# Metodología de ponderación del Censo 2023

Fecha de Publicación: 14 de Mayo de 2026

Instituto Nacional de Estadística

Esta página ha sido dejada intencionalmente en blanco

Por más información sobre este producto u otro del Instituto Nacional de Estadística

**Contacto**: Departamento de Difusión y Comunicación

Torre Ejecutiva Anexo, Piso 4 Liniers 1280, C.P.: 11.100

Tel: (598) 29027303, ints.: 7723, 7725

E-mail: difusion@ine.gub.uy

Sitio Web: https://www.ine.gub.uy

Twitter: https://twitter.com/ine_uruguay

**El siguiente documento fue elaborado por:**

Juan Pablo Ferreira - División Normalización, Investigación y Proyectos

Juan José Goyeneche - Consultor del Instituto Nacional de Estadística

**Agradecimientos**

Los autores agradecen especialmente los valiosos aportes y comentarios realizados por Yennyfer Feo (Unidad de Registros Administrativos) y Yohana Altez (División Normalización, Investigación y Proyectos) durante la elaboración y revisión de este documento.

**Nota:**

Con el objetivo de hacer la lectura de este documento más fluida se utiliza de forma indistinta las palabras estimador y estimación. Si bien existe una clara diferencia entre un estimador (variable aleatoria) y una estimación (realización del estimador), el objetivo de esta flexibilización en la terminología, es poder evitar frases complejas como ser: "el estimador de la varianza del estimador" y se remplaza, por ejemplo, por "la estimación de la varianza del estimador".

# Índice

**1. Introducción** **8**
1.1. Evaluación técnica de la cobertura y alcances del enfoque combinado 10
1.2. Estructura del documento 12
**2. Omisión censal** **14**
2.1. Omisión censal, tipos de omisión y relación con la no respuesta 14
2.2. Registros administrativos y sesgo de visibilidad institucional 16
2.3. El censo como muestra no probabilística 17
2.4. Límites de los enfoques tradicionales 18
2.5. Resumen del marco conceptual 19
**3. Diagnóstico empírico de la omisión censal** **21**
3.1. Estimación de la omisión a nivel de segmento censal 23
**4. Estrategia de corrección de la omisión censal** **27**
4.1. Objetivo de la ponderación 27
4.2. Unidad de análisis y unidad de ponderación 28
4.3. Muestra no probabilística 29
4.4. Enfoque doblemente robusto para la construcción de los ponderadores 30
4.4.1 Ponderación por probabilidad inversa (Inverse Probability Weighting, IPW) 31
4.4.2 Modelo superpoblacional 31
4.4.3 Estimador doblemente robusto 32
4.5. Variables auxiliares utilizadas en la construcción del estimador DR 34
4.6. Modelización de la probabilidad de respuesta 36
4.6.1 Especificación del modelo de propensión 38
4.6.2 Construcción de los ponderadores ajustados por omisión 39
4.7. Calibración a totales poblacionales del censo combinado 41
4.7.1 Calibración con ponderadores definidos a nivel de hogar 42
4.7.2 Unidades de inclusión forzosa 45
4.7.3 Unidades con ponderadores menor a uno. 46
4.7.4 Estimador Hayek a nivel de localidad 47
4.7.5 Controles de calidad y validación de los ponderadores 48
**5. Evaluación del impacto en indicadores censales** **49**
**6. Medidas de la calidad de las estimaciones** **50**
6.1. Intervalos de confianza 51
6.2. Coeficiente de Variación 52
**7. Supuestos y limitaciones del enfoque de ponderación** **54**
7.1 Supuestos del modelo de respuesta 54
7.2. Supuestos del modelo superpoblacional implícito en la calibración 54
7.3. Información auxiliar 55
7.4. Conclusiones 56
**8. Guía de uso de microdatos y cálculo de varianza en R** **57**
8.1. Conceptos básicos para la estimación con datos ponderados 57

8.1.1. Interpretación de los ponderadores 57
8.1.2. Consecuencias en el procesamiento y redondeo 58
8.2. Procedimiento de redondeo y preservación de totales 58
8.3. Cómputo de los errores estándar en R 60
**9. Glosario de términos y conceptos metodológicos 62**
**Abreviaturas 65**
**Bibliografía 66**

# Indice de Figuras

Figura 1: Diagrama Conceptual de la omisión censal

Figura 2: Estimación de la omisión según estrato socioeconómico para Montevideo

Figura 3: Estimación de la omisión según tramo de edad

Figura 4 : Omisión censal estimada mediante el MRTE por segmento censal. Montevideo

Figura 5 : Pobreza estimada mediante el modelo Fay- Harriot por segmento censal. Montevideo

Figura 6: Distribución de la estimación omisión por segmento censal

Figura 7: Distribución de la estimación omisión por segmento censal según departamento

Figura 8: Distribución por edad según distintas fuentes

Figura 9: Distribución de los factores de ajustes g provenientes de la calibración

Figura 10: Distribución de los factores de ajustes g por tramo de edad

Figura 11: Densidad bivariada entre los ponderadores ajustados por no respuesta y los ponderadores calibrados

Figura 12: Distribución del número de personas residentes en viviendas colectivas

Esta página ha sido dejada intencionalmente en blanco

# 1. Introducción

El Censo de Población, Hogares y Viviendas 2023, llevado adelante por el Instituto Nacional de Estadística (INE), estimó una población total de 3,499,451 personas, y presentó un nivel de omisión estimado en torno al 10,3% de la población, según los resultados de la Encuesta Nacional de Evaluación Censal (ENEC). Si bien este valor se encuentra dentro de los rangos observados en la última ronda de censos regionales y dentro de los aumentos crecientes de la omisión censal (United Nations, 2017), no resulta habitual en el contexto uruguayo, donde los censos previos se caracterizaron por niveles de cobertura más elevados. Esta magnitud de omisión plantea desafíos relevantes desde el punto de vista estadístico y analítico, particularmente en lo que respecta a la producción de estimaciones representativas de la población total y de sus principales características socioeconómicas.

En particular, la omisión censal no se distribuyó de forma homogénea en el territorio ni entre los distintos grupos poblacionales. La evidencia proveniente de la ENEC 2023 muestra que la omisión se concentró principalmente en zonas con mayores niveles de vulnerabilidad socioeconómica, caracterizadas por menores niveles educativos y mayores necesidades básicas insatisfechas, así como por un menor alcance efectivo de las modalidades de relevamiento implementadas, basadas en entrevistas presenciales asistidas por dispositivo (CAPI) y encuestas web autoadministradas (CAWI), con menor cobertura relativa en contextos de mayor vulnerabilidad. A modo ilustrativo, en Montevideo la tasa de omisión en el estrato socioeconómico bajo se situó en el entorno del 18%, mientras que en el estrato alto fue del orden del 3%, lo que evidencia una fuerte desigualdad en la cobertura censal. Adicionalmente, al analizar la distribución territorial, se observa que la omisión en el área rural duplica aproximadamente la omisión total, lo que constituye un rasgo especialmente relevante del proceso censal. Asimismo, la omisión presenta un patrón diferencial por edad y sexo, con una subrepresentación de personas jóvenes, especialmente en tramos etarios asociados a mayores niveles de movilidad residencial y participación en el mercado laboral (Groves & Couper, 1998), mientras que las tasas de omisión disminuyen progresivamente a edades más avanzadas. Estos patrones implican que la población efectivamente censada no reproduce de manera equilibrada la estructura demográfica y socioeconómica de la población total.

Con el objetivo de mitigar los efectos de esta omisión y garantizar una estimación adecuada del tamaño y la estructura de la población residente en el país, el INE adoptó un enfoque de *censo combinado*, integrando información proveniente del operativo censal con registros administrativos disponibles (Wallgren & Wallgren, 2014). A través de un *algoritmo* que utiliza, entre otros insumos, *señales de vida* (i.e. interacciones administrativas de las personas con fuentes de registros del estado), se incorporaron personas no censadas directamente en los microdatos del censo. Esta incorporación se limitó a un conjunto de variables básicas, tales como sexo, edad y una referencia geográfica aproximada, con el fin de alcanzar una población total coherente con las estimaciones de cobertura obtenidas a partir de la ENEC y otros métodos complementarios.

Las personas incorporadas a partir de registros administrativos cuentan únicamente con información básica a nivel individual como edad, sexo y ubicación geográfica, la cual, como se verá más adelante, presenta limitaciones en niveles territoriales finos, tales como localidades del interior y zonas dentro de Montevideo, debido a sesgos asociados al *algoritmo* utilizado para su asignación. Asimismo, no cuentan con información sobre las variables centrales del censo, tales como nivel educativo, condiciones de la vivienda y características del hogar. Adicionalmente, estas personas no se

encuentran integradas en hogares, es decir, no se reconstruyen hogares a partir de los registros administrativos, salvo en el caso de los niños y niñas de 15 años o menos, quienes fueron incorporados a hogares efectivamente censados, utilizando información proveniente del Certificado de Nacido Vivo (CNV).

En términos de magnitud, aproximadamente 360 mil personas fueron incorporadas a partir de registros administrativos, en línea con una omisión del orden del 10.3%. Sin embargo, la información disponible para estas personas es extremadamente limitada. Si bien este procedimiento permite ajustar la distribución por edad y sexo, no aporta información para el análisis de las variables clave del censo ni permite caracterizar adecuadamente a la población en términos socioeconómicos. En consecuencia, no es posible corregir el sesgo asociado a la omisión diferencial ni producir estimaciones confiables para aquellas variables que solo están disponibles para los hogares y personas efectivamente censadas.

Por otra parte, la omisión censal se concentra en segmentos específicos de la población y por lo tanto, el uso directo de los hogares y personas efectivamente censados, sin aplicar correcciones o ajustes adicionales, conduce a estimaciones sesgadas, representativas de una población con mejores condiciones socioeconómicas que las existentes en la realidad (Bethlehem, J., 2002). Este problema no puede resolverse únicamente mediante la imputación de información ni a través de la incorporación de personas provenientes de registros administrativos, ya que dichas fuentes también presentan sesgos estructurales, asociados a la mayor visibilidad administrativa de determinados grupos poblacionales (Zhang, L. C. , 2012).

Lo anterior vuelve necesario ajustar los datos censales mediante un sistema de ponderadores o pesos que permita corregir, en la medida de lo posible, el sesgo introducido por la omisión diferencial y recuperar la representatividad de la población total en las estimaciones de las variables de interés.

Es importante tener en cuenta que tanto la incorporación de personas a partir de registros administrativos como la estrategia de ponderación adoptada constituyen soluciones implementadas con posterioridad al operativo censal, en respuesta a los niveles y patrones de omisión observados. En particular, el uso de registros administrativos para la incorporación de personas no formó parte del diseño metodológico del censo, sino que surge como una estrategia de corrección orientada a reducir parcialmente el impacto de la omisión en los conteos poblacionales. Sin embargo, dicha estrategia no logra mitigar el problema de fondo asociado a la omisión diferencial, en la medida en que no incorpora información sobre las variables centrales del censo ni permite corregir los sesgos en la caracterización socioeconómica de la población.

De todas formas, la información derivada del censo combinado constituye un insumo relevante para el proceso de ponderación, en particular a través de los totales poblacionales por edad y sexo a nivel departamental. Sin embargo, su utilización se restringe a niveles de agregación en los que se considera que la calidad de la información es adecuada, dado que el INE identificó limitaciones y errores en la asignación geográfica y en algunas fuentes administrativas, especialmente a niveles territoriales más desagregados. En contraste, algunos países han integrado este tipo de herramientas desde la etapa de diseño metodológico del censo. Por ejemplo, en el Reino Unido se utilizan encuestas de evaluación censal de gran tamaño para estimar la omisión con un alto nivel de desagregación territorial, lo que permite ajustar las estimaciones mediante ponderación bajo un enfoque planificado (Office for National Statistics, 2021). Por su parte, en Nueva Zelanda se ha avanzado en el uso de registros administrativos no solo para la incorporación de personas, sino también para la reconstrucción de hogares y la agregación de múltiples variables, ampliando significativamente su potencial analítico.

Estas diferencias reflejan distintos enfoques metodológicos en la integración de fuentes de información, condicionados tanto por la disponibilidad de registros administrativos como por las decisiones de diseño adoptadas en cada contexto.

Previo a la adopción de la metodología final, el INE como lo ha hecho tradicionalmente durante toda su historia presentó y discutió la estrategia propuesta con distintos actores relevantes, incluyendo la academia, usuarios especializados de la información censal y organismos internacionales. En dichas instancias se presentaron y analizaron tanto los fundamentos conceptuales del enfoque como resultados empíricos preliminares, incluyendo la evaluación de estrategias tradicionales y alternativas para el tratamiento de la omisión.

Como resultado de este proceso, se desarrolló una estrategia específica de ponderación de los hogares y personas efectivamente censadas, con el objetivo de producir estimaciones con el menor sesgo posible para las variables centrales del censo. Para ello, el conjunto de hogares censados fue tratado conceptualmente como una *muestra no probabilística*, en la que el mecanismo de selección (i.e. la probabilidad de haber sido censado) es desconocido y depende de múltiples factores, tanto observables como no observables (Elliott & Valliant, 2017 & Chen, Li & Wu, 2020).

La metodología adoptada se basa en el uso de estimadores doblemente robustos, que combinan modelos para estimar la probabilidad de participación en el censo con modelos superpoblacionales (Chambers & Clark, 2012), i.e., modelos estadísticos que permiten predecir las variables de interés a partir de información auxiliar disponible, utilizando variables correlacionadas con dichas variables (como edad, sexo y región). En este esquema, si bien la base final de microdatos no incorpora directamente a las personas provenientes de registros administrativos, la información agregada derivada del censo combinado constituye un insumo clave para la calibración de los ponderadores. En particular, los totales poblacionales por sexo y edad a nivel departamental, obtenidos a partir de la integración del operativo censal con los registros administrativos, actúan como referencias externas para corregir la omisión diferencial con mayor precisión territorial, sin trasladar la incertidumbre asociada a la asignación geográfica individual a la base de datos final (Deville & Särndal, 1992).

## 1.1. Evaluación técnica de la cobertura y alcances del enfoque combinado

Es importante precisar que el sistema de ponderación desarrollado en este documento no sustituye los avances logrados con el enfoque de censo combinado (INE, 2024e), sino que los integra como el insumo fundamental para la etapa de ponderación. No obstante, en un ejercicio de transparencia metodológica, resulta fundamental precisar que la reducción del sesgo destacada en trabajos recientes sobre el Censo 2023 (Aboal & Segui, 2025) presenta inconsistencias conceptuales desde la perspectiva de la posibilidad de realizar inferencias válidas. En primer lugar, la afirmación de que el enfoque del censo combinado elimina el sesgo resulta metodológicamente insuficiente; si bien mejora la cobertura demográfica (i.e. el conteo), no resuelve la ausencia de información en las cerca de 100 variables sustantivas del censo para las personas incorporadas administrativamente. Como advierten Särndal y Lundström (2005), el conocimiento de los totales poblacionales no mitiga el sesgo de selección en las variables de interés (e.g. educación, empleo, condiciones de vida, etc.) si la propensión de respuesta está correlacionada con dichas variables.

De hecho, la sugerencia de resolver esta carencia de información mediante técnicas de imputación múltiple o esquemas de ponderación a medida (Aboal & Segui, 2025) entra en contradicción con el hito de "imputación cero" resaltado por los mismos autores. Imputar la totalidad de las variables socioeconómicas para las personas introducidas administrativamente no solo reintroduce la incertidumbre y la generación de "p*ersonas sintéticas*"[1] que se pretendía evitar, sino que, como advierten Little y Rubin (2019) y Reiter (2007), resulta inviable para preservar las complejas correlaciones y la coherencia lógica de los microdatos en una base de gran escala como es la del censo. Un proceso de este tipo introduce un sesgo de modelo incontrolable que amenaza la integridad de los datos. Asimismo, el uso de ponderadores específicos por variable fragmentaría la base de datos, rompiendo la coherencia interna que exige la relación técnica entre niveles de hogar y persona.

En segundo lugar, existe una contradicción técnica respecto a la precisión geográfica del algoritmo de asignación descrito por dichos autores. A diferencia de la visión de Aboal & Segui (2025), el análisis realizado en el INE identifica un sesgo de urbanicidad persistente en el algoritmo de asignación de personas vía registros administrativos, el cual tendió a subrepresentar sistemáticamente a las localidades pequeñas y áreas rurales. Este algoritmo presenta, además, un sesgo de visibilidad institucional vinculado al nivel socioeconómico: al supeditar la inclusión de personas a la existencia de "señales de vida" administrativas (e.g. empleo formal, salud, educación), el sistema tiende a ser más preciso en estratos medios y altos, pero subestima u omite sistemáticamente a la población en áreas de pobreza crítica y alta informalidad, donde las huellas o señales administrativas son escasas. Este fallo estructural en la distribución territorial obligó a la aplicación de ajustes adicionales para garantizar la verosimilitud de los datos a nivel local. Por más detalle ver 2.2. Registros administrativos y sesgo de visibilidad institucional.

