individuos que aparecen solo en una de las dos fuentes. Este término de interacción permite que el modelo sea más flexible y se ajuste mejor a situaciones reales donde los supuestos clásicos no se cumplen estrictamente. Por último, además de las restricciones del modelo básico, se debe imponer otras restricciones de suma cero para la interacción. Por ende, para cada j y para cada i, se debe garantizar que:

$$\sum\nolimits_i \lambda_{12}^{(ij)} = 0, \qquad \sum\nolimits_j \lambda_{12}^{(ij)} = 0$$

Esto asegura que la interacción represente únicamente la desviación de la independencia y no se confunda con los efectos principales. En este caso, el número esperado de individuos no observados en ninguna de las dos fuentes se obtiene como

$$\widehat{N}_{22} = exp\left(\hat{\lambda} + \hat{\lambda}_1^{(2)} + \hat{\lambda}_2^{(2)} + \hat{\lambda}_{12}^{(22)}\right).$$

Con los datos del ejemplo de este capítulo, suponga que, después de ajustar un modelo loglineal con interacción para estimar el tamaño total de la población a partir de los conteos del censo y la EPC, se encontró que $\hat{\lambda} = 12.0$, $\hat{\lambda}_1^{(2)} = -0.05$, $\hat{\lambda}_2^{(2)} = -0.06$ y $\hat{\lambda}_{12}^{(22)} = -0.32$. A partir de estos parámetros, podemos calcular el valor esperado de la celda que representa a los individuos no observados en ninguna fuente. Por tanto:

$$\widehat{N}_{22} = exp(12.0 - 0.05 - 0.02 - 0.32) = exp(11.61) = 110\,000$$

Recordando que el total poblacional estimado bajo el modelo loglineal se obtiene como la suma de todas las celdas de la matriz de contingencia, resulta entonces posible reconstruir todas las entradas de la matriz de contingencia a partir de los datos observados y de la estimación de $\widehat{N}_{22}$. El Cuadro 16 presenta la matriz estimada, a partir de un modelo loglineal con interacción.

Cuadro 16
Matriz de estimadores del modelo loglineal con interacción para el SED

	En la encuesta	Fuera de la encuesta	Total
En el censo	$\widehat{N}_{11} = 1\,200\,000$	$\widehat{N}_{12} = 280\,000$	$\widehat{N}_{1+} = 1\,480\,000$
Fuera del censo	$\widehat{N}_{21} = 100\,000$	$\widehat{N}_{12} = 110\,000$	$\widehat{N}_{12} = 210\,000$
Total	$\widehat{N}_{+1} = 1\,300\,000$	$\widehat{N}_{12} = 390\,000$	$\widehat{N}_{++} = 1\,690\,000$

Fuente: Elaboración propia.

Con base en los valores estimados por esta metodología, se obtienen los principales indicadores de cobertura censal. El error neto de cobertura (ENC), que mide la diferencia entre la población real y la censada, es de 210 000 personas, lo que evidencia una subenumeración del censo respecto de la población total. En términos relativos, la tasa del error neto de cobertura (TENC) asciende a 12,4%, indicando que cerca de una de cada ocho personas no fue correctamente registrada. El total de omisiones (O), que incluye tanto a las personas no contadas como a las enumeradas erróneamente, alcanza 230 000 personas, lo que representa una tasa de omisión (TO) de 13,6% sobre la población total. Por su parte, la tasa de emparejamiento (TE) resulta en 81,1%, reflejando que una proporción elevada de los registros censales se pudo vincular exitosamente con la EPC. Finalmente, la tasa de enumeraciones erróneas (TEE) es de 1,4%, lo que muestra que una fracción relativamente pequeña de los registros censales corresponde a personas que no pertenecen efectivamente a la población objetivo.

D. Comparación entre los estimadores

La elección del estimador más adecuado para calcular el tamaño poblacional en un SED requiere un análisis cuidadoso de los supuestos subyacentes a cada modelo y de las condiciones específicas de los datos. No todos los estimadores son igualmente robustos frente a las complejidades que pueden presentarse en la práctica, por lo que comprender estas limitaciones es fundamental para garantizar estimaciones confiables y válidas. La decisión dependerá tanto de las características del censo y de la EPC, como de los objetivos del estudio y del tipo de población que se desea estimar. Uno de los factores críticos a considerar es la dependencia entre fuentes. Este fenómeno ocurre cuando la probabilidad de

que un individuo sea registrado en el censo está relacionada con la probabilidad de aparecer en la EPC. Los modelos clásicos, como el de Petersen, asumen independencia entre ambos levantamientos; por lo tanto, la presencia de dependencia puede generar sesgos importantes si no se utiliza un estimador que ajuste explícitamente esta relación. Reconocer y corregir esta dependencia es esencial para evitar subestimaciones o sobreestimaciones del tamaño poblacional.

Otro elemento central es la heterogeneidad en las probabilidades de captura. En muchas aplicaciones, las personas no tienen la misma probabilidad de ser registradas en el censo o en la EPC debido a características individuales no observadas, como edad, ubicación geográfica, nivel de actividad o movilidad. Ignorar esta variabilidad puede conducir a estimaciones sesgadas, especialmente cuando existen grupos con probabilidades extremas de inclusión. Algunos estimadores modernos permiten modelar estas diferencias, mejorando la precisión en presencia de heterogeneidad. Finalmente, el diseño de muestreo complejo utilizado en la EPC también influye en la elección del estimador. La presencia de estratificación, conglomerados y factores de expansión requiere que la metodología de estimación incorpore estas características para reflejar correctamente la estructura de la población. El cuadro 17 resume los estimadores más utilizados, indicando cuáles consideran o ignoran cada tipo de complejidad y ofreciendo recomendaciones sobre su aplicabilidad según el contexto del estudio.

Cuadro 17
Resumen de las ventajas de algunos estimadores del SED

Estimador	Dependencia entre censo y EPC	Heterogeneidad de captura	Diseño complejo	Aplicación recomendada
Petersen	No	No	Sí (con ajustes)	Situaciones simples, censos y EPC relativamente completos e independientes
Razón	No	No	Sí	Estimaciones rápidas cuando las proporciones entre censo y EPC son estables
Posestratificación	No	No	Sí	Censos grandes con dominios definidos y diseño complejo
Chapman	No	No	Sí (con ajustes)	Censos y EPC pequeños a medianos, sin dependencia ni heterogeneidad fuerte
Nour	Sí	Sí	Sí (con ajustes)	Poblaciones con dependencia y heterogeneidad no modeladas por estimadores clásicos
Chao	Sí	Sí	Sí (con ajustes)	Poblaciones con heterogeneidad marcada y posible dependencia leve
Webster-Kemp	Sí	Sí	Sí (con ajustes)	Estimaciones robustas frente a dependencia moderada entre censo y EPC
Zelterman	Sí	Sí	Sí (con ajustes)	Situaciones con heterogeneidad extrema y pocas capturas dobles
Loglineales sin interacción	Sí	Sí	Sí	Cuando se dispone de tablas de contingencia y se requiere modelar dependencia entre fuentes sin interacciones complejas
Loglineales con interacción	Sí	Sí	Sí	Estimaciones precisas en censos con EPC y diseño complejo, considerando estratificación y posibles interacciones entre factores

Fuente: Elaboración propia.

VI. Modelamiento estadístico

La estimación de la omisión censal puede abordarse mediante el uso de modelos estadísticos que permitan explicar la probabilidad de que una persona haya sido omitida en el censo, a partir de características observables tanto del individuo como del entorno en que reside. Entre las alternativas disponibles, los modelos de regresión logística constituyen una herramienta especialmente adecuada, ya que permiten modelar una variable dicotómica (omisión o inclusión en el censo) en función de un conjunto de covariables explicativas. Estas covariables suelen incluir variables demográficas, socioeconómicas, y contextuales derivadas de la encuesta de cobertura o de fuentes administrativas, lo que posibilita identificar patrones sistemáticos de omisión asociados a determinadas condiciones o grupos poblacionales.

El uso de modelos estadísticos en la estimación de la omisión censal no solo permite obtener estimaciones ajustadas de la probabilidad de omisión a nivel individual o agregado, sino que además facilita la incorporación de efectos de área o grupo que reflejan las variaciones territoriales (o demográficas) en la cobertura censal. A partir de estos modelos, es posible generar estimaciones de la omisión para distintos dominios de estudio (como regiones, municipios o grupos demográficos) e incluso producir predicciones para áreas con información limitada. De esta manera, el modelamiento estadístico se convierte en un componente esencial del proceso de evaluación censal, complementando las estimaciones directas y permitiendo una comprensión más profunda de los factores que inciden en la cobertura del censo, permitiendo caracterizar estos fenómenos e identificar las subpoblaciones que tienen una mayor probabilidad de ser omitidas en el censo

En este capítulo se abordarán los elementos básicos para ajustar modelos estadísticos que expliquen la probabilidad de cada individuo de estar correctamente emparejado y de representar una enumeración correcta. Además, a partir de estos modelos es posible crear ponderadores que sean aplicados a la base de datos del censo para obtener estimaciones del tamaño de las poblaciones basadas en el SED.

A. Creación de ponderaciones para estimar la cobertura censal

Como se mencionó anteriormente, el método de emparejamiento entre el censo y la encuesta de posenumeración censal (EPC) permite estimar la omisión censal mediante diferentes tipos de estimadores. El estimador de Petersen constituye el punto de partida clásico del sistema de estimación dual (SED). No obstante, su aplicación directa supone que todos los registros censales son válidos y comparables, lo que rara vez ocurre en la práctica. En los censos reales, la calidad de los registros varía: algunos corresponden a enumeraciones correctas, otros contienen información incompleta o inconsistente, e incluso pueden existir duplicados. Por ello, es necesario incorporar mecanismos que ajusten las estimaciones para reflejar la calidad de los datos censales y la confiabilidad del emparejamiento resultante.

Con este propósito, es posible asignar a cada registro censal un peso proporcional a la probabilidad de que la persona haya sido correctamente enumerada. Para cada individuo i, se definen los siguientes eventos, en forma de variables dicotómicas que toman el valor de uno, si sucede el evento, o cero, en otro caso:

- c_i: indica si el registro corresponde a una enumeración correcta y completa; es decir, una persona que fue censada en el lugar donde debía ser contada y cuya ficha contiene toda la información básica requerida (sexo, edad, parentesco, etc.). Este tipo de registros se consideran los de mayor calidad dentro del censo, ya que cumplen simultáneamente con los criterios de cobertura y contenido.
- m_i: indica si el registro fue correctamente emparejado entre la EPC y el censo.

A partir de estos elementos, es posible modelar $p(c_i)$, la probabilidad de ocurrencia de una enumeración correcta y completa y $p(m_i)$, la probabilidad de un emparejamiento. Si pensamos en probabilidades, un registro que realmente debería contarse en el estimador es aquel que tiene toda la información básica y es una enumeración es correcta, condicionado a que emparejó. Al usar la regla de Bayes, esta probabilidad se puede escribir como:

$$p(m_i \mid c_i) = \frac{p(c_i \mid m_i) \cdot p(m_i)}{p(c_i)}$$

Ahora, si el registro emparejó efectivamente, entonces necesariamente es completo y válido (pues tenía toda la información completa y correspondía a una enumeración correcta), es decir que $p(\, c_i \mid m_i \,) = 1$. Por lo tanto, se tiene que:

$$p(m_i \mid c_i) = \frac{p(m_i)}{p(c_i)}$$

Luego, utilizando el inverso multiplicativo de esta probabilidad, se puede definir un factor de corrección por cobertura que pondera cada registro censal según su probabilidad de ser válido. Esta ponderación tomaría la siguiente forma:

$$a_i = \frac{p(c_i)}{p(m_i)}$$

Este ajuste permite que las estimaciones de la cobertura censal reflejen tanto la calidad del emparejamiento como la integridad de la información censal. Por lo tanto, recordando que, $\widehat{N}_{1+}^{0}$ denota la cardinalidad del conjunto censado (total censal bruto), un estimador del SED basado en probabilidades estaría definido por la siguiente expresión:

$$\widehat{N}_{++} = \sum_{i=1}^{\widehat{N}_{1+}^{0}} a_i = \sum_{i=1}^{\widehat{N}_{1+}^{0}} \frac{p(c_i)}{p(m_i)}$$