A las limitaciones mencionadas se suma una falencia estructural crítica: la incapacidad del algoritmo para reconstruir la unidad básica de análisis censal (i.e el hogar) en la mayoría de los registros administrativos. Esta desarticulación rompe la estructura de hogar para aproximadamente el 10% de la población, un fenómeno identificado en la literatura como una falla de vínculo (*linkage failure*) entre los registros de personas y de viviendas (Wallgren & Wallgren, 2014). Como se observó en experiencias internacionales similares (Stats NZ, 2019), esto genera individuos que solo pueden ser contabilizados a nivel de área, perdiendo la capacidad de representar la vivienda e impidiendo el cálculo de indicadores fundamentales como ser, por ejemplo, el hacinamiento.

En efecto, durante el proceso de revisión y construcción del sistema de ponderadores, se identificaron empíricamente estas inconsistencias, vinculadas a coberturas desiguales y a limitaciones en las *señales de vida* disponibles para ciertos grupos poblacionales. Asimismo, se detectaron y corrigieron errores de asignación geográfica en viviendas colectivas de gran escala antes de la generación final de los microdatos. Por lo tanto, la construcción de un sistema de ponderador único es la única vía para dotar al censo de consistencia y reducir los sesgos. Esta etapa es indispensable para que las unidades efectivamente observadas (las únicas con información socioeconómica completa y estructura de hogar definida) representen fielmente la heterogeneidad de la población total, corrigiendo tanto el sesgo de selección como las distorsiones territoriales, el sesgo de visibilidad institucional y la pérdida de estructura habitacional detectadas en la asignación administrativa.

[1] Siguiendo a Aboal & Segui, (2025), el término “personas sintéticas” refiere a registros generados íntegramente mediante procedimientos de imputación estadística para suplir la no respuesta en el campo (método utilizado en el censo 2011).

## 1.2. Estructura del documento

El presente documento describe en detalle la metodología de ponderación aplicada al Censo 2023, sus fundamentos conceptuales, los supuestos adoptados, las fuentes de información utilizadas y las principales limitaciones del enfoque, con el objetivo de proporcionar a los usuarios de los microdatos censales un marco claro y transparente para la correcta interpretación de las estimaciones producidas a partir de esta base.

El documento se estructura de la siguiente manera. En la Sección 2 se presenta el marco conceptual de la omisión censal, incluyendo su relación con la no respuesta, el rol de los registros administrativos y las limitaciones de los enfoques tradicionales. La Sección 3 aborda el diagnóstico empírico de la omisión, con especial énfasis en su estimación a nivel de segmento censal. En la Sección 4 se desarrolla la estrategia de corrección adoptada, describiendo en detalle la construcción de los ponderadores utilizando estimadores doblemente robustos, la modelización de la probabilidad de respuesta, la calibración a totales poblacionales y los principales controles de calidad implementados. La Sección 5 presenta la evaluación del impacto de la ponderación en los indicadores censales, mientras que la Sección 6 discute las medidas de calidad de las estimaciones. En, la Sección 7 expone los supuestos y limitaciones del enfoque adoptado; y en la Sección 8 se presenta una guía breve del uso de microdatos utilizando el sistema de ponderadores.

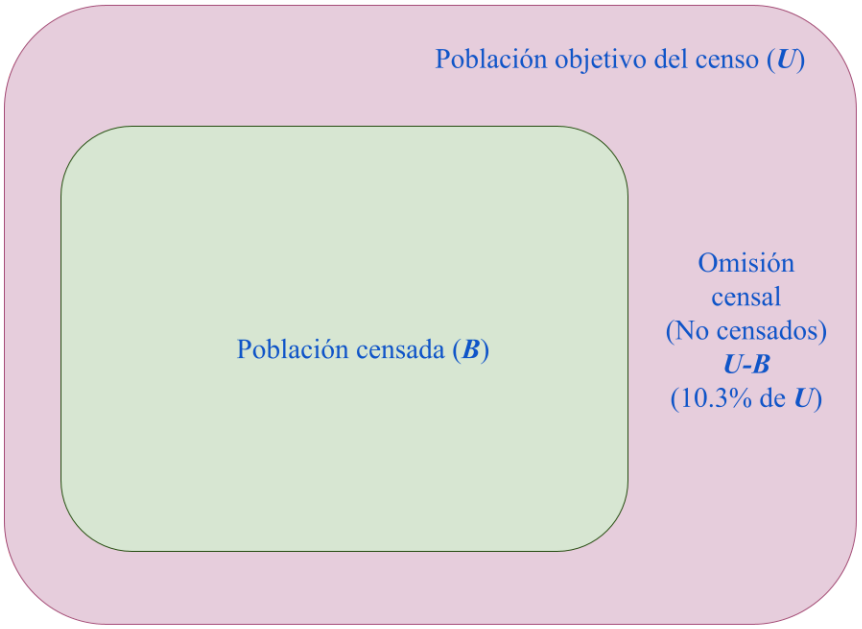
Esta página ha sido dejada intencionalmente en blanco

# 2. Omisión censal

## 2.1. Omisión censal, tipos de omisión y relación con la no respuesta

La omisión censal refiere a la situación en la que personas, hogares o viviendas pertenecientes al universo no son efectivamente relevados durante el operativo de campo. Este fenómeno puede deberse a múltiples factores, entre los que se incluyen dificultades de acceso territorial, barreras tecnológicas, problemas operativos y decisiones individuales que toman los hogares (y sus integrantes) de no participar en el censo (Groves, et al., 2002).

**Figura 1: Diagrama Conceptual de la omisión censal**




En la práctica, la omisión no es un fenómeno homogéneo, sino que comprende distintos tipos de situaciones, entre las que se destacan:

- *Moradores ausentes*, correspondientes a viviendas identificadas durante el operativo censal, pero en las que no se logró contactar a sus ocupantes dentro del período de relevamiento del censo.

- *Omisión propiamente dicha*, que refiere a personas, hogares o viviendas que no fueron identificados ni relevados durante el operativo, y cuya existencia sólo puede ser inferida a partir de fuentes externas o instrumentos de evaluación censal (United Nations, 2010).

- *Omisión de integrantes dentro de hogares censados,* cuando el hogar es efectivamente censado, pero no se registra a la totalidad de las personas que lo integran.

Bajo el enfoque de las inferencias por muestreo en poblaciones finitas, estas situaciones mencionadas anteriormente pueden interpretarse como distintas fuentes de no respuesta, tanto a nivel de la unidad primaria (hogares o viviendas) como a nivel de unidades secundarias (personas). En este sentido, la omisión censal puede asimilarse a un proceso de selección incompleto, en el que no todas las unidades del universo tienen la misma probabilidad de ser observadas (i.e. censadas). La omisión censal genera problemas cuando no ocurre de manera completamente aleatoria (i.e. todos los hogares tienen la misma probabilidad de ser censados), sino que se concentra en determinados grupos poblacionales. En estos casos, la omisión genera un sesgo de cobertura, dado que las unidades observadas no constituyen una representación adecuada del total de la población (Bethlehem, et al., 2011).

Bajo un enfoque determinístico, el sesgo de la media de los censados $\bar{Y}_B$ puede expresarse como el producto de la tasa de omisión $TO$ y la diferencia entre la $\bar{Y}_B$ y la media de la población omitida $\bar{Y}_O$:

$$\text{Sesgo}(\bar{Y}_B) = TO \times (\bar{Y}_B - \bar{Y}_O)$$

Esta relación muestra que el sesgo no solo depende de la magnitud de la omisión, sino fundamentalmente de la naturaleza diferencial de la misma. Complementariamente, bajo un enfoque estocástico de respuesta, el sesgo puede expresarse en función de la covarianza entre la probabilidad de ser censado $\phi$ y la variable de interés $y$:

$$\text{Sesgo}(\bar{Y}_B) \approx \frac{cov(\phi, y)}{\bar{\phi}}$$

donde $\bar{\phi}$ es la media de las probabilidades de responder en la población.

Ambas expresiones demuestran que, dado que en el Censo 2023 la omisión se concentró en estratos de mayor vulnerabilidad socioeconómica, la diferencia $(\bar{Y}_B - \bar{Y}_O)$ es significativa y la covarianza es distinta de cero, lo que invalida el uso directo de los datos sin ajustes adicionales.

En consecuencia, las estimaciones obtenidas a partir de los datos censales (i.e. los hogares efectivamente censados) sin ajustes adicionales pueden resultar en estimaciones sesgadas para ciertas variables de interés. A su vez, actualmente no es posible obtener estimaciones para la mayoría de los indicadores censales en términos absolutos (e.g. totales o conteos), debido al nivel de omisión observado.

En particular, el Censo 2023 registró una menor cantidad de personas que el censo anterior, el cual había contabilizado 3.252.091 personas con una omisión estimada del 4,07%, considerablemente menor a la observada en 2023, lo que vuelve inviable la estimación de los cambios netos en valores absolutos entre un censo y otro.

Es importante tener en cuenta que tanto los niveles absolutos de distintos indicadores (e.g. carencias de los hogares) como la estimación de los cambios netos intercensales constituyen insumos fundamentales para el diseño, monitoreo y evaluación de políticas públicas. Este problema no puede resolverse mediante el enfoque de censo combinado, sino que requiere una estrategia de ponderación aplicada a los hogares efectivamente censados.

Este marco conceptual resulta clave para comprender que el problema de la omisión censal no se limita únicamente a una cuestión de conteo, sino que afecta también directamente la calidad y validez de las estimaciones derivadas del censo, en particular cuando se analizan características socioeconómicas y demográficas que presentan una distribución diferencial entre la población censada y la población omitida.

## 2.2. Registros administrativos y sesgo de visibilidad institucional

Los registros administrativos (RRAA) constituyen una fuente de información valiosa para complementar los censos de hogares y personas, en particular para mejorar la estimación de los totales poblacionales. La utilización de los mismos puede permitir identificar personas que no fueron censadas directamente, contribuyendo a reducir exclusivamente la subenumeración. No obstante, la información disponible para estas personas se limita a variables básicas a nivel individual, fundamentalmente sexo, edad y una referencia de ubicación geográfica, sin incluir información sobre las variables centrales del censo, tales como educación, condiciones de la vivienda y características del hogar. En consecuencia, la inclusión de personas agregadas por medio de RRAA no tiene impacto en la reducción del sesgo en las estimaciones de los indicadores del censo, ya que no permite corregir la omisión diferencial en las variables centrales, limitándose a mejorar los conteos poblacionales.

Sin embargo, los RRAA no son “puros” desde el punto de vista estadístico. Su cobertura está generalmente condicionada por la interacción de las personas con el estado y otros sistemas formales, lo que da lugar a un “sesgo de visibilidad administrativa”. Tras revisar la documentación técnica interna de los modelos de aprendizaje automático (*Random Forest*) y de elección discreta (*Logit*) utilizados para la selección y localización de la población residente, se identifican falencias estructurales que comprometen la validez de las inferencias socioeconómicas. El problema central radica en que el modelo intenta predecir la residencia real utilizando como verdad absoluta (*target*) un censo que presenta sesgos de cobertura no aleatorios. Las falencias más críticas se detallan a continuación:

1. **Circularidad metodológica y retroalimentación del sesgo:** El algoritmo fue entrenado utilizando como variable dependiente la condición de "censado". Dado que el operativo de campo presentó una omisión concentrada en áreas de alta vulnerabilidad, el modelo "aprende" que esas características son predictoras de no residencia. En lugar de actuar como un corrector, el algoritmo institucionaliza el sesgo de campo (Hand, 2006), clasificando como no residentes a las personas que el operativo no logró contactar.

2. **El sesgo de las "señales de vida" como indicador de formalidad:** La predicción de residencia depende de la "huella administrativa" (BPS, Salud, Educación, UTE). Al condicionar la validación de un residente a su integración institucional, el modelo genera un *sesgo de exclusión sistemático*: los individuos en situación de precariedad e informalidad laboral son propensos a ser clasificados como "falsos negativos" debido a la ausencia de atributos en los registros formales (Wallgren & Wallgren, 2014).

3. **Redundancia de variables y penalización de la informalidad:** El diseño del modelo incluye tanto variables *dummy* por registro como una variable sintética que consiste en la suma de dichas señales de vida. Esta redundancia metodológica genera una multicolinealidad que sobrepondera la formalidad y penaliza doblemente a los sectores vulnerables. Si una persona carece de registros, el algoritmo recibe un impacto negativo multiplicado, reforzando el sesgo de visibilidad institucional por diseño.

4. **Inestabilidad en el entrenamiento y falta de calibración:** El entrenamiento con clases no balanceadas (donde la clase mayoritaria de "censados" domina sobre la omisión) y la ausencia de probabilidades calibradas (Kuhn & Johnson, 2013) provocan que el algoritmo sea incapaz de identificar correctamente a la población omitida. El uso de un umbral de probabilidad global para ajustar el total a la ENEC constituye una solución puramente numérica que no garantiza la representatividad sustantiva ni corrige la distorsión en la composición socioeconómica de la población resultante.

5. **Ruptura de la unidad de hogar y "parches" determinísticos:** El algoritmo procesa registros individuales, perdiendo la naturaleza relacional del censo. La posterior aplicación de reglas de excepción manuales para reconstruir núcleos familiares es una admisión de la incapacidad del modelo para lidiar con la falla de vínculo (*linkage failure*), rompiendo la consistencia del diseño original.

6. **Sesgo geográfico y distorsión urbana:** La localización depende de "anclas" administrativas. En áreas de urbanización irregular o ruralidad dispersa, donde estas anclas son débiles, el modelo tiende a asignar a los individuos a centros urbanos más densos. Esto explica el sesgo de urbanicidad detectado, donde el algoritmo "vacía" artificialmente la periferia o áreas dispersas, para inflar los conteos en zonas formales o con mayor población en el caso del interior del país.

Si bien los RRAA resultan adecuados para corregir desbalances en los conteos agregados de población (sexo, edad y departamento), no permiten, por sí solos, corregir los sesgos asociados a la omisión, ni sustituir la información relevada del censo para variables que no se encuentran disponibles en dichas fuentes. Esta limitación es particularmente relevante para las cerca de 100 variables centrales del censo. Este algoritmo debe entenderse como una herramienta de clasificación de registros y no como un modelo de inferencia poblacional. En consecuencia, el sistema de ponderación único es la única vía metodológicamente válida para devolverle la representatividad a la población que los mecanismos administrativos puros no logran capturar.

## 2.3. El censo como muestra no probabilística

En presencia de omisión censal heterogénea (i.e. tasa de omisión/no respuesta distintas), el conjunto de hogares y personas efectivamente censados (i.e. “los contados”) puede ser interpretado y por ende tratado, como una *muestra no probabilística*, en donde, el mecanismo de selección de las unidades observadas (hogares/personas) no responde a un diseño muestral (estrategia de selección) conocido,

sino a un proceso de generación de respuesta complejo y desconocido (Meng, X. L. ,2018), en el que intervienen factores observables y no observables. La probabilidad/propensión/voluntarismo de que un hogar o una persona haya sido censada depende tanto de características propias de los hogares ( y sus integrantes) así como de condiciones externas, como ser, aspectos operativos del relevamiento censal. Dado que este mecanismo de selección es desconocido, las estrategias tradicionales de inferencia basada en el diseño muestral resultan insuficientes para garantizar una muestra "representativa" que sea capaz de brindar estimaciones "insesgadas" (Elliott, M. R., & Valliant, R., 2017). Este enfoque abre un camino el uso de metodologías desarrolladas originalmente para el análisis de datos provenientes de fuentes no probabilísticas, en donde se modela de manera implícita el proceso de selección de los hogares y se incorporan variables auxiliares con el objetivo de reducir el sesgo y por ende mejorar la calidad de las estimaciones producidas por el censo.

## 2.4. Límites de los enfoques tradicionales

En la práctica, tradicionalmente, la omisión, y en particular la asociada a moradores ausentes, ha sido abordada mediante procedimientos de imputación, usualmente basados en la asignación de donantes seleccionados de forma aleatoria (*Hot deck*) entre las unidades efectivamente censadas que cumplían ciertas características (Fellegi, I. P., & Holt, D. , 1976). En este enfoque, cada unidad con información faltante recibe los valores observados de una unidad "donante", seleccionada dentro de un conjunto de unidades similares (denominado pool de donantes), definido a partir de variables auxiliares disponibles. En general, los donantes se seleccionan entre los "vecinos" más cercanos o entre hogares similares dentro de la misma área geográfica, priorizando aquellos que también presentaban mayores dificultades de respuesta, con el objetivo de conformar un pool de donantes más comparable a los hogares omitidos. Dependiendo de la implementación, la selección puede realizarse con o sin reemplazo, y puede incorporar distintos criterios de estratificación o emparejamiento (matching), lo que introduce variabilidad adicional en los resultados. Este tipo de estrategias ha sido ampliamente utilizado con el objetivo de completar la información faltante y asegurar la coherencia interna de las bases de datos censales. Sin embargo, desde un punto de vista estadístico, estos procedimientos no corrigen el sesgo de cobertura cuando la omisión no ocurre de manera aleatoria. En la medida en que los hogares y personas omitidas presentan, en promedio, características distintas a las de los hogares y personas efectivamente censadas, los donantes utilizados en los procesos de imputación tienden a reproducir el comportamiento de la población observada y no el de la población omitida. En particular, el método asume implícitamente que, condicionado a las variables utilizadas para definir el pool de donantes, las unidades observadas son representativas de las no observadas, supuesto que no se cumple en presencia de omisión diferencial. Como resultado, el método *Hot Deck* no garantiza reducir el sesgo (al menos de forma significativa) para las variables de interés del censo.

En el marco del Censo 2023, el INE recorrió también estos caminos metodológicos tradicionales para la reducción del sesgo producto de la omisión, evaluando el impacto de la imputación de moradores ausentes sobre las principales estimaciones censales. Los resultados obtenidos mostraron que las estimaciones construidas a partir de los hogares efectivamente censados son, en términos generales, similares a aquellas que incorporan hogares censados más imputación de moradores ausentes, lo que refuerza la evidencia de que estos procedimientos, al menos para el caso del censo de Uruguay, no logran corregir o reducir el sesgo asociado a la omisión. Este diagnóstico llevó a adoptar un enfoque alternativo, en el que el conjunto de hogares y personas censados es tratado conceptualmente como una muestra no probabilística, reconociendo que el mecanismo de participación en el censo es desconocido y depende de múltiples factores observables y no observables. A partir de este enfoque,

se incorporan metodologías desarrolladas para la inferencia a partir de muestras no probabilísticas, con el objetivo de reducir el sesgo de las estimaciones y preservar la utilidad analítica de los datos censales en un contexto de tasas de respuesta decrecientes.

La incorporación de nuevas metodologías para mejorar la calidad de los productos estadísticos (e.g. censos, encuestas e indicadores derivados) forma parte de las buenas prácticas internacionales en la producción estadística oficial. Las oficinas nacionales de estadística revisan y actualizan de manera continua sus métodos, supuestos y procedimientos de estimación a medida que cambian los patrones de respuesta, las tecnologías y nuevas metodologías disponibles entre otras cosas. El INE no es ajeno a este proceso de mejora continua y, por lo tanto, la adopción de nuevas estrategias de ponderación y ajuste constituye una respuesta metodológica necesaria frente a los desafíos actuales de cobertura y omisión censal

## 2.5. Resumen del marco conceptual

El marco conceptual presentado deja en evidencia que la omisión censal constituye un fenómeno complejo, con múltiples manifestaciones y con implicancias directas sobre la calidad de las estimaciones derivadas del censo. Cuando la omisión no ocurre de forma aleatoria (en la práctica nunca lo es) y se concentra en determinados grupos poblacionales, tales como hogares en contextos de mayor vulnerabilidad socioeconómica, personas jóvenes y áreas con menor cobertura efectiva del operativo censal, incluyendo zonas rurales, el uso directo de los datos censales sin ningún tipo de ajustes (e.g. pesos distintos a los hogares) conduce por definición a estimaciones sesgadas.

Si bien los enfoques tradicionales basados en la imputación de moradores ausentes permiten “completar” parte de la información faltante y asegurar la coherencia interna de las bases de datos, estos procedimientos no resultan suficientes para corregir los sesgos derivados de la omisión diferencial. En particular, los métodos de imputación tipo *Hot Deck* asignan información a partir de unidades observadas, seleccionadas dentro de un conjunto de donantes con características similares, lo que implica que las unidades imputadas reproducen la distribución de la población efectivamente censada. Dado que esta población difiere sistemáticamente de la población omitida, estos procedimientos no siempre permiten recuperar las características de los grupos subrepresentados, ni corregir el sesgo en las variables de interés.

De igual forma, la estrategia previamente utilizada por el INE basada en la incorporación de personas a partir de registros administrativos, si bien contribuye a mejorar la estimación de los conteos poblacionales (e.g. cantidad de personas por edades simples), tiene un efecto nulo en la reducción del sesgo en aquellas variables censales que no se encuentran disponibles en dichas fuentes, ya que no incorpora información sobre las variables centrales del censo ni permite corregir la omisión diferencial en dichas dimensiones. Esto se debe a que las personas incorporadas mediante registros administrativos cuentan únicamente con información básica, fundamentalmente edad, sexo y una referencia geográfica, sin información sobre las variables centrales del censo, y no se encuentran integradas en hogares censados, lo que impide su utilización para caracterizar adecuadamente la estructura socioeconómica de la población y, por ende, corregir los sesgos derivados de la omisión diferencial.

En este contexto, el tratamiento del censo como una muestra no probabilística y la adopción de metodologías desarrolladas para la inferencia a partir de fuentes no probabilísticas constituyen una

respuesta metodológica adecuada e innovadora frente a los desafíos que enfrentan los censos modernos. Este enfoque también es posible debido a los importantes avances metodológicos que se han producido durante los últimos diez años en materia de inferencia a partir de muestras no probabilísticas, estimadores doblemente robustos, modelos de respuesta, técnicas de calibración, entre otros.

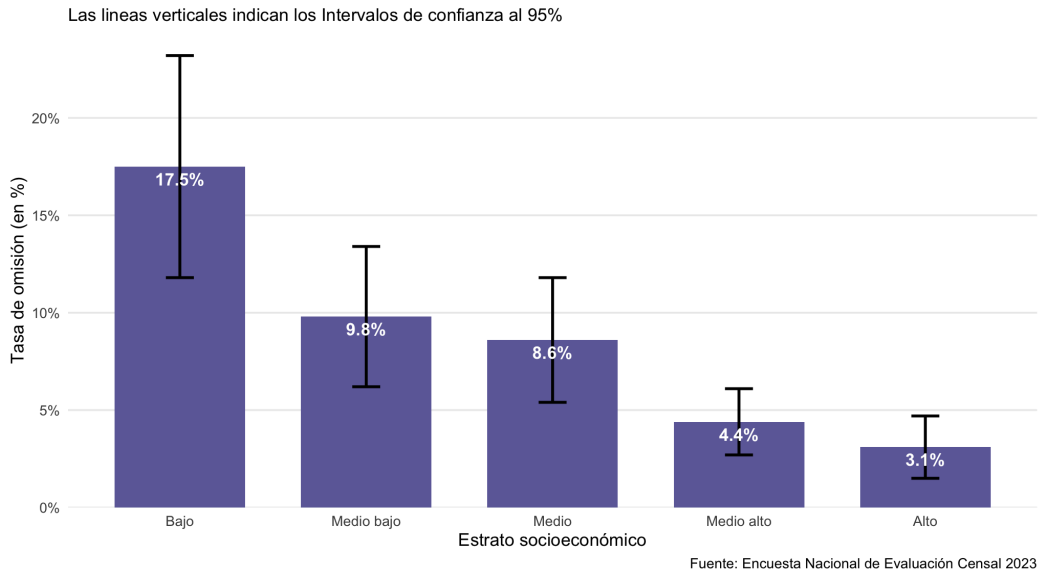
Si bien el Censo 2011 también presentó niveles de omisión, estimados en 4,07% y como ocurre habitualmente esta no fue homogénea entre distintos grupos poblacionales y territorios, su magnitud no era suficientemente elevada como para comprometer de manera importante la utilidad de las estimaciones censales. De todos modos, hubiera sido de interés contar con un sistema de ponderadores que permitiera corregir parcialmente esos sesgos. Sin embargo, en ese momento no existía ni el desarrollo metodológico actual ni la disponibilidad de información auxiliar suficiente para construir este tipo de ponderadores de manera robusta. En cambio, el Censo 2023 dispone de mayores fuentes de información complementaria que permiten implementar estrategias de ajuste y ponderación considerablemente más sofisticadas.

# 3. Diagnóstico empírico de la omisión censal

El proceso de ponderación desarrollado para el Censo 2023 se aplica exclusivamente sobre los hogares efectivamente contados y sus integrantes. El ajuste por no respuesta se implementa a nivel de segmento censal (*conjunto de manzanas en el área urbana y territorios en el área rural*) y se interpreta en un sentido amplio, ya que busca compensar simultáneamente distintas manifestaciones de la omisión censal: los hogares con moradores ausentes, las omisiones propiamente dichas (hogares o viviendas que no llegaron a ser censados) y los casos en los que, aun habiéndose censado al hogar, se omitió la declaración de alguna persona. Desde la perspectiva de la teoría del muestreo en poblaciones finitas, estas situaciones se modelan como un problema de no respuesta, en el que los hogares y personas efectivamente observados no constituyen un subconjunto aleatorio del universo poblacional. En este marco, los factores de ajuste por no respuesta redistribuyen el peso de los hogares observados dentro de cada segmento censal, de modo que estos representen tanto a los hogares (y sus integrantes) efectivamente censados como a aquellas que quedaron fuera del censo (i.e. omitidas) por cualquiera de las situaciones mencionadas.

La principal fuente de información para el diagnóstico empírico de la omisión censal es la Encuesta Nacional de Evaluación Censal (ENEC) 2023, instrumento post-censal diseñado específicamente para estimar la cobertura del operativo censal y caracterizar los patrones de omisión/no respuesta. La ENEC permite obtener estimaciones confiables de la omisión para dominios amplios de población, así como analizar su comportamiento según territorio, nivel socioeconómico, edad y sexo. Estos resultados constituyen la base empírica fundamental para justificar y diseñar los ajustes por no respuesta aplicados al censo. De acuerdo con los resultados de la ENEC, la omisión censal en Montevideo, donde reside aproximadamente el 40% de la población del país, se concentró con mayor intensidad en las zonas geográficas en las que residen las poblaciones más vulnerables. En particular, se observa un claro gradiente socioeconómico en la omisión: los estratos bajos presentan los niveles más elevados, mientras que la omisión disminuye sistemáticamente a medida que aumenta el nivel socioeconómico del área.

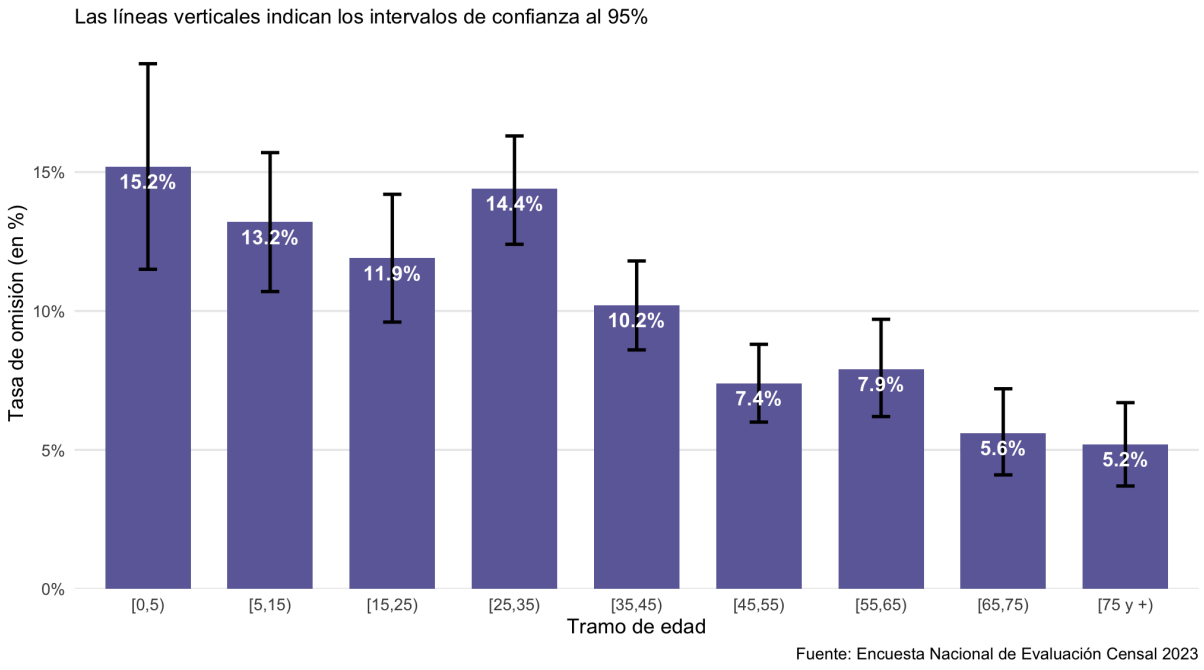
**Figura 2: Estimación de la omisión según estrato socioeconómico para Montevideo**



Los estratos socioeconómicos[2] más bajos presentan tasas de omisión del orden del 18% mientras que en los estratos altos estos valores descienden hasta aproximadamente 3%. Este patrón resulta consistente con la distribución territorial de la pobreza en Montevideo, ya que las áreas con mayor incidencia de pobreza coinciden con aquellas donde la ENEC estimó las tasas más elevadas de omisión. Por lo tanto, la evidencia territorial refuerza la conclusión de que la omisión censal tuvo una marcada dimensión socioeconómica, y que los hogares ubicados en contextos de mayor vulnerabilidad presentaron una probabilidad o propensión mucho menor de ser efectivamente censados.

A nivel del total del país, la ENEC permite mostrar que la omisión censal presenta también un componente diferencial por edad y sexo. En la siguiente figura se puede observar una subrepresentación de personas jóvenes, especialmente en tramos etarios caracterizados por una mayor participación en el mercado laboral, entre otras características. Las tasas de omisión se sitúan en torno al 14–15% en los tramos de 25 a 35 años, mientras que disminuyen progresivamente a medida que aumenta la edad, alcanzando valores cercanos al 5–8% en los tramos de 55 años o más. Esta estructura implica que la población efectivamente censada no reproduce de manera equilibrada la pirámide poblacional (esperada) del país, sobrerrepresentando a personas de mayor edad y sub representando a los grupos jóvenes.

**Figura 3: Estimación de la omisión según tramo de edad**




La evidencia presentada (la cual se ampliará en el resto del documento) muestra que la omisión censal observada no fue aleatoria, sino que se concentró (entre otras cosas) en determinados grupos poblacionales, territorios y áreas con distintos niveles socioeconómicos. En este contexto, si se

[2] Los estratos socioeconómicos son construidos utilizando técnicas de aprendizaje no supervisado (clustering y más precisamente k-means) en donde la información se encuentra definida a nivel de segmento censal y se utilizan como variables de entrada para el algoritmo, el ingreso promedio del hogar, la proporción de ocupados, desocupados e inactivos. Dicha información proviene de hacer un pool de varios años de la Encuestas Continua de Hogares (ECH).

trabajara exclusivamente con la población efectivamente censada, sin aplicar ajustes por omisión o no respuesta, las estimaciones resultantes estarían sistemáticamente sesgadas.

Por lo tanto, la población censada no constituye un subconjunto aleatorio de la población total, sino que está sobrerrepresentada por hogares con menores niveles de vulnerabilidad socioeconómica y por personas de mayor edad, mientras que los sectores más pobres y los grupos jóvenes quedan subrepresentados. Como consecuencia, los indicadores sociales y demográficos derivados directamente de los datos censales sin ajustes tenderían a describir una población "más envejecida" y con mejores condiciones de vida que las existentes en la realidad.

## 3.1. Estimación de la omisión a nivel de segmento censal

Si bien la ENEC permite estimar la omisión para dominios amplios de población, no cuenta con un tamaño muestral suficiente para producir estimaciones precisas en niveles territoriales muy desagregados, como el segmento censal. Las encuestas de cobertura pueden utilizarse para estimar las propensiones de los hogares a responder el censo, pero para que ello sea posible, las Oficinas Nacionales de Estadística (ONEs) deben diseñar la encuesta con ese propósito (por ejemplo, aumentando el tamaño de muestra para obtener estimaciones más desagregadas de la omisión).

En el caso del INE, la ENEC no fue concebida con ese objetivo, sino para obtener estimaciones confiables de la omisión (tasa de no respuesta) a nivel nacional y para dominios amplios de la población. Por este motivo, se adopta un *modelo predictivo* que permite estimar tasas de respuesta a un mayor nivel de desagregación territorial (e.g. a nivel de segmento censal). A partir de este punto, dicho enfoque será denominado *Modelo de Respuesta Censal Basado en la Tasa de Enlace*[3] (*MRTE*).

El *MRTE* se basa en la naturaleza del operativo censal, que combinó modalidades de recolección mixta, CAWI y CAPI. Independientemente de que el hogar respondiera inicialmente vía CAWI, debía ser contactado presencialmente por un censista para verificar la dirección y enlazar el cuestionario con la vivienda correspondiente. En este esquema, la probabilidad de respuesta está determinada principalmente por la posibilidad de establecer contacto efectivo con el hogar. Este planteamiento se fundamenta en la teoría clásica de la participación en encuestas de Groves y Couper (1998), en donde la probabilidad de respuesta $P(R)$ se puede expresar como producto de dos componentes:

$$P(R) = P(C) \times P(K|C),$$

donde $P(C)$ es la probabilidad de establecer contacto con el hogar y $P(K|C)$ es la probabilidad de cooperación del hogar una vez contactado.

El supuesto central del *MRTE* es que, en el contexto del censo, la tasa de enlace de los hogares CAWI constituye un *proxy* adecuado de la tasa de respuesta de los hogares relevados exclusivamente de forma presencial. Esta aproximación se basa en que el principal determinante de la participación no es el rechazo del hogar, sino la capacidad de establecer contacto.

[3] El Modelo *MRTE* fue desarrollado por el consultor del INE Marcelo Perera y fue concebido originalmente como un mecanismo adicional, complementario a la ENEC, para la estimación de la omisión censal.

La evidencia operativa y estudios sobre no respuesta en censos (Bates & Mulry, 2011) indican que la tasa de rechazo es relativamente baja debido a la obligatoriedad y relevancia del operativo. Por lo tanto, se asume que la cooperación es casi total, i.e.,

$$P(K|C) \approx 1,$$

lo que permite que la probabilidad de respuesta sea explicada esencialmente (o en una gran parte) por el mecanismo de contacto, i.e.,

$$P(R) \approx P(C).$$

Una vez logrado el contacto, la probabilidad de respuesta completa es elevada, independientemente del modo de recolección de los datos.