Esta ponderación se aplica de manera individual a cada registro censal presente en la base de datos, de modo que cada observación contribuye al estimador final según su probabilidad de ser válida, completa y emparejada. En el caso ideal, cuando todas estas probabilidades toman el valor máximo ($p(c_i) = p(m_i) = 1$), cada registro aporta exactamente una unidad al total, y el estimador $\widehat{N}_{++}$coincide con el conteo bruto de registros censales, porque el censo sería perfecto. Sin embargo, en escenarios más realistas, estas probabilidades pueden ser menores que 1, reflejando imperfecciones en la cobertura o en el emparejamiento; por ejemplo, un registro incompleto ($p(c_i) < 1$) reduce proporcionalmente su contribución, mientras que un registro con coincidencia incierta ($p(m_i) < 1$) aumenta su peso al dividirse entre una probabilidad menor. Por supuesto, en la vida real, estas probabilidades no se pueden conocer para cada individuo censado, por lo que tienen que ser estimadas o predichas usando los datos de la EPC.

B. Modelamiento de las probabilidades mediante regresión logística

Las probabilidades $p(c_i)$ y $p(m_i)$ que intervienen en el factor de corrección de cobertura pueden estimarse mediante modelos de regresión logística. Estos modelos permiten representar la probabilidad del evento de interés (ser una enumeración correcta, tener datos completos o constituir un emparejamiento) en función de un conjunto de covariables observadas, tanto a nivel individual como contextual. Entre las covariables más utilizadas se incluyen variables demográficas (edad, sexo, parentesco con el jefe del hogar), socioeconómicas (nivel educativo, condición de ocupación), y geográficas (tipo de área, región).

Un aspecto importante es que el modelo $p(m_i)$ se ajusta utilizando los datos de la muestra de la enumeración (muestra E), la cual contiene registros cuidadosamente emparejados entre el censo y la encuesta. Esto permite estimar la probabilidad de que un registro específico corresponda efectivamente a la misma unidad en ambas fuentes. Por otro lado, la muestra P proporciona información sobre la calidad y completitud de los datos censales básicos, lo que permite estimar las probabilidades $p(c_i)$ asociadas a cada registro. Al combinar estos conjuntos de datos, los modelos logran capturar tanto la precisión del emparejamiento como la integridad de la información censal, ajustando de manera individual el aporte de cada registro al estimador final. Esta estrategia asegura que las estimaciones reflejen de forma más realista la cobertura efectiva del censo, compensando las deficiencias en los registros individuales y proporcionando una medida robusta y ajustada de la población total

El modelo de regresión logística permite predecir la probabilidad de que una enumeración censal sea correcta y la probabilidad de coincidencia en el emparejamiento. Además, permite incluir variables continuas y utilizar únicamente los términos de interacción estadísticamente significativos, lo que contribuye a reducir tanto el sesgo por heterogeneidad como el error de muestreo. Por ejemplo, considere la variable binaria m_i que en la muestra P se puede definir como:

$$m_i = \begin{cases} 1, & \textit{si el registro i es un emparejado} \\ 0, & \textit{si el registro i es una omisión} \end{cases}$$

La probabilidad aproximada del suceso se expresa mediante la función logística En términos generales, para una probabilidad de interés $p(i)$, se especifica un modelo logístico de la forma:

$$p(m_i) = \frac{exp(x_i\beta)}{1 + exp(x_i\beta)}$$

En donde x_i es el vector de covariables asociadas al registro i, y β es el vector de parámetros del modelo correspondiente. A partir de la estimación de los parámetros, se obtienen las probabilidades predichas $\hat{p}(i)$, usando técnicas apropiadas que incluyan el diseño de muestreo complejo en la inferencia (Heeringa, West, y Berglund, 2017), la probabilidad estimada de que la variable de interés tome el valor uno, que a su vez es también la esperanza de la variable de interés, en un modelo de regresión logística es la siguiente:

$$\hat{p}(m_i) = \frac{exp\ (x_i\hat{\beta})}{1 + exp\ (x_i\hat{\beta})}$$

Usando un enfoque similar es posible ajustar un modelo para estimar $\hat{p}(c_i)$. El uso de modelos logísticos presenta varias ventajas. En primer lugar, permite incorporar información auxiliar proveniente del censo, de la EPC o de otras fuentes, mejorando la precisión de las estimaciones. En segundo lugar, facilita la identificación de patrones sistemáticos de error, como mayores tasas de omisión en determinados grupos poblacionales o áreas geográficas. Finalmente, la estimación puede extenderse a modelos con efectos aleatorios o estructuras jerárquicas, lo que permite capturar la variabilidad entre dominios geográficos o estratos censales y obtener estimaciones más estables en presencia de tamaños muestrales pequeños.

Este tipo de modelos constituye una alternativa más flexible que la posestratificación y ha demostrado reducir de manera más efectiva el sesgo en las estimaciones de la población total (Olson y Sands 2012). Esto reduce el sesgo de correlación sin necesidad de recurrir a estructuras excesivamente complejas. Otra ventaja importante es que este enfoque admite la incorporación de un mayor número de covariables, tanto categóricas como continuas, ampliando así el conjunto de predictores potenciales que pueden contribuir a mejorar la precisión del ajuste. Asimismo, facilita procesos de selección de variables y comparación de modelos, lo que refuerza su uso en aplicaciones de estimación poblacional.