Bajo este supuesto, la tasa de enlace observada en los hogares CAWI captura el componente dominante del proceso de respuesta. Como señalan Bethlehem, et al (2011), en encuestas con una alta cooperación, los factores que explican el contacto con el hogar (accesibilidad, ubicación, tipo de vivienda) son los que definen el sesgo de selección.

En consecuencia, los factores que explican el enlace en CAWI se interpretan como representativos de los mecanismos que determinan la inclusión de los hogares en el censo en general, incluyendo aquellos relevados exclusivamente mediante CAPI. Este supuesto fue validado empíricamente mediante la comparación de las estimaciones del modelo con las obtenidas por la ENEC a nivel nacional y para dominios amplios, observándose una elevada concordancia entre ambas.

Es importante tener en cuenta que este procedimiento se utiliza exclusivamente para la construcción de los factores de ajuste por no respuesta y no para la producción directa de las estimaciones.

En las siguientes figuras se muestra la distribución espacial de la omisión estimada mediante el *MRTE* y la pobreza a nivel de segmento censal en Montevideo. Se observa una clara superposición geográfica y una correlación espacial positiva entre ambos fenómenos: los mapas de calor revelan que los núcleos de mayor pobreza, concentrados fundamentalmente en la periferia de Montevideo, coinciden con las áreas de mayor omisión censal.

**Figura 4 : Omisión censal estimada mediante el MRTE por segmento censal. Montevideo**

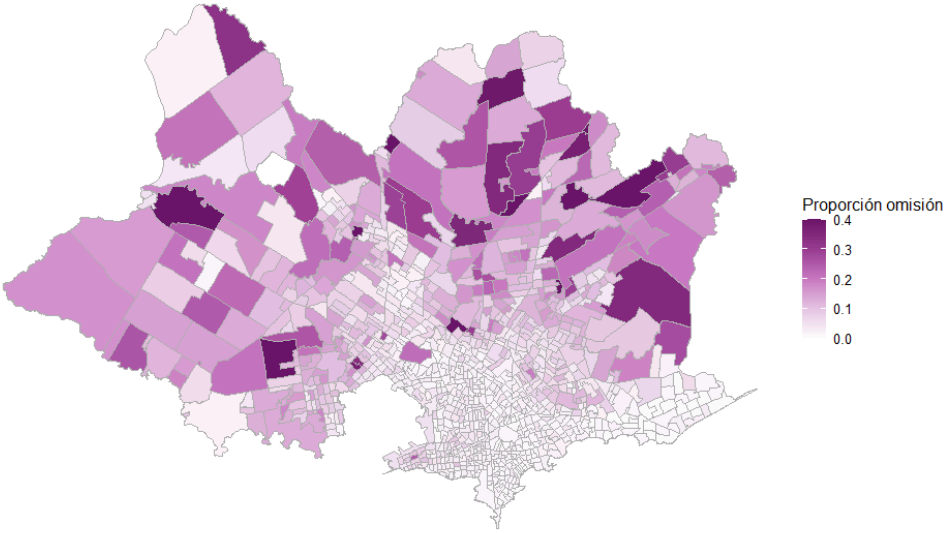
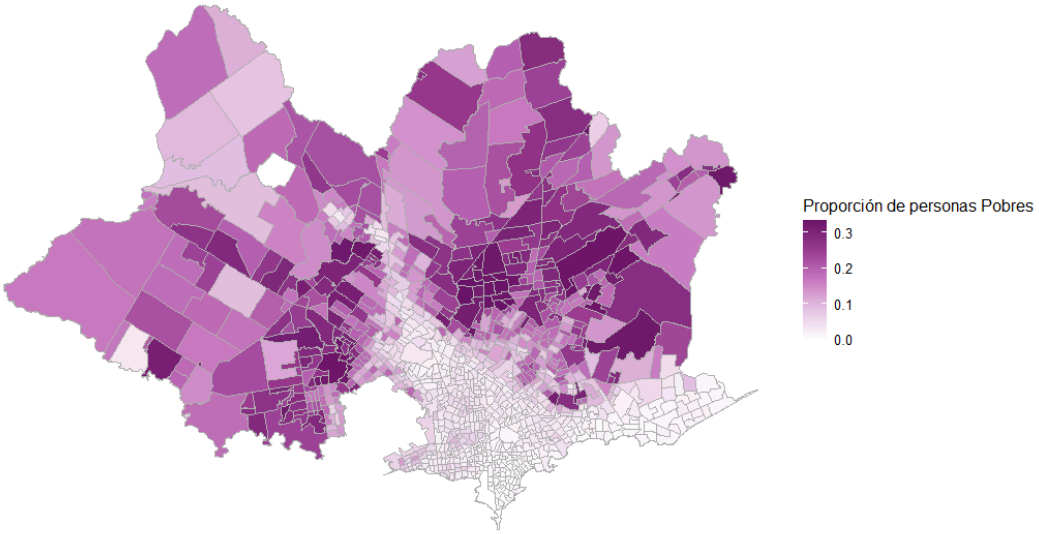



**Figura 5** : **Pobreza estimada mediante el modelo Fay- Harriot por segmento censal. Montevideo**




Esta congruencia territorial demuestra que la omisión no se distribuye de manera aleatoria, sino que responde a un patrón vinculado a la vulnerabilidad (Groves, 2006). En términos estadísticos, esto evidencia un sesgo en la omisión, de no corregirse, subestiman sistemáticamente la incidencia de la pobreza. Por lo tanto, el uso del *MRTE* en la construcción de los ponderadores resulta fundamental: permite construir ajustes de forma granular que captura heterogeneidades espaciales y socioeconómicas imposibles de corregir mediante estimaciones agregadas (como las de la ENEC) (Pfeffermann, 2013), garantizando así la representatividad y precisión de los resultados censales finales.

Para validar la consistencia de este patrón visual, se realizó un análisis de correlación entre ambas dimensiones. Los resultados arrojan un coeficiente de 0.67, lo que indica una asociación positiva y robusta entre el nivel pobreza y la omisión censal censal. Esta relación es estadísticamente significativa (*p-valor* < 0.001), descartando que la coincidencia espacial sea producto del azar. Asimismo, el modelo de regresión lineal presenta una pendiente positiva de 0.47, confirmando que a medida que aumenta la incidencia de la pobreza en el territorio, en promedio, se incrementa la omisión.

# 4. Estrategia de corrección de la omisión censal

## 4.1. Objetivo de la ponderación

La evidencia presentada anteriormente muestra que la omisión observada en el Censo no fue aleatoria, sino que se concentró (al menos) en determinados grupos poblacionales y regiones geográficas. En particular, la omisión resultó más elevada en áreas con mayores niveles de vulnerabilidad socioeconómica, así como en ciertos tramos de edad y en zonas con mayores dificultades operativas durante el relevamiento censal. Por lo tanto, la población efectivamente censada no constituye un subconjunto "representativo" de la población del Uruguay, lo que genera sesgos en las estimaciones obtenidas directamente a partir de los datos observados (i.e. hogares y/o personas censadas), especialmente para las variables centrales (core) del censo (Bethlehem, J., 2002), tales como educación, condiciones de vivienda, composición del hogar y otras características socioeconómicas.

Con el fin de poder corregir estos desbalanceos, el INE desarrolló una estrategia de ponderación de los hogares y personas efectivamente censados, cuyo objetivo es producir estimaciones lo más cercanas posible a las que se obtendrían bajo un censo con cobertura "teórica" completa, reduciendo el sesgo asociado a la omisión censal para las variables centrales del censo. El ajuste se realiza creando un sistema de ponderadores (pesos o factores de expansión) de los hogares censados, de modo que estos representen tanto a los hogares observados como a aquellos que no fueron relevados debido a moradores ausentes u omisiones propiamente dichas (United Nations, 2010).

La ponderación se aplica exclusivamente sobre los hogares efectivamente censados y sus integrantes. No se incorporan a la base de microdatos las personas provenientes de registros administrativos, ni se imputan hogares no observados. En cambio, el ajuste se realiza creando un sistema de ponderadores (pesos o factores de expansión) de los hogares censados, de modo que estos representen tanto a los hogares observados como a aquellos que no fueron relevados debido a moradores ausentes, omisiones propiamente dichas o falta de declaración de integrantes dentro de hogares contactados.

Desde el punto de vista estadístico, el problema se puede abordar tratando el censo como una muestra no probabilística, en donde la probabilidad o propensión de que un hogar o una persona haya sido efectivamente censada no es conocida y depende de múltiples factores, tanto observables como no observables (Elliott & Valliant, 2017; Chen, Li & Wu, 2020). En este contexto, el conjunto de hogares censados se interpreta como una muestra no probabilística, y la corrección por omisión se realiza mediante la modelización del mecanismo de selección y modelando las variables target del censo.

La estrategia adoptada por el INE combina dos componentes. Por un lado, se modela la probabilidad de respuesta al censo (i.e. voluntarismo), utilizando información auxiliar disponible (omisión a nivel de segmento censal). Por otro lado, se asume un modelo *superpoblacional* para las variables de interés del censo,que permite utilizar información externa para reducir el sesgo (Chambers & Clark, 2012). En particular, se utilizan los totales o conteos poblacionales obtenidos a partir del censo combinado, el cual, integra el operativo censal con registros administrativos; y se utilizan como información auxiliar para la calibración de los ponderadores, reproduciendo la distribución de la población por variables básicas como sexo, edad y grupos geográficos (Deville & Särndal, 1992).

La combinación de un modelo de respuesta con un modelo *superpoblacional* conduce a la utilización de un *estimador doblemente robusto*, el cual permite reducir el sesgo de las estimaciones aun cuando uno de los dos modelos no esté completamente bien especificado (Kang & Schafer, 2007; Kim & Haziza, 2014). Este enfoque resulta especialmente adecuado en contextos como el del censo 2023, el cual como ya se ha dicho, presenta una omisión diferencial y el mecanismo de inclusión de las unidades (hogares y personas) no responde a un mecanismo de selección conocido.

El objetivo final de la ponderación es, por lo tanto, producir estimaciones coherentes con la población total estimada (por el censo combinado) y, al mismo tiempo, mantener la consistencia interna de los datos censales, reduciendo el sesgo provocado por la omisión diferencial y permitiendo obtener estimaciones insesgadas (en la medida de lo posible) para las variables centrales del censo, lo cual, es "imposible" de realizar con el censo combinando al menos que se asuman (entre otras cosas) mecanismos de respuesta *naive*.

## 4.2. Unidad de análisis y unidad de ponderación

La estrategia de ponderación adoptada en el Censo 2023 se define a partir de la distinción entre la unidad observada, la unidad de inferencia y la unidad de ponderación. La información censal se releva a nivel de hogar y de persona, mientras que el objetivo de la estimación es representar correctamente a la población total residente, incluyendo tanto a las unidades efectivamente censadas como a aquellas que no fueron relevadas debido a la omisión. En el enfoque adoptado, la unidad de ponderación primaria es el hogar censado, y se crea un sistema de ponderadores para cada hogar efectivamente censado. Este criterio responde a la propia estructura del operativo censal, en el que la recolección de la información se organiza en torno a la vivienda y el hogar, y en el que la omisión se manifiesta principalmente como la falta de respuesta de hogares completos o como la omisión de integrantes dentro de hogares contactados.

Por lo tanto, a todos los integrantes de un mismo hogar se les asigna el mismo ponderador, correspondiente al hogar al que pertenecen. Esta decisión se adopta con el objetivo de mantener la consistencia interna de la información censal, evitando que dentro de un mismo hogar coexistan personas con diferentes ponderadores, lo que podría generar incoherencias en la estimación de variables relacionadas a nivel hogares y a nivel de persona (Alexander, 1987) . Si bien los ponderadores se definen a nivel de hogar, la calibración final se realiza utilizando información auxiliar disponible a nivel de persona, en particular los conteos de la población por sexo, edad y dominio geográfico, obtenida a partir del censo combinado.

Lo anterior implica que los ponderadores se ajustan de modo que la suma de las personas ponderadas reproduzca los totales poblacionales conocidos para estas variables, aun cuando el ponderador sea común para todos los integrantes del hogar (Lemaitre & Dufour, 1987). Esta estrategia es ampliamente utilizada en las encuestas a hogares y personas, por ejemplo, la Encuesta Continua de Hogares (ECH) exige que los ponderadores de los integrantes de un mismo hogar sean iguales con el fin de brindar coherencia con las estimaciones.

Este procedimiento permite compatibilizar dos objetivos de forma simultánea. Por un lado, asegurar la coherencia interna de los datos censales, manteniendo la estructura observada de los hogares. Por otro

lado, garantizar que las estimaciones agregadas reproduzcan la distribución demográfica de la población total, corrigiendo el sesgo asociado a la omisión diferencial del censo .

La construcción de los ponderadores se realiza en dos etapas jerárquicas. Como primer paso, se aplica un ajuste por no respuesta dentro de dominios geográficos definidos, modelando la probabilidad de que un hogar haya sido efectivamente censado. En segundo lugar, los ponderadores resultantes se someten a un proceso de calibración a conteos poblacionales conocidos (Deville & Särndal, 1992), utilizando como restricciones externas la estructura demográfica obtenida del censo combinado. La combinación de ambas etapas permite construir ponderadores consistentes con el marco conceptual de estimación doblemente robusta descrito anteriormente (Chen et al., 2020).

## 4.3. Muestra no probabilística

Como ya se dijo, debido a la existencia de un patrón de omisión heterogéneo, el conjunto de hogares y personas efectivamente relevados en el Censo 2023 no puede considerarse como una muestra "representativa" de la población. En la medida en que la probabilidad o propensión de ser censado varía entre hogares y personas; y depende de múltiples factores, por lo tanto, el censo observado puede tratarse como una muestra no probabilística de la población total. En este contexto, el mecanismo que determina la inclusión de un hogar en el censo no responde a una estrategia de selección (diseño muestral) conocida, sino a un proceso de respuesta complejo en el que intervienen factores operativos, territoriales, socioeconómicos, entre otros (Meng, 2018; Elliott & Valliant, 2017).

La probabilidad de que un hogar haya sido efectivamente censado depende tanto de características propias del hogar (y sus integrantes), como su localización geográfica o sus condiciones socioeconómicas, así como de factores externos asociados al operativo censal, tales como la accesibilidad del territorio, la modalidad de recolección utilizada, el número de visitas estipuladas en el censo según la ubicación geográfica del hogar o la posibilidad de contacto efectivo por parte del censista (Groves & Couper, 1998).

Esta situación puede modelarse mediante una probabilidad de respuesta al censo, que representa la probabilidad de que una unidad de la población haya sido efectivamente observada (i.e. censada). Dado que esta probabilidad no es conocida, debe ser estimada asumiendo un mecanismo o modelo de respuesta y utilizando información auxiliar disponible. Este enfoque adoptado combina la modelización de este mecanismo de respuesta con la utilización de información externa sobre la estructura de la población.

En particular, se asume un modelo superpoblacional para las variables de interés, en el cual la distribución de dichas variables en la población total puede ser aproximada utilizando información auxiliar disponible para todas las unidades, como la edad, el sexo y la localización geográfica. En el caso del Censo 2023, esta información auxiliar proviene de los totales poblacionales obtenidos a partir del censo combinado, que integra el operativo censal con registros administrativos y permite estimar con mayor precisión la distribución de la población por variables demográficas básicas. Estos totales se utilizan como referencia en el proceso de calibración de los ponderadores, de modo que las estimaciones finales reproduzcan la estructura conocida de la población aun cuando la base analítica se construya únicamente a partir de los hogares efectivamente censados.

La combinación de un modelo para modelar el mecanismo de la probabilidad de respuesta en conjunto con un modelo superpoblacional para las variables de interés conduce al uso de un estimador doblemente robusto, el cual permite reducir el sesgo de las estimaciones incluso en situaciones en las que uno de los dos modelos no esté completamente bien especificado (Chen et al., 2020; Kim & Haziza, 2014). Este enfoque resulta especialmente adecuado en el contexto actual, caracterizado por niveles de omisión no despreciables y por la ausencia de un mecanismo de selección probabilístico conocido.

En las secciones siguientes se describe de forma detallada este procedimiento, incluyendo la estimación de la probabilidad de respuesta, el ajuste por no respuesta y la calibración de los ponderadores a conteos demográficos conocidos.

## 4.4. Enfoque doblemente robusto para la construcción de los ponderadores

En presencia de omisión censal diferencial y en ausencia de un mecanismo de selección o respuesta conocido, la construcción de ponderadores requiere modelar explícitamente dicho mecanismo de inclusión en el censo y utilizar información auxiliar sobre la estructura de la población. Para ello se adopta un enfoque basado en *estimadores doblemente robustos* (Chen et al., 2020), ampliamente utilizados en la inferencia a partir de muestras no probabilísticas (Kim & Haziza, 2014) y en el tratamiento de problemas de no respuesta en las encuestas basadas en el diseño muestral (i.e. en las muestras aleatorias). Para una revisión detallada de estos enfoques metodológicos, se recomienda consultar (Wu, C., 2022).

Sea $U$ la población objetivo, de tamaño $N$, y sea $B \subset U$ el conjunto de unidades efectivamente observadas en el censo. Debido a la omisión censal, el conjunto $B$ puede interpretarse como una muestra no probabilística de la población total.

Sea $y_i$ la variable de interés definida para cada unidad $i \in U$, y sea el total poblacional

$$Y = \sum_{i \in U} y_i \,,$$

el parámetro que se desea estimar.

Se define la variable indicadora de respuesta al censo para la unidad $i$ como $R_i = 1$ si la unidad fue censada, y $R_i = 0$ en caso contrario. Luego, se asume que el mecanismo de no respuesta satisface el supuesto de *Missing at Random* (MAR) (Rubin, 1976), i.e., la probabilidad de respuesta depende únicamente de variables auxiliares observadas y no de la variable de interés no observada. El supuesto MAR es estándar en la literatura sobre tratamiento de la no respuesta en encuestas probabilísticas y no probabilísticas (Little, R. J. & Rubin, D. B, 2019, Valliant, R., et al, 2018).

Formalmente,

$$P(R_i = 1 \mid y_i, x_i) = P(R_i = 1 \mid x_i) \,,$$

donde $y_i$ representa la variable de interés y $x_i$ es un vector de variables auxiliares disponibles para todas las unidades o para dominios definidos, tales como edad, sexo y regiones geográficas. Bajo este supuesto, se define la probabilidad de respuesta de la unidad $i$ como:

$$\phi_i = P(R_i = 1 \mid x_i),$$

la cual puede modelarse utilizando información auxiliar proveniente de registros administrativos, del propio censo o de otras fuentes disponibles para toda la población. Este enfoque permite corregir el sesgo de no respuesta mediante técnicas de ponderación, calibración o modelización de propensiones de respuesta.

### 4.4.1 Ponderación por probabilidad inversa (*Inverse Probability Weighting, IPW*)

Si las probabilidades de respuesta $\phi_i$ fueran conocidas, un estimador del total poblacional puede obtenerse mediante ponderación por probabilidad inversa (*Inverse Probability Weighting*, *IPW*). Este enfoque consiste en ponderar cada unidad respondente por el inverso de su probabilidad de respuesta (Seaman & White, 2013), y puede interpretarse como una extensión del estimador de *Horvitz–Thompson* (*HT*), en la que las probabilidades de inclusión se ajustan incorporando el mecanismo de no respuesta (Särndal et al. 1992, Kim & Shao, 2013)

$$\hat{Y}_{IPW} = \sum_{i \in B} \frac{y_i}{\phi_i}$$

En la práctica, $\phi_i$ es desconocida y se reemplaza por una estimación $\hat{\phi}_i$ obtenida a partir de un modelo asumido de respuesta,

$$\hat{\phi}_i = \hat{P}(R_i = 1 \mid \mathbf{x}_i)$$

El estimador resultante es:

$$\hat{Y}_{IPW} = \sum_{i \in B} \frac{y_i}{\hat{\phi}_i} = \sum_{i \in B} w_i^{nr} \times y_i$$

Es importante tener en cuenta que este estimador reduce el sesgo cuando el modelo de respuesta está correctamente especificado (Bethlehem, J., 2020), pero puede resultar sesgado o inestable si dicho modelo no describe adecuadamente el mecanismo de respuesta.

### 4.4.2 Modelo superpoblacional

Como alternativa, puede asumirse un modelo *superpoblacional* para la variable de interés. Bajo este enfoque, la población finita se considera una realización de una población infinita hipotética, donde los valores $y_i$ se tratan como variables aleatorias vinculadas a las variables auxiliares (Särndal, et al, 1992). Se define entonces:

$$E(y_i \mid x_i) = m(x_i)$$

donde *m* es un método/modelo que relaciona la variable de interés con las variables auxiliares. A este modelo se le denomina habitualmente *working model* en la literatura de datos faltantes e inferencia, resaltando que la validez del estimador final depende de su correcta especificación (Haziza & Beaumont, 2017).

Bajo este enfoque, el total poblacional puede estimarse mediante un estimador basado en modelos (*Model-Based, MB*), dado por

$$\hat{Y}_{MB} = \sum_{i \in B} y_i + \sum_{i \in U-B} \hat{m}(\mathbf{x}_i)$$

El modelo *m* que vincula la variable de interés con las variables auxiliares, y se estima utilizando la muestra $B$, es decir, el conjunto de hogares y personas efectivamente censados.

A diferencia de los estimadores asistidos por modelos, la insesgadez del estimador *MB* depende de que el modelo estimado con la muestra $B$ sea válido para toda la población, en particular para las unidades no censadas (i.e., los omisos) (Valliant et al., 2000). En otras palabras, se requiere que la relación especificada por el modelo se mantenga también fuera de la muestra observada (Chambers & Clark, 2012). Por lo tanto, este estimador es consistente cuando el modelo *superpoblacional* está correctamente especificado, pero puede ser sesgado si dicho modelo no representa adecuadamente la relación entre $y_i$ y $x_i$.

### 4.4.3 Estimador doblemente robusto

El *estimador doblemente robusto* (*DR*) combina el estimador *IPW* y el estimador *MB*, dando a lugar a un estimador doblemente robusto (Chen et al., 2020),

$$\hat{Y}_{DR} = \sum_{i \in B} \frac{y_i - \hat{m}(\mathbf{x}_i)}{\hat{\phi}_i} + \sum_{i \in U} \hat{m}(\mathbf{x}_i)$$

Este estimador presenta la propiedad de doble robustez, en el sentido de que es consistente si se cumple al menos una de las siguientes condiciones (Kang & Schafer, 2007):

- El modelo de respuesta $\phi_i$ está correctamente especificado.
- El modelo superpoblacional $m(x_i)$ está correctamente especificado

Esta propiedad resulta especialmente relevante en el contexto del censo, donde el mecanismo de omisión no es completamente conocido.

Para demostrar formalmente la propiedad de doble robustez, es necesario analizar la estructura del sesgo del estimador. Si denotamos como $y_i$ al valor observado y $m(x_i)$ a la predicción del modelo *superpoblacional*, el sesgo de $\hat{Y}_{DR}$ puede expresarse aproximadamente como:

$$\text{Sesgo}(\hat{Y}_{DR}) \approx \sum_{i \in U} \left(1 - \frac{R_i}{\phi_i}\right)(y_i - m(\mathbf{x}_i))$$

En términos de valor esperado, esta expresión demuestra que el error sistemático es el producto de los errores de ambos modelos:

$$E[\text{Sesgo}(\hat{Y}_{DR})] \approx \sum_{i \in U} E\left[\frac{\pi_i - \phi_i}{\phi_i}\right] E[y_i - m(\mathbf{x}_i)]$$

,

donde $\pi_i$ representa la verdadera (y desconocida) probabilidad de respuesta y $\phi_i$ es la propensión estimada por el modelo. Por lo tanto, el sesgo tiende a cero si el modelo de propensión está correctamente especificado $(\phi_i = \pi_i)$ o si el modelo *superpoblacional* describe fielmente la realidad $(m(x_i) = E[y_i])$. Lo anterior garantiza una reducción drástica del riesgo de sesgo frente a otros estimadores simples.

Si el *working model* elegido es un modelo de regresión lineal, se tiene que:

$$m(\mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta},$$

y el estimador DR del total poblacional puede expresarse como una suma ponderada (Kim & Haziza, 2014)

$$\hat{Y}_{DR} = \sum_{i \in B} w_i y_i$$

donde

$$w_i = w_i^{nr} \times g_i$$

en donde el factor $g_i$ viene dado como:

$$g_i = 1 + (\mathbf{X} - \hat{\mathbf{X}}_{IPW})^T (\sum_{i \in B} w_i \times \mathbf{x}_i \mathbf{x}_i^T)^{-1} \mathbf{x}_i$$

con:

$$\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$$

$$\hat{\mathbf{X}}_{IPW} = \sum_{i \in B} w_i^{nr} \mathbf{x}_i$$

En este caso, puede demostrarse que los ponderadores resultantes satisfacen ecuaciones de balance (llamadas también *ecuaciones de calibración* o *benchmarking*) tales que reproducen exactamente los totales (o conteos) de las variables auxiliares utilizadas como predictores en el modelo (Kim & Haziza, 2014), es decir,

$$\sum_{i \in B} w_i \mathbf{x}_i^T = \sum_{i \in U} \mathbf{x}_i^T$$

Entonces, cuando el modelo basado es de regresión lineal, el estimador *DR* puede interpretarse como un estimador calibrado, en el sentido de que los ponderadores se ajustan para que la muestra reproduzca los totales poblacionales de las variables auxiliares (Deville & Särndal, 1992). Sin embargo, es importante tener en cuenta que el fundamento teórico es diferente. En el enfoque de calibración, los ponderadores se obtienen imponiendo restricciones sobre variables auxiliares conocidas, mientras que en el enfoque doblemente robusto los pesos surgen como consecuencia de combinar un modelo para la variable de interés con un modelo para el mecanismo de respuesta. A pesar de esta diferencia conceptual, ambos enfoques conducen al mismo tipo de estimador cuando el modelo utilizado para las variables de interés es lineal.

## 4.5. Variables auxiliares utilizadas en la construcción del estimador *DR*

La implementación del estimador *DR* requiere la utilización de variables auxiliares que permitan modelar tanto la propensión de respuesta al censo como la distribución de las variables de interés en la población total (i.e. las variables auxiliares para la construcción del modelo de predicción). En el caso del Censo, la selección de estas variables se realizó considerando criterios teóricos, empíricos y operativos, con el objetivo de reducir el sesgo asociado a la omisión diferencial y mejorar la calidad de las estimaciones para las variables centrales del censo (Little & Vartivarian, 2005).

En términos generales, las variables auxiliares utilizadas cumplen al menos una de las siguientes funciones:

1. Explicar la probabilidad o propensión de que un hogar haya sido efectivamente censado.
2. Estar correlacionadas con las variables de interés del censo.
3. Estar disponibles para la población total a través del censo combinado o de otras fuentes auxiliares de información (Haziza & Beaumont, 2017).

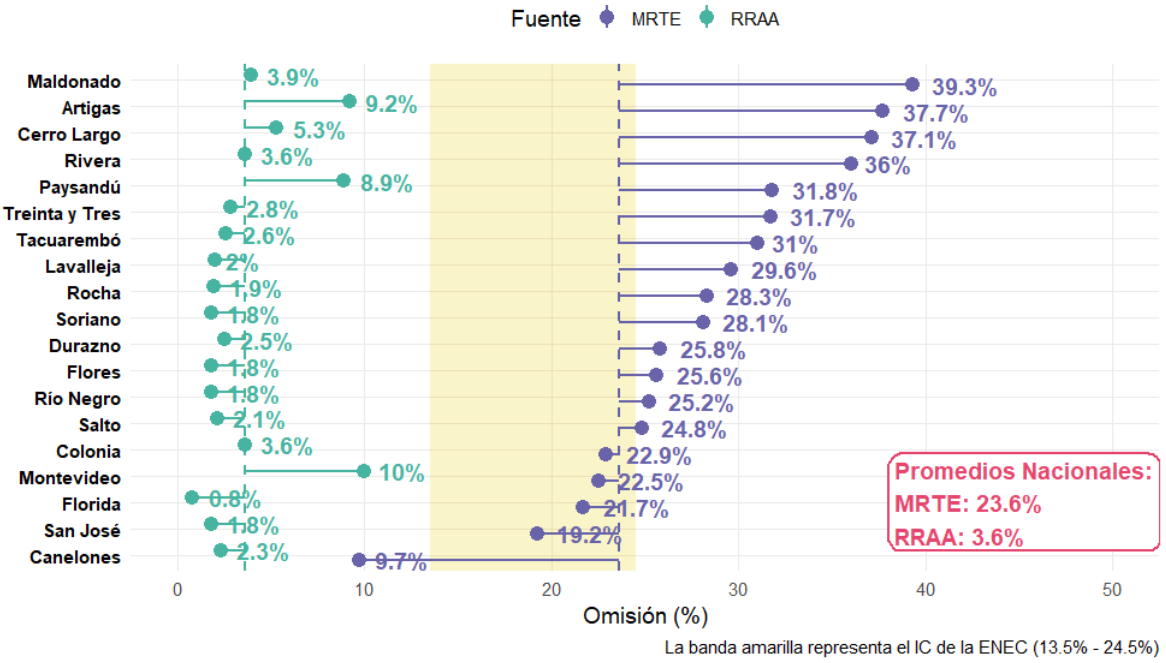
El uso simultáneo de las variables auxiliares relacionadas con el mecanismo de respuesta y con las variables de interés del censo es fundamental para garantizar las propiedades del estimador doblemente robusto.

Las principales variables auxiliares utilizadas en la etapa de construcción del modelo de las variables de interés (i.e. para la calibración) corresponden a la distribución de la población por características demográficas básicas obtenidas a partir del censo combinado, el cual integra la información relevada en el operativo censal con registros administrativos. En particular, se utilizaron totales poblacionales por: sexo, edades simples y departamento de residencia.

Estos totales se consideran de buena calidad, ya que resultan de la integración del censo con registros administrativos que contienen "*señales de vida*" de la población. La calibración de los ponderadores a estos totales permite asegurar que la población estimada reproduzca correctamente la estructura demográfica conocida, aun cuando la base analítica se construya únicamente a partir de los hogares efectivamente censados. Si bien era posible utilizar información de residencia a un nivel más fino (e.g, localidad, barrio, etc), se decidió no incorporar debido a que, durante el proceso de ponderación, se detectó la presencia de un "*sesgo de urbanicidad"* en el algoritmo utilizado para integrar las personas

provenientes de RRAA al censo combinado. En particular, se observó que las personas agregadas por registros tenían mayor probabilidad de ser asignadas a localidades de mayor tamaño, independientemente de su lugar real de residencia. Este comportamiento generaba inconsistencias en los conteos de población por localidad, especialmente en el área rural dispersa. En dicha área, la omisión estimada a partir de la ENEC, así como la evidencia proveniente de la teoría general de encuestas y de la experiencia operativa del INE, indican niveles de omisión aproximadamente dobles que en el área urbana (i.e 20% aproximadamente). Sin embargo, la estimación ad hoc basada en el censo combinado arrojaba niveles de omisión inferiores al 4%, lo cual resulta operativamente implausible dadas las características del territorio y el hecho de que en el área rural se realizó, en general, una única visita censal. En la siguiente figura se presenta la omisión por departamento utilizando los distintos métodos para aproximar la omisión del censo:

**Figura 5: Porcentaje de omisión censal por departamento según fuente de datos (MRTE y RRAA) y comparación con el intervalo de confianza de la ENEC.**




Adicionalmente, en Montevideo se detectó otro efecto del mismo algoritmo: las personas con menor vinculación a registros administrativos, es decir, aquellas con mayor nivel de vulnerabilidad o con menor presencia en sistemas formales, presentaban menor probabilidad de ser incorporadas al censo combinado. Como consecuencia, la omisión estimada ad hoc a partir del censo combinado resultaba artificialmente más baja precisamente en las áreas donde, debido al mayor uso de modalidad CAPI y complejidad de acceso, la ENEC estimó niveles más altos de omisión. Este patrón reforzó la evidencia de que los conteos derivados directamente del censo combinado no eran adecuados para calibraciones a niveles geográficos finos. Por esta razón, se optó por no calibrar a niveles geográficos más desagregados que aquellos para los cuales los totales disponibles presentaban suficiente consistencia y calidad. Esta situación también motivó que el INE revisará posteriormente los conteos de población por localidad y procediera a realizar una nueva estimación de la cantidad de habitantes a ese nivel geográfico, tal como se describe más adelante.

Finalmente, el uso de estas variables resulta especialmente importante debido a que la ENEC estimó que la omisión presenta un componente diferencial por edad, sexo y territorio. Además de las variables demográficas, para la construcción del estimador DR se incorporó información auxiliar vinculada al territorio y al proceso de relevamiento, con el objetivo de modelar el mecanismo de respuesta al censo. Entre estas variables se incluyen:

- segmento censal (conjunto de zonas censales <manzanas>)
- modalidad de relevamiento (CAWI / CAPI)
- tasa de enlace de cuestionarios web en el segmento

La inclusión del segmento censal como unidad básica en la modelización del mecanismo de respuesta permite capturar heterogeneidad territorial fina en la tasa de omisión, la cual no puede ser observada utilizando únicamente dominios geográficos amplios (e.g. departamentos). La tasa de enlace de cuestionarios web se utilizó como proxy de la probabilidad de contacto efectivo en el segmento, bajo el supuesto de que los segmentos con menor tasa de enlace presentan también mayor probabilidad de omisión en el relevamiento presencial. La utilización de una desagregación tan fina se debe a la evidencia proveniente de la ENEC la cual mostró que la omisión censal se concentra en zonas con mayores niveles de vulnerabilidad socioeconómica. Por lo tanto, utilizar el segmento censal como unidad para modelar la no respuesta permite mejorar la estimación de la probabilidad de respuesta y contribuyen a reducir el sesgo en la estimación de variables censales fuertemente asociadas al nivel socioeconómico, como educación, condiciones de vivienda y características del hogar.

Las variables auxiliares seleccionadas se utilizan de manera conjunta en las dos componentes del procedimiento de estimación: i) en la modelización de la propensión o voluntarismos de participar en el censo, a través de modelos que estiman la propensión a ser censado, y ii) en la calibración de los ponderadores, mediante la imposición de restricciones que aseguren la coherencia con los totales poblacionales conocidos.