Una vez estimadas las probabilidades $\hat{p}(c_i)$ y $\hat{p}(m_i)$ mediante los respectivos modelos logísticos, es posible estimar el factor de corrección por cobertura de cada registro censal, de la siguiente manera:

$$\hat{a}_i = \frac{\hat{p}(c_i)}{\hat{p}(m_i)}$$

En particular, $\hat{a}_i$ actúa como un ponderador de calidad que corrige las imperfecciones del censo y del proceso de emparejamiento, garantizando que los registros más confiables tengan mayor influencia en la estimación de la cobertura. Luego, con los factores $\hat{a}_i$ disponibles, el total ajustado de la población de personas en la población se estima mediante la siguiente suma ponderada:

$$\widetilde{N}_{++} = \sum_{i=1}^{\widehat{N}_{1+}^{0}} \hat{a}_i = \sum_{i=1}^{\widehat{N}_{1+}^{0}} \frac{\hat{p}(c_i)}{\hat{p}(m_i)}$$

Así mismo, dado que cada uno de los registros en la base de datos censal tiene un ponderador asociado, entonces es posible estimar el tamaño de cualquier subpoblación U_g, de la siguiente forma:

$$\widetilde{N}_{g,++} = \sum_{i \in U_g} \frac{\hat{p}(c_i)}{\hat{p}(m_i)}$$

En este caso, es importante resaltar que los registros que se emparejan correctamente entre el censo y la EPC pueden encontrarse en distintos lugares geográficos, lo que tiene implicaciones distintas para el cálculo de las tasas de omisión.

- A nivel nacional, cualquier registro que se haya emparejado correctamente se considera como observado, por lo que la variable indicadora de emparejamiento toma el valor de uno.

Esto permite que la estimación nacional refleje de manera más precisa la cobertura censal general, sin importar la ubicación específica de los registros.

- Sin embargo, a nivel subnacional la situación es diferente. Un registro correctamente emparejado que se encuentre fuera del área geográfica de búsqueda no se contabiliza como emparejado dentro de esa unidad subnacional, por lo que la variable indicadora tomaría el valor de cero. Como consecuencia, se pueden producir tasas de omisión subnacionales más altas que las observadas a nivel nacional, incluso cuando la cobertura general es buena.

Esta diferencia subraya la importancia de considerar la ubicación de los emparejamientos al realizar estimaciones subnacionales. No todos los registros que se emparejan correctamente a nivel nacional mejoran automáticamente la estimación local, y factores como la migración interna, la localización incorrecta de los registros o errores en la asignación geográfica pueden afectar la calidad de las estimaciones regionales. Por ello, al interpretar tasas de omisión subnacionales, es fundamental tener en cuenta que la cobertura efectiva puede variar significativamente de una región a otra, aun cuando la tasa nacional indique un alto nivel de coincidencia entre censo y encuesta.

C. Ejemplo aplicado

En esta sección abordaremos de manera detallada un ejemplo de la aplicación de los modelos estadísticos propuestos para comprender sus particularidades. En el ejemplo de la EPC, el conteo bruto censal fue $N_{1+}^{0} = 1\,500\,000$ y, tras identificar $\widehat{N}_{EE} = 20\,000$ enumeraciones incorrectas, el total censal corregido es $N_{1+} = 1\,480\,000$. Para modelar la probabilidad de que un registro individual i corresponda a una enumeración correcta y completa, se puede definir la variable indicadora c_i, que toma el valor de uno si la enumeración es correcta y cero si es errónea. Suponiendo que la calidad de la enumeración puede depender de ciertas características del individuo y su hogar, un modelo logístico plausible sería:

$$logit(\hat{p}(c_i)) = \alpha + \beta_1 \cdot Edad_i + \beta_2 \cdot Sexo_i + \beta_3 \cdot Zona_i + \beta_4 \cdot NivelEduc_i$$

Donde las covariables estarían dadas por la edad del individuo, ya que personas muy jóvenes o mayores podrían estar subrepresentadas, el sexo (hombre o mujer), considerando posibles diferencias en cobertura, la zona (urbana o rural), dado que la densidad y accesibilidad afectan la calidad del registro y el nivel educativo (del jefe de hogar o del mismo individuo), que puede correlacionarse con la probabilidad de aparecer correctamente en el censo. Numéricamente, usando la proporción de enumeraciones correctas detectada, el promedio estimado de $\hat{p}(c_i)$ sería:

$$\hat{p}(c_i) \approx \frac{N_{1+}}{N_{1+}^{0}} = \frac{1\,480\,000}{1\,500\,000} \approx 0.9867$$

A partir del modelo logístico es posible ajustar esta probabilidad según las covariables de cada individuo, capturando heterogeneidad individual en la calidad de la enumeración y ofreciendo una base más robusta para el análisis de la cobertura censal y estimación de errores de enumeración. Suponga que a partir del modelo anterior, se estimaron los coeficientes de regresión del cuadro 18.

Según los resultados del modelo ajustado, el intercepto indica que, para un individuo la edad tiene un efecto negativo pequeño, mostrando que los individuos más jóvenes o muy mayores tienen mayor riesgo de errores de enumeración. El sexo muestra un efecto positivo leve, indicando que los hombres tienen ligeramente más probabilidad de ser correctamente enumerados. Vivir en zona rural reduce la probabilidad, reflejando dificultades logísticas y de cobertura. Finalmente, un mayor nivel educativo del jefe del hogar se asocia con un incremento en la probabilidad de enumeración correcta, posiblemente porque facilita la recolección de información completa y precisa.

Cuadro 18
Coeficientes de regresión estimados para el modelo de enumeraciones correctas

Covariable	Coeficiente estimado $(\hat{\beta})$
Intercepto	2,000
Edad (por año)	-0,015
Sexo (1=Hombre)	0,050
Zona (1=Rural)	-0,100
Nivel Educativo (por nivel)	0,120

Fuente: Elaboración propia.

Para ilustrar cómo las covariables influyen en la probabilidad de emparejamiento, consideremos un individuo de 40 años, hombre, que vive en zona rural y cuyo jefe de hogar tiene nivel educativo 3 (secundaria completa). En este caso, se observa que:

$$logit(\hat{p}(c_i)) = 2.0 + (-0.015 \cdot 40) + (0.05 \cdot 1) + (-0.10 \cdot 1) + (0.12 \cdot 3) = 1.71$$

Aplicando la función logística, se obtiene la probabilidad correspondiente:

$$\hat{p}(c_i) = \frac{e^{1.71}}{1 + e^{1.71}} \approx 0.847$$

Por consiguiente, para este individuo, la probabilidad estimada de que su registro sea una enumeración correcta es aproximadamente 84.7 %, ligeramente superior a la probabilidad promedio del 81 % estimada para todo el censo corregido.