Cuando las variables utilizadas en ambos componentes están relacionadas tanto con el mecanismo de respuesta y si a su vez las mismas se encuentran correlacionadas con las variables de interés, el estimador resultante presenta las propiedades de doble robustez, lo que permite reducir el sesgo aun cuando uno de los modelos no esté completamente bien especificado. En el caso del Censo 2023, la combinación de variables demográficas provenientes del censo combinado, variables territoriales del operativo censal permitió construir un conjunto de ponderadores consistente con el marco conceptual adoptado y adecuado para la producción de estimaciones con menor sesgo para las variables centrales del censo.

## 4.6. Modelización de la probabilidad de respuesta

Como se describió anteriormente, el enfoque adoptado para la construcción de los ponderadores requiere (entre otras cosas) poder estimar la probabilidad de que una unidad (hogar) haya sido efectivamente observada en el censo. Esta probabilidad o propensión, denotada por $\phi_i$, representa la propensión de inclusión en la base censal y constituye un componente fundamental en el ajuste por omisión. Como ya se dijo, dado que el Censo no responde a un diseño muestral aleatorio, las probabilidades de inclusión no son conocidas y deben ser estimadas a partir de información auxiliar disponible. Para ello se especifica un modelo de respuesta que aproxima la probabilidad de inclusión asumiendo que el mecanismo de respuesta es MAR (Missing at random), es decir, que la propensión

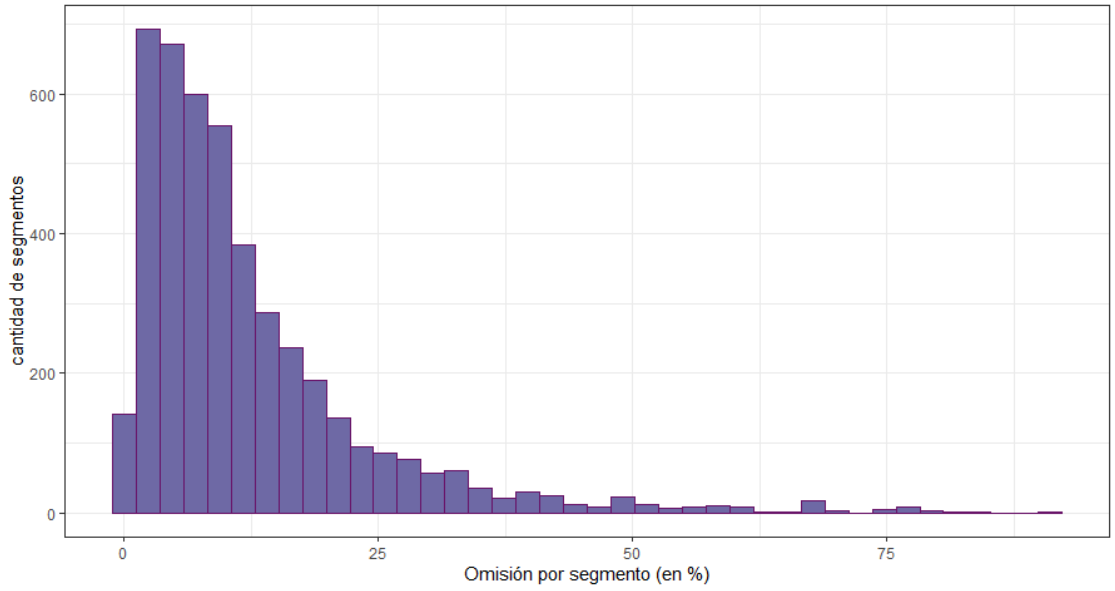
de un hogar a responder depende de variables $x$ conocidas y no de las propias variables de interés $y$ que mide el censo. Si este supuesto se cumple, obtendrán estimaciones insesgadas de los parámetros de interés que aborda el censo.

La modelización de la probabilidad de respuesta se realiza a nivel de hogar, dado que la omisión censal se manifiesta principalmente como la falta de respuesta de hogares completos o como dificultades de contacto durante el operativo de campo. En consecuencia, se asume que todos los integrantes de un mismo hogar comparten la misma probabilidad de inclusión en el censo y, por lo tanto, reciben el mismo factor de ajuste por no respuesta. El modelo de respuesta incorpora variables auxiliares asociadas tanto a las características demográficas de la población como a condiciones territoriales y operativas del relevamiento. Entre las variables consideradas se incluyen:

- segmento censal
- modalidad de relevamiento (CAWI / CAPI)
- tasa de enlace de cuestionarios web en el segmento

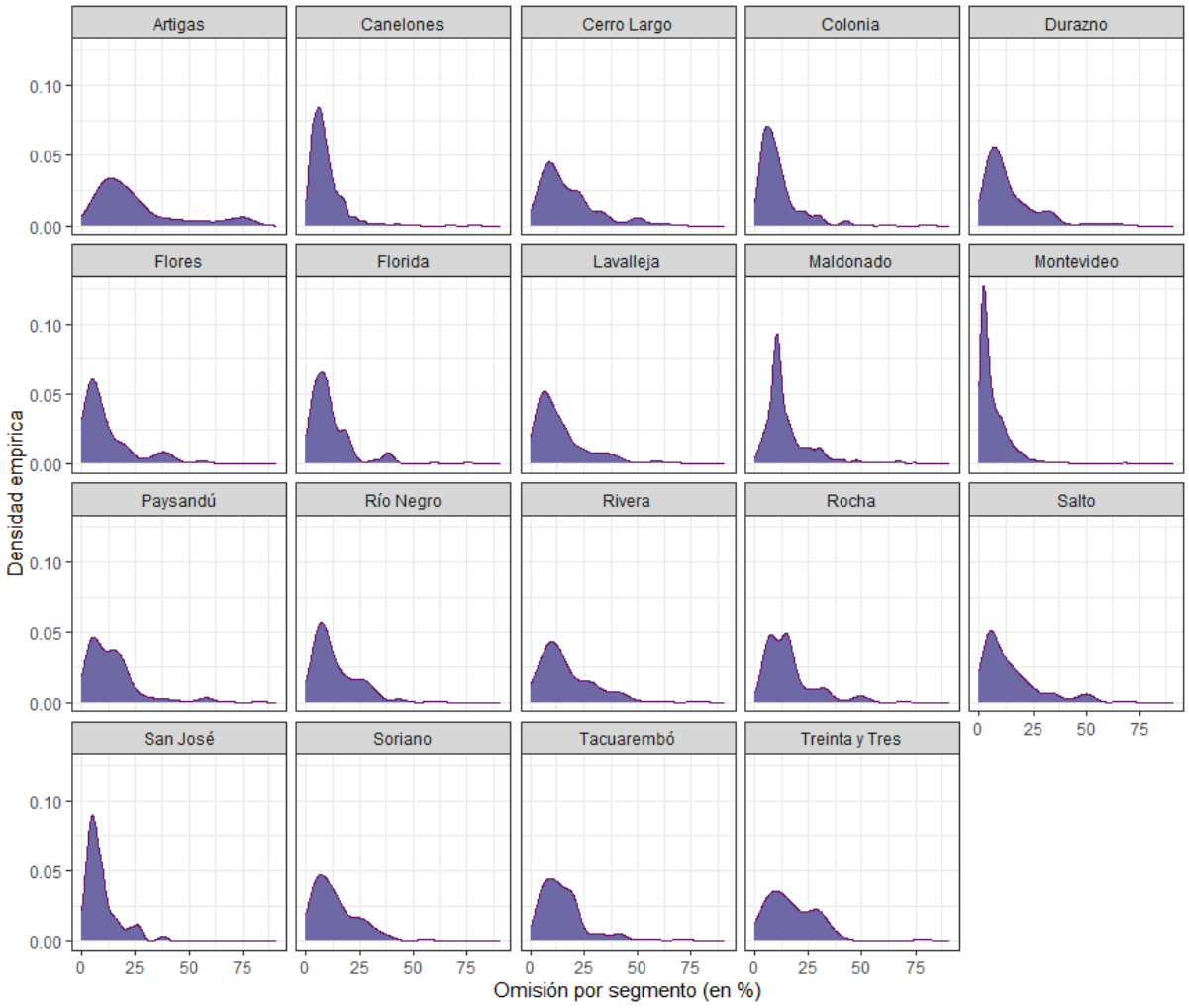
Como ya se dijo, el modelo *MRTE* utiliza la tasa de enlace de cuestionarios web como indicador de la probabilidad de contacto efectivo en el segmento, bajo el supuesto de que los segmentos con menor tasa de enlace presentan mayores dificultades operativas y, por lo tanto, mayores niveles de omisión. Esto refuerza que los hogares efectivamente censados no se comportan como una muestra "seleccionada" bajo un muestreo con iguales probabilidades de selección. En la siguiente figura se presenta la distribución de la omisión estimada por el modelo *MRTE* a nivel de segmento censal:

**Figura 6: Distribución de la estimación omisión por segmento censal**




Por otra parte, en la siguiente figura se puede apreciar que la distribución de la omisión por segmento tiene distribuciones marcadamente distintas según el departamento.

**Figura 7: Distribución de la estimación omisión por segmento censal según departamento**




### 4.6.1 Especificación del modelo de propensión

Se utilizan clases de no respuesta para el ajuste para estimar las propensiones de respuesta de los hogares, las cuales, se definen a partir de los segmentos censales. Esta decisión se basa en el supuesto de que, si es posible conformar grupos de hogares (y personas) con probabilidades de respuesta similares, el sesgo por no respuesta se reduce sustancialmente (Little, 1986). A su vez, resulta deseable que las variables utilizadas para formar dichas clases de no respuesta también estén relacionadas con las principales variables de interés del censo, dado que esto mejora la capacidad del ajuste para corregir sesgos (Little & Vartivarian, 2005).

En este sentido, los segmentos censales constituyen una unidad particularmente adecuada, dado que agrupan hogares con características geográficas, socioeconómicas y de acceso relativamente homogéneas, lo que permite aproximar tanto comportamientos de respuesta similares como perfiles de hogares comparables en términos de las variables de interés (Bethlehem et al., 2011). De esta manera, dentro de cada segmento censal se asume que los hogares presentan propensiones de respuesta relativamente homogéneas. El objetivo de las clases de ajuste es reunir unidades con propensiones de respuesta similares y, en la medida de lo posible, también con valores esperados similares en las variables de interés (Haziza & Beaumont, 2017). Aunque en la práctica las variables de interés no se

observan para los no censados (i.e. los no respondentes), el uso de segmentos censales permite aprovechar información auxiliar disponible para aproximar ambas dimensiones.

Sea $c = 1, ..., C$ las clases de no respuesta, definidas a partir de los segmentos censales. La propensión de respuesta del hogar $i$ que pertenece al segmento censal $c$ se modela como $\phi_i = \phi_c \quad \forall i \in c$. Lo anterior (como ya se dijo) realiza el supuesto (*hipótesis*) de que los hogares dentro de un mismo segmento censal presentan comportamientos de respuesta similares, se asume que todos los hogares pertenecientes a la clase $c$ tienen la misma propensión de respuesta (Valliant et al., 2018).

El uso de segmentos censales como clases de ajuste por no respuesta busca optimizar el equilibrio entre sesgo y varianza. La varianza del estimador IPW depende de la dispersión de las propensiones dentro de la población. La varianza de los ponderadores mediante la maximización de la homogeneidad intra-clase y viene dada como:

$$V(w^{nr}) = \frac{1}{N} \sum_{c=1}^{C} N_c \left( \frac{1}{\hat{\phi}_c} - 1 \right),$$

donde $N_c$ es el tamaño de la clase (segmento) $c$ y $\hat{\phi}_c$ es la propensión estimada.

Lo anterior, justifica técnicamente el uso del segmento censal como de clase de no respuesta: al maximizar la homogeneidad intra-clase, se minimiza la variabilidad interna de los ponderadores. A su vez, esta expresión fundamenta la necesidad de establecer umbrales máximos para los factores de ajuste, con el fin de evitar que valores de propensión excesivamente bajos disparen la varianza global y comprometan la estabilidad de las estimaciones.

### 4.6.2 Construcción de los ponderadores ajustados por omisión

Las probabilidades estimadas se utilizan para construir ponderadores ajustados por omisión (no respuesta), bajo el enfoque del estimador *IPW*. El ponderador ajustado por no respuesta para el hogar $i$ queda definido como:

$$w_i^{nr} = \frac{1}{\hat{\phi}_i}$$

donde $\hat{\phi}_i = 1 - \widehat{\mathrm{TO}}_c$ y la tasa de omisión estimada en un segmento censal cualquiera $c$ se denota como $\widehat{\mathrm{TO}}_c$.

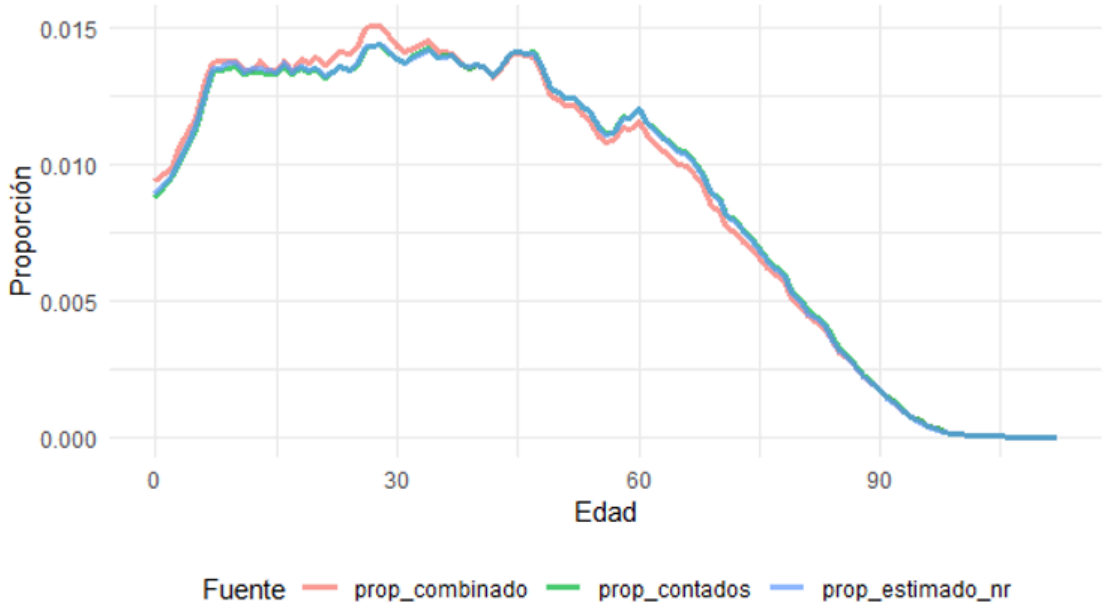
Es importante tener en cuenta que estos ponderadores asignan mayor peso a las unidades con menor probabilidad estimada, es decir, para hogares que residen en segmentos con mayor nivel de omisión compensando así la subrepresentación de ciertos grupos poblacionales. A su vez, en esta etapa no se permiten ponderadores ajustados por no respuesta mayor a 2.5 (Valliant, et al, 2018) . Esto se debe a un *trade-off* entre sesgo y varianza dado que ajustes muy grandes pueden aumentar de forma excesiva la variabilidad de los ponderadores (Valliant, et al, 2018) y así brindar estimaciones inestables (sobre todo a niveles más finos de la población).

Por otra parte, es importante destacar que como era esperable, los ponderadores $w^{nr}$ estiman de forma coherente el tamaño de la población estimada en el censo., i.e. utilizando suma de los ponderadores $w^{nr}$ .

$$\hat{N} = \sum_{i \in B} w_i^{nr} \approx 3,510,000$$

Este primer paso (ajuste) en la construcción de los ponderadores del censo, busca reducir el sesgo territorial, el cual se encuentra asociado a las condiciones de vida. Sin embargo, como puede apreciarse en la siguiente figura, donde se presenta la proporción de personas por edad entre los individuos efectivamente censados, las estimaciones obtenidas utilizando el ponderador ajustado por no respuesta y los datos provenientes del censo combinado, este ajuste no logra reducir el sesgo por edad.

**Figura 8: Distribución por edad según distintas fuentes**




Por lo tanto, el uso de ponderadores inversamente proporcionales a la probabilidad estimada de respuesta constituye el primer componente del estimador doblemente robusto. El segundo componente corresponde a la calibración de los ponderadores a totales o conteos poblacionales conocidos. La combinación de ambos componentes permite obtener estimaciones consistentes aun cuando el modelo de respuesta no esté completamente bien especificado, siempre que las variables auxiliares utilizadas en la calibración estén relacionadas con las variables de interés. Este enfoque resulta particularmente adecuado en el contexto del Censo, donde la omisión censal presenta un patrón diferencial por territorio, características socioeconómicas, edad y sexo, y donde la disponibilidad de información proveniente del censo combinado permite imponer restricciones externas que mejoran la calidad de las estimaciones. Los ponderadores obtenidos en esta etapa constituyen factores iniciales que corrigen la omisión diferencial dentro de cada segmento censal, mediante el uso de ponderadores inversamente proporcionales a la probabilidad estimada de respuesta. Este ajuste representa el primer componente del estimador doblemente robusto y permite reducir el sesgo asociado a diferencias en las tasas de respuesta entre segmentos con distintas características territoriales y socioeconómicas.