Así mismo, para modelar las enumeraciones correctas pero incompletas a nivel individual, definimos la variable indicadora m_i que toma el valor de uno si el registro del individuo fue correctamente emparejado, y toma el valor de cero, en otro caso. En este modelo, las covariables reflejan factores que podrían influir en la probabilidad de que un registro individual sea correctamente emparejado entre el censo y la encuesta. Para ilustrar, consideramos las siguientes covariables: la edad, dado que los extremos de edad podrían presentar mayor dificultad para un emparejamiento preciso; el sexo, por posibles diferencias en la calidad de los registros entre hombres y mujeres; el tipo de hogar, ya que hogares unipersonales o con jefatura femenina podrían presentar registros más difíciles de vincular; el material de la vivienda, pues residencias construidas con materiales precarios podrían asociarse con información parcial o errores en el emparejamiento; el acceso a servicios básicos, dado que la disponibilidad de electricidad, agua o internet puede facilitar o dificultar la correcta identificación de los registros; y, finalmente, la zona (rural o urbana), ya que la distancia y la accesibilidad influyen en la calidad y la completitud del emparejamiento. Siendo así, un posible modelo estaría dado por:

$$logit(\hat{p}(m_i)) = \alpha + \beta_1 \cdot Edad_i + \beta_2 \cdot Sexo_i + \beta_3 \cdot TipoHogar_i + \beta_4 \cdot MaterialVivienda_i + \beta_5 \cdot AccesoServicios_i + \beta_6 \cdot Zona_i$$

Usando los datos del ejemplo, tras eliminar las enumeraciones incorrectas detectadas por la EPC, el total censal corregido es $N_{1+} = 1\,480\,000$ y, de estos, $N_{11} = 1\,200\,000$ registros coinciden con la encuesta postcensal. Por lo tanto, el promedio estimado de $\hat{p}(m_i)$ se calcula como la proporción de coincidencias dentro del censo:

$$\hat{p}(m_i) \approx \frac{N_{11}}{N_{1+}} = \frac{1\,200\,000}{1\,480\,000} \approx 0.8108$$

Esto indica que, aproximadamente, el 81 % de los registros censales corregidos corresponde efectivamente a la misma unidad observada en la encuesta. Suponga que a partir del modelo anterior, se estimaron los coeficientes de regresión del cuadro 19.

En este modelo logístico, el intercepto refleja la probabilidad base de que un registro se empareje correctamente para un individuo promedio con características de referencia. La edad tiene un efecto ligeramente negativo, indicando que personas más jóvenes o mayores tienen mayor dificultad de emparejamiento. El sexo muestra un efecto positivo pequeño, sugiriendo que los hombres, en este ejemplo, tienen una probabilidad ligeramente mayor de coincidir correctamente entre censo y encuesta. Asimismo, el tipo de hogar y el material de la vivienda muestran efectos negativos más pronunciados, reflejando que los hogares unipersonales o con construcción precaria tienden a presentar registros más difíciles de emparejar. Por el contrario, un buen acceso a servicios incrementa la probabilidad de coincidencia, probablemente porque facilita la recolección de información completa. Finalmente, residir en una zona rural disminuye ligeramente la probabilidad de emparejamiento, debido a factores de accesibilidad y dispersión geográfica.

Cuadro 19
Coeficientes de regresión estimados para el modelo de emparejamiento

Covariable	Coeficiente estimado $(\hat{\beta})$
Intercepto	1,50
Edad (por año)	-0,01
Sexo (1=Hombre)	0,05
Tipo de hogar (1=unipersonal)	-0,20
Material de vivienda (1=precario)	-0,30
Acceso a servicios (1=bueno)	0,25
Zona (1=rural)	-0,15

Fuente: Elaboración propia.

De manera ilustrativa, consideremos el mismo individuo de 40 años, hombre, que vive en un hogar ubicado en una zona rural, que además es un hogar unipersonal, en una vivienda de materiales buenos, y con buen acceso a servicios. En este caso, se tiene que:

$$logit(\hat{p}(m_i)) = 1.5 - 0.4 + 0.05 - 0.2 + 0 + 0.25 - 0.15 = 1.05$$

Para obtener la probabilidad de interés, se aplica la función inversa:

$$\hat{p}(m_i) = \frac{e^{1.05}}{1 + e^{1.05}} \approx 0.74$$

Según los coeficientes del modelo logístico, este individuo tiene una probabilidad estimada de aproximadamente 74 % de que su registro censal coincida correctamente con la EPC. Esto refleja cómo ciertos factores, como residir en un hogar unipersonal y en zona rural, reducen ligeramente la probabilidad de emparejamiento, mientras que ser hombre y contar con buen acceso a servicios incrementa dicha probabilidad. Comparado con la probabilidad promedio del 81 % observada en todo el censo corregido, este ejemplo permite ver de manera concreta cómo la heterogeneidad individual puede influir en la calidad del emparejamiento de registros

A partir de las probabilidades estimadas de enumeración correcta $\hat{p}(c_i)$ y de coincidencia entre censo y encuesta $\hat{p}(m_i)$ es posible estimar el ponderador de cobertura $\hat{a}_i = \hat{p}(c_i)/\hat{p}(m_i)$ para ajustar el peso de cada registro individual en la base de datos censal. Por un lado, la probabilidad de que la enumeración sea correcta aumenta el valor de registros que podrían haber sido omitidos o mal

registrados; por otro lado, la división por la probabilidad de que un registro coincida entre las fuentes corrige el efecto de las coincidencias parciales o errores de emparejamiento.