Sin embargo, estos ponderadores no garantizan por sí solos la coherencia con los totales poblacionales conocidos ni eliminan completamente el sesgo remanente asociado a diferencias en las tasas de respuesta por edad, sexo u otras características demográficas. Por este motivo, los ponderadores ajustados por no respuesta se utilizan como punto de partida para una etapa posterior de calibración, en la cual se imponen restricciones externas provenientes del censo combinado, de forma que las estimaciones reproduzcan la estructura demográfica conocida de la población.

Esta calibración constituye el segundo componente del estimador doblemente robusto. La combinación entre el ajuste por no respuesta dentro de segmento y la calibración a totales poblacionales permite obtener estimaciones consistentes aun cuando el modelo de respuesta no esté completamente bien especificado, siempre que las variables auxiliares utilizadas en la calibración estén relacionadas con las variables de interés.

Este enfoque resulta particularmente adecuado en el contexto del Censo 2023, donde la omisión censal presenta un patrón diferencial por territorio, características socioeconómicas, edad y sexo, y donde la disponibilidad de información proveniente del censo combinado permite imponer restricciones externas que mejoran la calidad de las estimaciones.

## 4.7. Calibración a totales poblacionales del censo combinado

El ajuste por no respuesta dentro del segmento descrito en la sección anterior permite corregir parte del sesgo asociado a la omisión. Sin embargo, los ponderadores obtenidos en esa etapa no garantizan que las estimaciones reproduzcan exactamente la estructura demográfica de la población total. Para asegurar esta coherencia, se realiza una etapa adicional de calibración de ponderadores utilizando como referencia los totales poblacionales obtenidos a partir del censo combinado.

Es fundamental señalar que los totales de control utilizados en este proceso (sexo, edad y departamento) provienen de los conteos finales del Censo Combinado. En el marco de la presente metodología, esta información auxiliar se define como totales poblacionales conocidos, asumiéndose, por tanto, como libres de error de muestreo. Esta decisión dota al Censo Combinado del estatus de Marco de Referencia Oficial, garantizando que el proceso de calibración no solo reduzca el sesgo de selección, sino que asegure la convergencia absoluta de los microdatos con las cifras demográficas institucionales.

El censo combinado integra la información relevada en el operativo censal con registros administrativos, permitiendo estimar con mayor precisión la distribución de la población por variables demográficas básicas. En particular, y como ya se dijo, se dispone de totales poblacionales por sexo, edad y departamento, los cuales, como ya se dijo, se consideran de buena calidad y se utilizan como restricciones externas en el proceso de calibración. El vector/set para calibrar/modelar $\mathbf{x}_i = (x_{1i}, ...., x_{Ji})^T$ son variables de pertenencia (indicadoras) a distintos grupos demográficos. Más precisamente edades simples, sexo y departamento de residencia.

La calibración bajo una distancia lineal se puede ver como un estimador de regresión, en este paso, el modelo *superpoblacional m* queda definido como:

$$E_m(y_i) = \mathbf{x}_i^\top \boldsymbol{\beta} = \mu + \text{edad}_j + \text{sexo}_k + \text{dpto}_l + (\text{edad} \times \text{sexo} \times \text{dpto})_{jkl},$$

El modelo anterior corresponde al estimador post-estratificado que es un caso particular de un estimador calibrado. Por lo tanto, la calibración de los ponderadores se realiza de forma que las estimaciones obtenidas a partir de los hogares censados reproduzcan exactamente estos totales.

Es importante tener en cuenta que cuando las variables utilizadas en la calibración están relacionadas tanto con el mecanismo de respuesta como con las variables de interés, el estimador resultante presenta la propiedad de doble robustez, es decir, produce estimaciones consistentes aun cuando uno de los dos modelos no esté completamente bien especificado. En el contexto del Censo 2023, la calibración a totales de sexo, edad y departamento equivale a asumir que las variables de interés están relacionadas con estas variables demográficas, lo cual resulta razonable dado el fuerte vínculo entre dichas variables y la mayoría de las características sociales y económicas relevadas en el censo. Esta interpretación es fundamental dentro del enfoque doblemente robusto, ya que la combinación del modelo de respuesta con el modelo superpoblacional implícito en la calibración permite reducir el sesgo aun cuando uno de los dos modelos no esté completamente bien especificado.

De forma más formal, sea $w_i^{nr}$ el ponderador inicial obtenido luego del ajuste por no respuesta dentro de cada segmento. El objetivo de la calibración es obtener nuevos sistema de ponderadores $w_i = g_i w_i^{nr}$ que sean lo más cercanos posible a $w_i^{nr}$, pero que cumplan las restricciones impuestas por los totales poblacionales conocidos (Valliant, et al, 2018). Los ponderadores calibrados se determinan de modo que cumplan la ecuación de calibración:

$$\sum_{i \in B} w_i \mathbf{x}_i = \mathbf{X}$$

El procedimiento de calibración se formula como un problema de optimización en el cual se minimiza una función de distancia entre los ponderadores calibrados y los ponderadores ajustados por no respuesta, sujeto a que cumplan las restricciones anteriores. Este enfoque corresponde al estimador de calibración propuesto por Deville & Särndal (1992), ampliamente utilizado en encuestas oficiales llevadas a cabo por las ONEs.

### 4.7.1 Calibración con ponderadores definidos a nivel de hogar

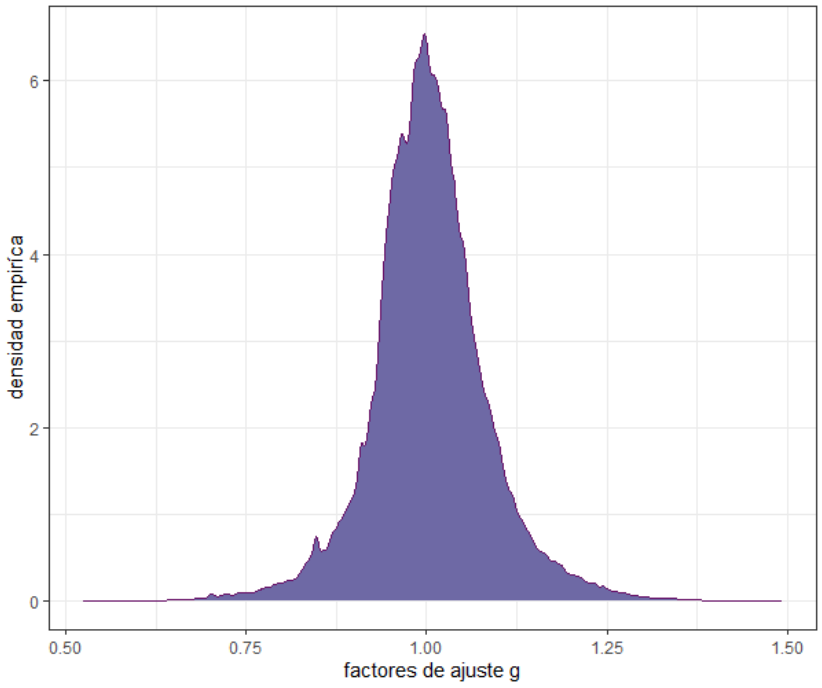
Si bien la información auxiliar utilizada en la calibración se encuentra definida a nivel de persona, los ponderadores se construyen a nivel de hogar, de modo que todos los integrantes de un mismo hogar tengan el mismo ponderador. Este criterio se adopta con el objetivo de preservar la coherencia interna de la información censal (Alexander, C. H. 1987), evitando que dentro de un mismo hogar existan personas con distintos ponderadores, lo que podría generar inconsistencias en la estimación de variables relacionadas con los microdatos de personas y hogares.

En consecuencia, la calibración se realiza de manera que los ponderadores asignados a los hogares permitan reproducir los totales poblacionales a nivel de persona, respetando al mismo tiempo la restricción de que todos los integrantes del hogar tengan el mismo ponderador. Para obtener esto último se utiliza el *método de integración*. Este indica que las variables auxiliares deben ser previamente promediadas a nivel del hogar antes de realizar la calibración final de los ponderadores (Lemaitre, G., & Dufour, J. 1987). Si bien este método proporciona coherencia entre las estimaciones de personas y hogares, produce un leve aumento en la variabilidad de los ponderadores (i.e. aumenta el *efecto de diseño de Kish*) (Kish, L. 1965), lo que se traduce en un leve aumento de la variabilidad de

los errores estándar de las estimaciones en comparación a la obtenida con ponderadores finales distintos para las personas de un mismo hogar.

En la siguiente figura se presentan los factores de ajuste provenientes de la calibración en donde se puede apreciar que la distribución de los mismos es en forma de campana y centrada (i.e. la media) aproximadamente en uno.

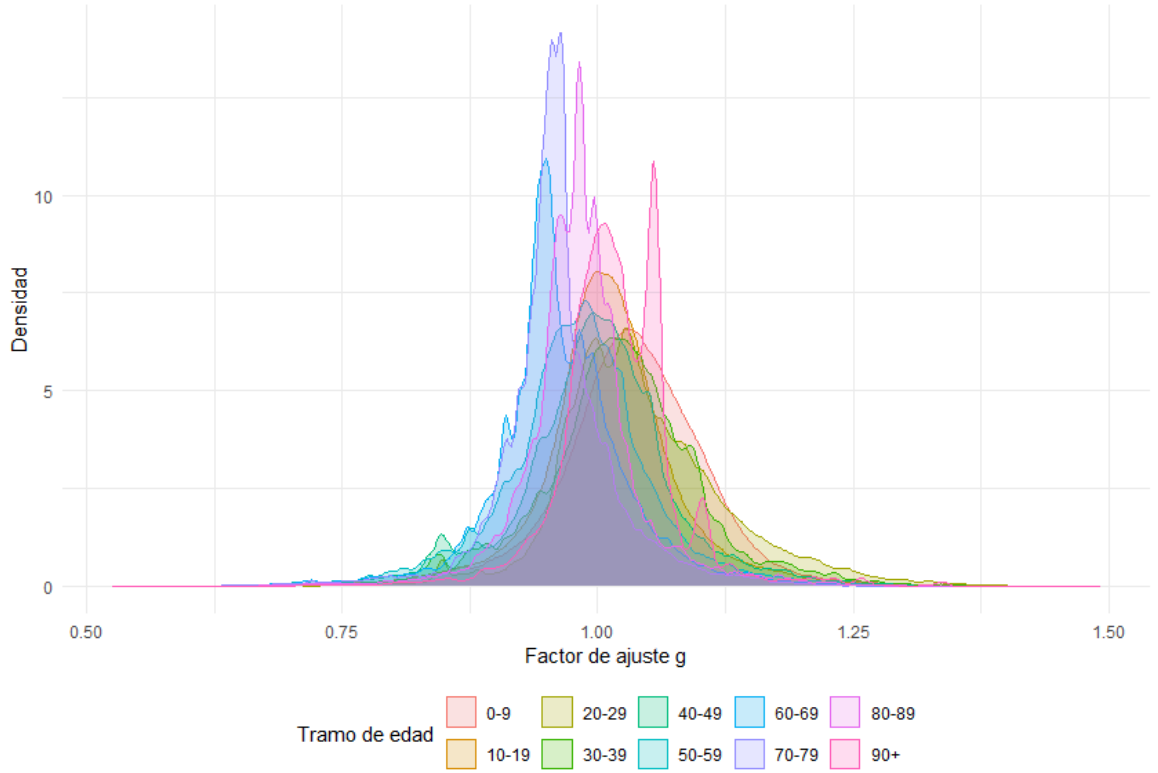
**Figura 9: Distribución de los factores de ajustes *g* provenientes de la calibración**




De igual forma, se presentan los factores de ajuste provenientes de la calibración segmentados por grupos de edad. En todos los tramos etarios se observa una distribución aproximadamente acampanada y centrada cerca de 1, lo que indica que, en promedio, los ajustes aplicados no se alejan sustancialmente del ponderador original.

Sin embargo, existen diferencias en la dispersión de las distribuciones entre grupos de edad. Algunos tramos presentan distribuciones más concentradas alrededor de 1, mientras que otros muestran curvas más achatadas y con colas más pesadas. Esto refleja que, para ciertos grupos etarios, la calibración debió aplicar ajustes más fuertes para corregir situaciones de subrepresentación o sobrerepresentación en variables como edad y sexo. En particular, los grupos de edades intermedias parecen presentar una mayor dispersión de los factores $g_i$, lo que sugiere una necesidad más frecuente de aplicar correcciones relativamente grandes. Por el contrario, en los grupos de mayor edad las distribuciones tienden a estar más concentradas en torno a 1, indicando ajustes más moderados.

**Figura 10: Distribución de los factores de ajustes *g* por tramo de edad**

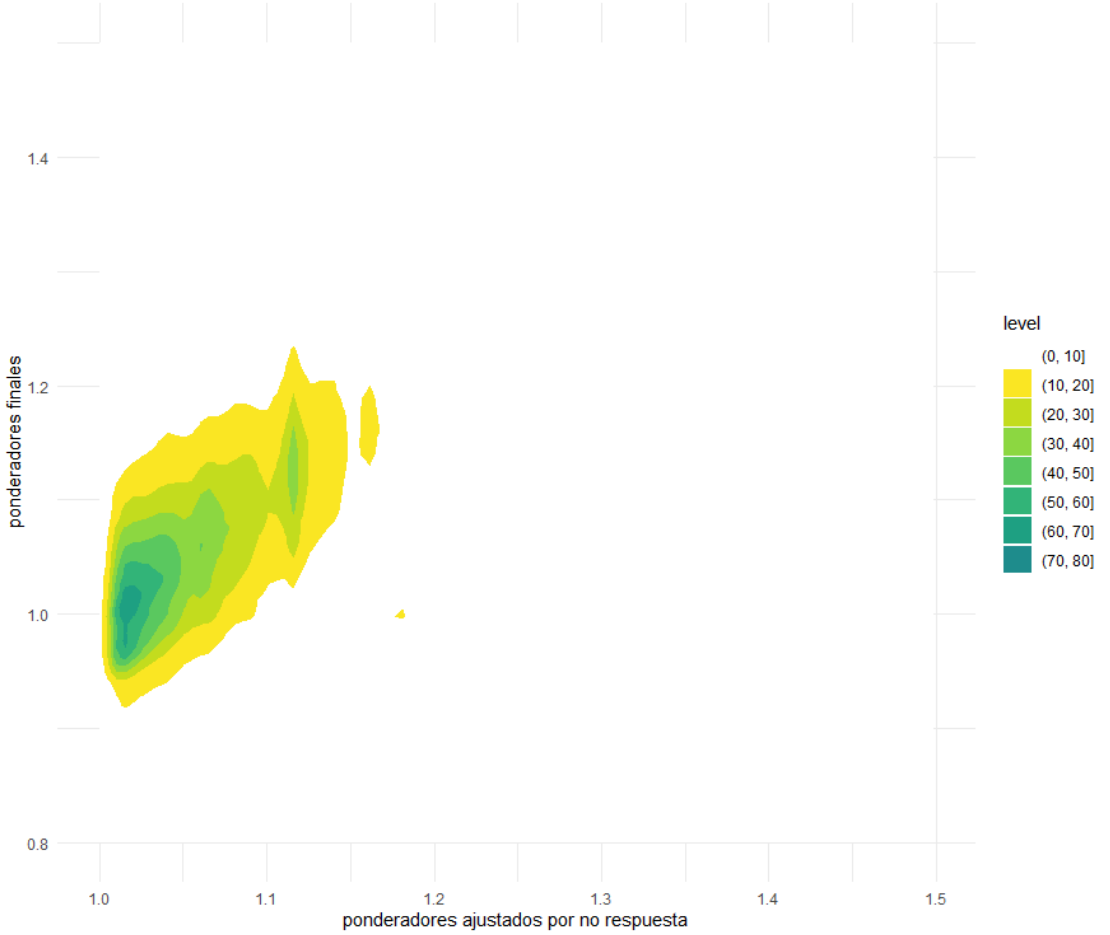



Finalmente, en la siguiente figura se presenta la distribución conjunta entre los ponderadores ajustados por no respuesta ($w^{nr}$) y los ponderadores finales ($w$) obtenidos luego de la calibración doblemente robusta.

La intensidad de color representa la densidad de observaciones en cada región del gráfico: las zonas más oscuras corresponden a combinaciones de ponderadores más frecuentes, mientras que las zonas más claras representan combinaciones menos frecuentes.

Por ejemplo, se puede observar que la mayor parte de los casos se concentra en valores de $w^{nr}$ cercanos a 1.01–1.05 y ponderadores finales $w$ entre aproximadamente 0.98 y 1.05. En términos generales, esto indica que la calibración final no introduce cambios "abruptos" sobre los ponderadores ajustados por no respuesta, aunque sí modifica algunos grupos específicos de observaciones.