Para ilustrarlo, usando los valores promedio estimados previamente, el factor de ajuste individual sería aproximadamente:

$$\hat{a}_i = \frac{\hat{p}(c_i)}{\hat{p}(m_i)} \approx \frac{0.9867}{0.8108} \approx 1.217.$$

Esto significa que cada registro censal corregido contribuye en promedio un 21.7 % más al total estimado para compensar tanto las enumeraciones incorrectas como la incompletitud del emparejamiento. Por supuesto, dependiendo de los modelos y de sus covariables asociadas, cada individuo tendrá un valor diferente en el ponderador de cobertura $\hat{a}_i$. Por ejemplo, para el mismo individuo de 40 años, hombre, que vive en un hogar ubicado en una zona rural, que además es un hogar unipersonal, en una vivienda de materiales buenos, y con buen acceso a servicios, se había estimado que la probabilidad de enumeración correcta era aproximadamente 0.847 y que su probabilidad de emparejamiento era de 0.741. Por ende, su factor de ajuste a_i será:

$$\hat{a}_i = \frac{0.847}{0.741} \approx 1.143$$

Es decir que su registro contribuye alrededor de un 14.3 % más al total estimado, ajustando tanto por la probabilidad de que la enumeración sea correcta como por la probabilidad de emparejamiento. Por ende, sumando estos valores sobre todos los registros del censo, se obtendría $\widetilde{N}_{++}$, un estimador ajustado del total poblacional que integra la información de la EPC y del censo de manera coherente, a través de modelos logísticos.

VII. Conclusiones y recomendaciones

Este documento se elaboró con el propósito de reunir en un solo texto la fundamentación y racionalidad de los métodos estadísticos que intervienen en el diseño y análisis de las encuestas de posenumeración censal (EPC), con un enfoque integral para el diseño, ejecución y análisis, el cual debe estar basado siempre en el Sistema de Estimación Dual (SED). En este sentido, es muy importante comprender la importancia de los supuestos que sustentan los métodos de captura y recaptura, así como las propiedades de los estimadores empleados. La correcta aplicación de estos principios asegura que la información obtenida a través de la EPC sea exacta, precisa, consistente y confiable para estimar la cobertura censal.

La planificación de la EPC, incluyendo la construcción del cuestionario, el diseño muestral, la determinación del tamaño de muestra y el operativo de recolección, constituye un elemento estratégico que condiciona la calidad de los resultados. Este documento se destaca que la selección de los elementos de la muestra y la definición de los objetivos de la encuesta deben estar alineadas con las condiciones requeridas por el SED. La vinculación de los registros censales con los de la encuesta depende críticamente de estas decisiones, de manera que errores en la planificación pueden amplificarse en las etapas posteriores, afectando tanto la validez de los emparejamientos como la estimación de la omisión censal.

La definición y uso de las muestras E y P es otro componente clave. Estos conceptos permiten establecer con claridad cuándo una enumeración es correcta, incorrecta o incompleta, y constituyen la base sobre la cual se construyen los emparejamientos estadísticos. La correcta identificación de los registros en estas muestras permite reconstruir los hogares y personas tal como existían en el momento de la recolección censal, minimizando errores derivados de duplicaciones o enumeraciones parciales. El análisis detallado de las muestras y la aplicación de criterios estandarizados de clasificación facilitan un enfoque sistemático, aumentando la confiabilidad de las estimaciones de cobertura.

El proceso de emparejamiento estadístico, que incluye limpieza de datos, estandarización de textos, búsqueda en bloques, clasificación y evaluación, es fundamental para garantizar la consistencia de los registros. Cada etapa del flujo de trabajo cumple un rol específico: la limpieza asegura la coherencia interna de los datos, la estandarización facilita la comparación entre registros de ambas fuentes, la búsqueda en bloques agiliza el emparejamiento, la clasificación mediante modelos de vinculación probabilística permite tener mayor precisión y, por último, la evaluación y revisión clerical

permite identificar errores y ajustar procedimientos. Este enfoque estructurado garantiza que los registros sean vinculados de manera precisa, minimizando errores que podrían afectar tanto las estimaciones nacionales como las subnacionales.

La aplicación de los estimadores del SED, junto con los factores de expansión, el cálculo de varianzas y los intervalos de confianza, proporciona una base sólida para la inferencia estadística. La incorporación de modelos estadísticos más flexibles permite manejar situaciones en las que algunos supuestos del SED no se cumplen estrictamente, ofreciendo herramientas robustas para estimar el tamaño poblacional, junto con medidas de calidad.

Por otro lado, los modelos logísticos para estimar la probabilidad de omisión permiten identificar patrones sistemáticos asociados con características demográficas, sociales y territoriales. Covariables como la edad, el sexo, el tipo de hogar, los materiales de la vivienda, el acceso a servicios, la zona geográfica y el nivel educativo pueden usarse para predecir la probabilidad de que una enumeración sea correcta o que un registro se encuentre emparejado entre ambas fuentes. Este análisis facilita la identificación de grupos de mayor riesgo de omisión, permitiendo diseñar estrategias específicas de recolección y supervisión que optimicen la cobertura censal. Asimismo, la creación de factores de ajuste individuales, calculados con estos modelos, reflejan la heterogeneidad de los registros y aseguran que cada unidad contribuya al total de manera proporcional a la calidad de su información, permitiendo obtener estimaciones del tamaño de la población directamente desde la base de datos del censo.

Una recomendación muy importante es que los análisis diferenciales por subgrupos de interés se realicen con precaución, especialmente los concernientes con la estimación de la omisión a nivel subnacional. A nivel nacional, los registros correctamente emparejados contribuyen plenamente a la estimación de la cobertura, mientras que a nivel subnacional un registro fuera del área geográfica considerada aumenta la tasa de omisión local. Este hecho evidencia que una cobertura nacional adecuada puede coexistir con problemas de omisión localizados, subrayando la necesidad de analizar los resultados en distintos niveles geográficos y de ajustar los factores de ponderación según la ubicación de los registros.

Finalmente, se resalta que la correcta aplicación del SED y de los modelos estadísticos asociados no solo mejora la estimación de la omisión censal, sino que también proporciona información valiosa para la planificación de futuras EPC y censos. La identificación de subgrupos específicos como factores de riesgo de omisión censal, la ponderación adecuada de los registros y la consideración de la ubicación geográfica de los emparejamientos permiten obtener una visión más completa y precisa de la cobertura censal. Este enfoque integrado constituye una base metodológica sólida para fortalecer la calidad de la información estadística, orientar decisiones operativas y políticas, y asegurar que las estimaciones poblacionales reflejen con fidelidad la realidad de la población.

Bibliografía

American Association for Public Opinion Research (AAPOR). (2016). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys (9th ed.). AAPOR.

Baffour, Bernard, Thomas King, and Paolo Valente. (2013). "The Modern Census: Evolution, Examples and Evaluation." International Statistical Review 81 (3): 407–25.

Binder, David A. (2011). "Estimating Model Parameters from a Complex Survey Under a Model-Design Randomization Framework." Pakistan Journal of Statistics 27 (4): 371–90.

Bishop, Yvonne M, Stephen E Fienberg, y Paul W Holland. (2007). Discrete Multivariate Analysis: Theory and Practice. Springer Science & Business Media.

Borges, G. y Queiros, J. (2025). Evaluación de cobertura del Censo de Brasil 2022: metodología y resultados preliminares. Notas de Población 120, Santiago, Comisión Económica para América Latina y el Caribe (CEPAL).

Bureau, US Census. (2022). "2020 PostEnumeration Survey Estimation Design."

Comisión Económica para América Latina y el Caribe CEPAL. (1999). América Latina: aspectos conceptuales de los censos del 2000, Serie Manuales.

Comisión Económica para América Latina y el Caribe CEPAL. (2014). Los datos demográficos: alcances, limitaciones y métodos de evaluación. Serie Manuales.

Comisión Económica para América Latina y el Caribe CEPAL. (2023). Diseño y Análisis Estadístico de Las Encuestas de Hogares de América Latina. Metodologías de La CEPAL.

Chackiel, J. (2010), "Evaluación post-empadronamiento de la cobertura en los censos de población", Notas de Población, N° 91 (LC/G.2484-P), Santiago, Comisión Económica para América Latina y el Caribe (CEPAL).

Chao, Anne. (1987). "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability." Biometrics, 783–91.

Chao, Anne. (1989). "Estimating Population Size for Sparse Data in Capture-Recapture Experiments." Biometrics, 427–38.

Chapman, Douglas George. (1951). "Some Properties of the Hypergeometric Distribution with Applications to Zoological Censuses." Univ. Calif. Stat. 1: 60–131.

Christen, Peter. (2012). Data Matching Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer.

Cormack, Richard M. (1989). "Log-Linear Models for Capture-Recapture." Biometrics, 395–413.
de Bruin, J. (2019). *recordlinkage: A toolkit for record linkage and duplicate detection in Python*.
Dunn, H. L. (1946). "Record Linkage." American Journal of Public Health and the Nations Health 36 (12): 1412–16. https://doi.org/10.2105/AJPH.36.12.1412
Enamorado, T., Fifield, B., & Imai, K. (2019). *fastLink: Fast Probabilistic Record Linkage*. R package version 0.6.0.
Fellegi, Ivan P, y Alan B Sunter. (1969). "A Theory for Record Linkage." Journal of the American Statistical Association 64 (328): 1183–1210.
Fienberg, Stephen E. (1972). "The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables." Biometrika 59 (3): 591–603.
Gregg, F. & Eder, D. (2022). *Dedupe: Python library for accurate and scalable fuzzy matching, record deduplication and entity-resolution*.
Griffin, Robert. (2000). "Accuracy and Coverage Evaluation Survey: Definition and Explanation of Correlation and Related Heterogeneity Bias." Q-35. DSSD Census 2000 Procedures and Operations Memorandum Series. U.S. Census Bureau.
Gutiérrez, Hugo Andrés. (2016). Estrategias de Muestreo: Diseño de Encuestas y Estimación de Parámetros. Segunda edición. Ediciones de la U.
Hawthorn, F. (2020). *Dedupe: Python library for accurate and scalable data deduplication and entity resolution*.
Heeringa, S., West, B. T., & Berglund, P. A. (2017). Applied survey data analysis (Second edition). CRC Press, Taylor & Francis Group.
Herzog, Thomas N, Fritz J Scheuren, y William E Winkler. (2007). Data Quality and Record Linkage Techniques. Vol. 1. Springer.
Hogan, Howard. (2003). "The Accuracy and Coverage Evaluation: Theory and Design." Survey Methodology 29 (2): 129–38. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20030026444.
Larsen, Michael D, y Donald B Rubin. (2001). "Iterative Automated Record Linkage Using Mixture Models." Journal of the American Statistical Association 96 (453): 32–41.
Lincoln, F. C. (1930). "Calculating Waterfowl Abundance on the Basis of Banding Returns." Circular 118: 1–4.
Lynch, Billy T, William L Arends, et al. (1977). "Selection of a Surname Coding Procedure for the SRS Record Linkage System." Washington, DC: US Department of Agriculture, Sample Survey Research Branch, Research Division.
Menestrina, David, Steven Euijong Whang, y Hector Garcia-Molina. (2010). "Evaluating Entity Resolution Results." Proceedings of the VLDB Endowment 3 (1-2): 208–19.
Mulry, Mary H, y Bruce D Spencer. (1991). "Total Error in EPC Estimates of Population." Journal of the American Statistical Association 86 (416): 839–55.
Nauman, Felix, y Melanie Herschel. (2022). An Introduction to Duplicate Detection. Springer Nature.
Navarro, Gonzalo. (2001). "A Guided Tour to Approximate String Matching." ACM Computing Surveys (CSUR) 33 (1): 31–88.
Newcombe, H. B., J. M. Kennedy, S. J. Axford, y A. P. James. (1959). "Automatic Linkage of Vital Records: Computers Can Be Used to Extract" Follow-up" Statistics of Families from Files of Routine Records." Science 130 (3381): 954–59. https://doi.org/10.1126/science.130.3381.954
Nour, El-Sayed. (1982). "On the Estimation of the Total Number of Vital Events with Data from Dual Collection Systems." Journal of the Royal Statistical Society Series A: Statistics in Society 145.
Odell, Margaret, y Robert Russell. (1918). "The Soundex Coding System." US Patents 1261167: 9.
Olson, D., y R. Sands. (2012). "2010 Census Coverage Measurement Estimation Report: Net Coverage Comparison with PostStratification." DSSD 2010 Census Coverage Measurement Memorandum Series 2010-G-12. U.S. Census Bureau.
Petersen, Carl G. J. (1896). "The Yearly Immigration of Young Plaice into the Limfjord from the German Sea." Report of the Danish Biological Station 6: 1–48.
Philips, Lawrence. (1990). "Hanging on the Metaphone." Computer Language 7 (12): 39–43.
Robinson, D., Bryan, J., Elias, J. (2020). *fuzzyjoin: Join Tables Together on Inexact Matching*. CRAN package "fuzzyjoin".
Sadinle, Mauricio. (2009). "Transformed Logit Confidence Intervals for Small Populations in Single Capture–Recapture Estimation." Communications in Statistics-Simulation and Computation 38 (9): 1909–24.

Sariyar, M., Borg, A., & Schnell, R. (2022). *RecordLinkage: Record Linkage in R*. R package version 0.4-13. Available at: https://CRAN.R-project.org/package=RecordLinkage

Särndal, Carl-Erik, Bengt Swensson, y Jan Wretman. (2003). Model Assisted Survey Sampling. Springer Science; Business Media.

Schnabel, Z. E. (1938). "The Estimation of the Total Fish Population of a Lake." American Mathematical Monthly 45: 348–52.

Seber, George Arthur Frederick. (1982). The Estimation of Animal Abundance and Related Parameters.

Sekar, C. Chandra, y W. Edwards Deming. (1949). "On a Method of Estimating Birth and Death Rates and the Extent of Registration." Journal of the American Statistical Association 44 (245): 101–15.

United Nations. (2010). Post Enumeration Surveys: Operational Guidelines. 2010 World Population and Housing Census Programme.

United Nations. (2025). Principles and recommendations for population and housing censuses (Revision 4). United Nations.

van der Loo, M. (2014). *The stringdist package for approximate string matching*. The R Journal, 6(1), 111–122.

Webster, Anthony J, y Richard Kemp. 2013. "Estimating Omissions from Searches." The American Statistician 67 (2): 82–89.

Winkler, William E. (2000). Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. US Bureau of the Census Washington, DC.

Whitford, D. C. y J. P. Banda. (2002). ¿Vale la pena hacer encuestas de post-empadronamiento censal?, Notas de Población, N° 75 (LC/G.2186-P), Santiago, Comisión Económica para América Latina y el Caribe (CEPAL).

Wolter, Kirk M. (1986). "Some Coverage Error Models for Census Data." Journal of the American Statistical Association 81 (394): 338–46. https://doi.org/10.2307/2289222

Zamora, J. (2022). "2020 PostEnumeration Survey: Estimation Design." 2020-J-03. DSSD 2020 PostEnumeration Survey Memorandum Series. U.S. Census Bureau.

Zelterman, Daniel. (1988). "Robust Estimation in Truncated Discrete Distributions with Application to Capture-Recapture Experiments." Journal of Statistical Planning and Inference 18 (2): 225–37.

NACIONES UNIDAS

CEPAL

Serie

Estudios Estadísticos

Números publicados

Un listado completo, así como los archivos pdf están disponibles en www.cepal.org/publicaciones

111. Encuestas de posenumeración censal: fundamentos estadísticos para su diseño y análisis, Andrés Gutiérrez y Giovany Babativa-Márquez (LC/TS.2025/117), 2026
110. La situación de las estadísticas, indicadores y cuentas ambientales en América Latina y el Caribe, 2023, Georgina Alcantar López, Alberto Malmierca y Analía Pérez Quesada (LC/TS.2025/112), 2025
109. Estimación en áreas pequeñas de los indicadores de pobreza en América Latina: una aplicación basada en modelos de regresión multinivel con posestratificación, Andrés Gutiérrez, Xavier Mancero, Stalyn Guerrero (LC/TS.2024/137), 2025
108. Servicios de intermediación financiera medidos indirectamente: medición en un contexto de tasas de interés subsidiadas y existencia de encajes bancarios, Federico Dorin, Lourdes Erro, Salvador Marconi y Juan Carlos Propatto (LC/TS.2024/17), 2024
107. Índice de mala calidad del empleo: una exploración de la última década en América Latina, Mauricio Apablaza, Pablo Villatoro, Pablo González, Kirsten Sehnbruch y Xavier Mancero (LC/TS.2023/199), 2024
106. Efectos de diseño para indicadores sociales en América Latina: función generalizada de varianza para estimadores directos provenientes de encuestas de hogares, Andrés Gutiérrez y Giovany Babativa-Márquez (LC/TS.2023/95), 2023.
105. Modelos de unidad para la generación de mapas de pobreza a nivel subnacional, Andrés Gutiérrez, Xavier Mancero, Gabriel Nieto, Felipe Molina y Diego Lemus (LC/TS.2022/191), 2022.
104. Cambio de año de referencia de los agregados regionales anuales en las cuentas nacionales, C. de Camino y otros (LC/TS. 2022/158), 2022.
103. Predicciones agregadas de pobreza con información a escala micro y macro: evaluación, diagnóstico y propuestas, Walter Sosa Escudero y Magdalena Cornejo (LC/TS.2022/95), 2022.
102. La medición de la discriminación en base al autorreporte: estado de situación y desafíos, Pablo Villatoro (LC/TS. 2021/87), 2021.

ESTUDIOS ESTADÍSTICOS

Números publicados:

111 Encuestas de posenumeración censal
Fundamentos estadísticos para su diseño y análisis
Andrés Gutiérrez y Giovany Babativa-Márquez

110 La situación de las estadísticas, indicadores y cuentas ambientales en América Latina y el Caribe, 2023
Georgina Alcantar López, Alberto Malmierca Castaño y Analía Pérez Quesada

109 Estimación en áreas pequeñas de los indicadores de pobreza en América Latina
Una aplicación basada en modelos de regresión multinivel con posestratificación
Andrés Gutiérrez, Xavier Mancero y Stalyn Guerrero

CEPAL

Comisión Económica para América Latina y el Caribe (CEPAL)
Economic Commission for Latin America and the Caribbean (ECLAC)
www.cepal.org

Versión digital disponible online

https://bit.ly/CEPAL2025-117S