**Figura 11: Densidad bivariada entre los ponderadores ajustados por no respuesta y los ponderadores finales**



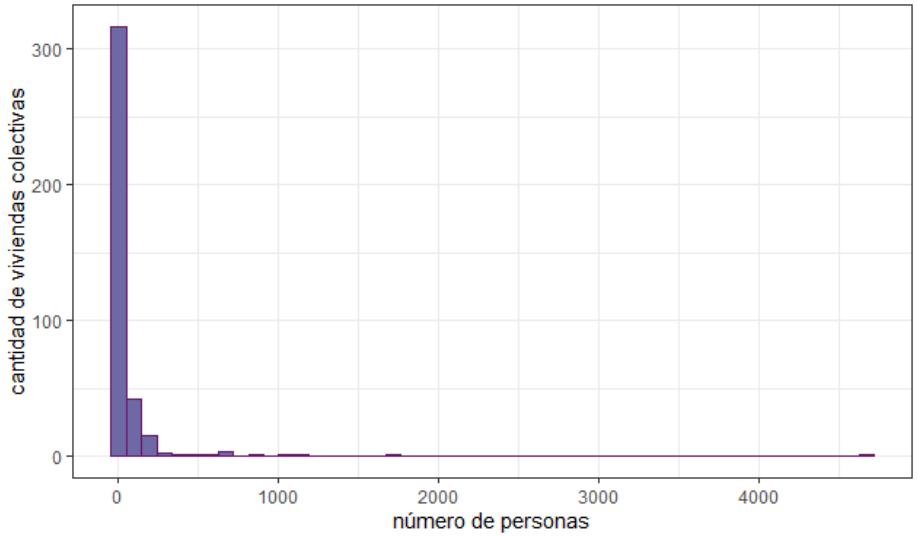

## 4.7.2 Unidades de inclusión forzosa

No todas las personas contadas en el censo entran en el proceso de ponderación descrito anteriormente. Existen casos particulares a los cuales se les asigna un ponderador igual a uno (i.e. dichas unidades se representan a sí mismas). Esto se debe ya sea a que no existe información auxiliar para corregir posibles omisiones o a que sus características las hacen únicas. Existen dos casos: las personas en situación de calle y las personas que residen en viviendas colectivas grandes (e.g. centros penitenciarios).

Las primeras provienen del relevamiento de personas en situación de calle realizado por el Ministerio de Desarrollo Social (MIDES) en el año 2023. Si bien habitualmente este operativo se denomina "censo de calle", estrictamente no constituye un censo en el mismo sentido que el Censo de Población y Vivienda, sino un relevamiento exhaustivo realizado en un momento determinado. Debido a la ausencia de información auxiliar y a las particularidades de esta población, no se realizan ajustes por omisión sobre estas personas.

En el caso de las personas que residen en viviendas colectivas, se fijó un umbral, basado en la distribución del tamaño de estos establecimientos, la cual se presenta en el siguiente gráfico y muestra una clara asimetría positiva, a partir del cual las personas residentes en viviendas colectivas con 50 personas o más reciben un ponderador igual a 1.

**Figura 12: Distribución del número de personas residentes en viviendas colectivas**




Esto se debe a que estos establecimientos son relativamente pocos, concentran una gran cantidad de personas y presentan características particulares que hacen poco razonable que representen a otras unidades más allá de sí mismos. En otras palabras, se considera más apropiado que estas personas se representen únicamente a sí mismas, en lugar de asignarles ponderadores mayores a 1.

A su vez, es importante tener en cuenta que, previo a este paso, se detectó que gran parte de los establecimientos penitenciarios asignados por vía administrativa se encontraban georreferenciados de forma incorrecta. Esto implicó que algunas de las personas relevadas fueran clasificadas erróneamente entre áreas urbanas y rurales, por lo que fue necesario corregir previamente dicha localización.

### 4.7.3 Unidades con ponderadores menor a uno.

La calibración utilizando una distancia chi cuadrado, o de forma equivalente el estimador *MB* basado en un modelo de regresión lineal, puede dar lugar a ponderadores menores a uno. Si bien esto puede resultar contraintuitivo en la práctica, ya que implica que existan unidades, por ejemplo personas, que "no se representan ni a sí mismas", ello no afecta las propiedades estadísticas de los estimadores.

En el caso de los ponderadores finales del censo, existe una proporción importante de valores menores a uno. Esto se debe a que el ajuste por no respuesta, orientado a corregir sesgos territoriales y socioeconómicos, permitía estimar de forma adecuada el tamaño total de la población, pero no lograba corregir los sesgos asociados a la edad y al sexo, tal como se mostró en la Sección 4.5.2. Por lo tanto, aquellos grupos de edad y sexo que se encontraban sobrerepresentados debían recibir, en promedio, ponderadores menores a uno para reproducir correctamente la estructura poblacional y reducir la parte del sesgo explicada por estas variables. Si bien en encuestas de hogares, como por ejemplo la ECH, también pueden aparecer ponderadores menores a uno, esto suele ocurrir en una proporción muy

reducida. En esos casos, a veces se les asigna posteriormente un valor igual a uno por razones prácticas o estéticas, más que por cuestiones teóricas.

### 4.7.4 Estimador *Hayek* a nivel de localidad

Previo a la calibración de los ponderadores, estos fueron ajustados utilizando el *estimador de Hayek*, considerando como dominios a las localidades. Este estimador, al basarse en una razón, es técnicamente preferible en contextos donde los tamaños de los dominios son aleatorios o presentan problemas de cobertura local. Esto se debe a que los tamaños de localidad previamente publicados presentaban errores, asociados tanto al algoritmo utilizado para incorporar personas mediante registros administrativos, como a problemas de georreferenciación en viviendas colectivas grandes. En particular, el algoritmo de asignación de personas vía RRAA presentaba un sesgo de urbanicidad, en el sentido de que las localidades pequeñas y las áreas rurales tendían a quedar subrepresentadas y de que el censo presentaba niveles de omisión relativamente más elevados en estos territorios que en las localidades de mayor tamaño.

A lo anterior, se sumó el problema de algunas viviendas colectivas grandes, especialmente establecimientos penitenciarios, que se encontraban georreferenciadas de forma incorrecta, tal como se señaló en la sección anterior. Independientemente del método utilizado, ya sea el censo combinado o el censo ponderado, es importante tener presente que la cantidad de personas por localidad constituye una estimación y no una medida exacta o definitiva . Por este motivo, el INE implementó el proceso de ponderación en dos etapas.

En primer lugar, una vez corregidos los microdatos correspondientes a las viviendas colectivas georreferenciadas erróneamente, se realizó la ponderación mediante el estimador doblemente robusto, tal cual fue descrito en las secciones anteriores. Posteriormente, las estimaciones obtenidas de los tamaños de localidad se compararon con los resultados del censo combinado, así como con distintas fuentes externas, tales como la cantidad de viviendas relevadas en el censo anterior, el número de contadores residenciales según UTE y otros indicadores territoriales disponibles. Este proceso de validación mediante indicadores auxiliares externos es una práctica estandarizada en la estimación de áreas pequeñas para corregir inconsistencias en niveles de alta desagregación (Rao & Molina, 2015).

A partir de esta comparación, aquellas localidades cuyas estimaciones se apartaban significativamente de las fuentes externas o resultaban poco coherentes respecto a lo esperable, recibieron un ajuste leve adicional en los ponderadores corregidos por no respuesta. Para ello, se aplicó un factor de ajuste definido como el cociente entre un tamaño de localidad de referencia y el tamaño estimado utilizando únicamente los ponderadores ajustados por no respuesta. Este procedimiento constituye un ajuste de tipo sintético (Brackstone, 1987), que permite estabilizar las estimaciones locales

El tamaño de referencia de cada localidad se construyó combinando la estimación obtenida a partir de la ponderación con la información proveniente de fuentes externas. En este proceso no se permitieron variaciones excesivas, ni en términos absolutos ni relativos, respecto a los valores observados en el Censo 2011.

### 4.7.5 Controles de calidad y validación de los ponderadores

La construcción de los ponderadores descrita en las secciones anteriores fue acompañada por un conjunto de controles de calidad y procedimientos de validación destinados a asegurar la coherencia interna de los resultados y la consistencia de las estimaciones obtenidas a partir de la base de microdatos ponderada. Estos controles se realizaron en distintas etapas del proceso, incluyendo la estimación de la probabilidad de respuesta, el ajuste por no respuesta dentro de cada segmento y la calibración a totales poblacionales. El objetivo de esta etapa es verificar que los ponderadores finales logren reducir el sesgo asociado a la omisión censal.

Como primer paso, se analizó la distribución de los ponderadores finales con el fin de detectar valores extremos o inestabilidades que pudieran afectar la calidad de las estimaciones. Se evaluaron, entre otros, los siguientes indicadores:

- Efecto de diseño de Kish
- media de los ponderadores
- desviación estándar
- valores máximos y mínimos
- percentiles de la distribución
- variación dentro de cada segmento censal

La presencia de ponderadores excesivamente grandes puede indicar problemas en la estimación de la probabilidad de respuesta o en la etapa de calibración. En caso de detectarse valores extremos, se revisaron los segmentos correspondientes y se verificó la consistencia de la información auxiliar utilizada. Este análisis permitió confirmar que los ponderadores obtenidos presentan una distribución estable y coherente con los niveles de omisión observados. A su vez, se verificó que las estimaciones obtenidas a partir de la base ponderada reproduzcan exactamente los totales utilizados en la calibración. En particular, se comprobó la consistencia de los resultados para:

- población total
- población por sexo
- población por edades simples
- población por departamento

Este control garantiza que la base de microdatos ponderada sea coherente con la estructura demográfica proveniente del censo combinado.

# 5. Evaluación del impacto en indicadores censales

Con el objetivo de evaluar los efectos de la ponderación sobre la estructura de la población censada, se compararon distintos indicadores sociales y demográficos antes y después de aplicar la ponderación. Este análisis permite verificar si la ponderación corrige de manera coherente los sesgos asociados a la omisión censal diferencial. En particular, se consideraron indicadores vinculados a situaciones de vulnerabilidad socioeconómica, características territoriales y composición demográfica, entre ellos: necesidades básicas insatisfechas, residencia en asentamientos, nivel educativo, estructura por edad, ascendencia afrodescendiente y población rural.

En términos generales, los resultados muestran que la ponderación incrementa la representación de grupos que presentaban mayores niveles de omisión en el censo. Esto ocurre especialmente entre personas que residen en contextos de mayor vulnerabilidad, así como en determinados grupos de edad y sexo cuya propensión a responder fue menor. El siguiente cuadro resume algunos ejemplos a modo ilustrativo para mostrar el efecto de la ponderación en la estimación de los indicadores. Además de mostrar las diferencias entre las estimaciones utilizando los datos del censo combinado y ponderados, se incorpora una estimación de qué proporción de la población omitida pertenece a cada grupo. Este último indicador permite dimensionar cuánto incide cada característica en la omisión total estimada.

***Cuadro 1. Comparación de algunos indicadores según censo combinado, censo ponderado, y estimación de su participación en la población omitida***

| | **Censo Combinado** | | **Censo Ponderado** | | **Población omitida (estimación)** | |
|---|---|---|---|---|---|---|
| **Indicador** | **Cantidad personas** | **%** | **Cantidad personas** | **%** | **Diferencia Absoluta** | **%** |
| *Materialidad* | 534.561 | 17,3 | 622.165 | 18,1 | 87.604 | 24,3 |
| *Hacinamiento (más de 2 personas por habitación para dormir)* | 416.066 | 13,5 | 489.576 | 14,3 | 73.510 | 20,4 |
| *Hacinamiento (más de 3 personas por habitación para dormir)* | 113.055 | 3,7 | 134.824 | 3,9 | 21.769 | 6,0 |
| *Espacio apropiado para cocinar* | 63.359 | 2,1 | 74.809 | 2,2 | 11.450 | 3,2 |
| *Calefacción* | 364.408 | 11,8 | 417.051 | 12,2 | 52.643 | 14,6 |
| *Conservación de alimentos* | 109.191 | 3,5 | 126.913 | 3,7 | 17.722 | 4,9 |
| *Calentador de agua para el baño* | 196.666 | 6,4 | 231.872 | 6,8 | 35.206 | 9,7 |
| *Abastecimiento agua potable/Origen y llegada de agua a la vivienda* | 135.124 | 4,4 | 156.892 | 4,6 | 21.768 | 6,0 |
| *Servicio higiénico/Acceso y calidad del servicio higiénico (saneamiento)* | 142.443 | 4,6 | 167.657 | 4,9 | 25.214 | 7,0 |
| *Residencia en asentamiento* | 158.727 | 4,5 | 193.260 | 5,5 | 34.533 | 9,6 |
| *Población rural* | 142.745 | 4,1 | 174.765 | 5 | 32.020 | 9,2 |

# 6. Medidas de la calidad de las estimaciones

Bajo el paradigma de inferencia basado en el diseño muestral (i.e. utilizando muestras aleatorias), la calidad de la estimación de un indicador se cuantifica mediante el error estándar (*SE*), el cual mide la variabilidad de las estimaciones entre muestra y muestra. El SE depende, entre otros factores, del diseño muestral y del método de estimación utilizado. En cambio, cuando se trabaja con muestras no probabilísticas o de voluntarios, como ocurre con los hogares efectivamente censados, las estimaciones pueden presentar sesgo debido a que la probabilidad de inclusión no es conocida ni controlada por diseño. En este contexto, las metodologías utilizadas (como los estimadores *DR*) buscan reducir este sesgo mediante modelos de respuesta y calibración. Por este motivo, la calidad de una estimación se evalúa conceptualmente a través del error cuadrático medio (*ECM*), que combina tanto la variabilidad como el sesgo residual del estimador.

$$ECM(\hat{\theta}) = \mathrm{SE}^2(\hat{\theta}) + \mathrm{Sesgo}(\hat{\theta})^2,$$

En la práctica, el sesgo remanente es desconocido y muy difícil de estimar. Por ello, las medidas de precisión suelen construirse únicamente a partir de la varianza (SE al cuadrado) estimada del estimador, bajo el supuesto de que el sesgo residual representa una fracción relativamente pequeña del error cuadrático medio. El desafío consiste en obtener una aproximación razonable del SE para calcular medidas de calidad como el coeficiente de variación (CV) e intervalos de confianza.

En aplicaciones habituales, los SEs de estimadores DR suelen calcularse mediante métodos de remuestreo, tales como Bootstrap o Jackknife, con el objetivo de capturar todas las fuentes de variabilidad involucradas en la estimación, incluyendo la asociada al modelo de respuesta y al modelo superpoblacional. Sin embargo, en el caso del Censo 2023 existen dos dificultades principales. En primer lugar, el ajuste por no respuesta se basa en indicadores operativos observados a nivel de segmento (como la tasa de enlace CAWI), los cuales actúan como parámetros de ajuste predefinidos en el proceso. Al no requerir una reestimación iterativa de modelos probabilísticos complejos, la variabilidad asociada a esta etapa presenta una estabilidad que hace que los métodos de remuestreo no aporten ganancias significativas en la captura de varianza adicional.

En segundo lugar, el tamaño de la base de datos, superior a los tres millones de registros, hace que los métodos de remuestreo resulten altamente demandantes tanto para su generación como para su posterior utilización por parte de los usuarios. Por este motivo, para aproximar la varianza de los estimadores se adopta un enfoque basado en un diseño muestral equivalente, construido de forma de reflejar las principales fuentes de variabilidad presentes en los datos. En particular, se consideran los siguientes elementos:

- La unidad de conglomeración se define a nivel de hogar.
- Dado que la tasa de cobertura censal (en término de muestreo) es muy alta, se incorpora la corrección por población finita.
- El ajuste por no respuesta se realizó a nivel de segmento censal, los cuales presentan distintas tasas de respuesta.
- La diferencia entre ponderadores agregan una fuente extra de variabilidad .

Teniendo en cuenta estas fuentes, se adopta un diseño equivalente de muestreo por conglomerados, estratificado a nivel de segmento censal y con probabilidades de selección distintas. Este enfoque

permite obtener una aproximación razonable de la varianza de los estimadores finales. El método anterior es conocido como el método del último conglomerado (Hansen, et al., 1953). Este procedimiento es ampliamente utilizado en la práctica estadística para simplificar el cálculo de varianza en diseños complejos, permitiendo que la variabilidad entre las unidades primarias de muestreo (UPM) capture la varianza total del diseño (Wolter, 2007).

Por lo tanto, la estimación de la varianza ($\mathrm{SE}^2$) del total de una variable cualquiera viene dada como:

$$\mathrm{SE}^2(\hat{Y}) = \sum_{c=1}^{C} \frac{m_c(1-\hat{\phi}_c)}{m_c(m_c-1)} \sum_{i \in B_c} (\hat{Y}_{ci} - \hat{\bar{Y}}_c)^2$$

donde:

$\hat{Y} =$ es la estimación del total de la variable $Y$

$m_c =$ es la cantidad de hogares censados pertenecientes al segmento $c$

$B_c =$ indica la muestra no probabilistica (i.e. los contados) pertenecientes al segmento c

$\hat{\phi}_c =$ es la tasa de muestreo estimada en el segmento $c$

$w_i$ es el ponderador de la unidad (hogar) $i$ y

$$\hat{Y}_{ci} = \sum_{k \in B_{ci}} w_{ik} \times y_{ik}$$

donde $y_{ik}$ es el valor que toma la variable $y$ para la persona $k$ perteneciente al hogar $i$

y

$$\hat{\bar{Y}}_c = m_c^{-1} \sum_{i \in B_c} \hat{Y}_{ci}$$

## 6.1. Intervalos de confianza

El error estándar (SE) es el insumo fundamental para la construcción de los Intervalos de Confianza (IC), los cuales permiten cuantificar la incertidumbre asociada a una estimación puntual y realizar inferencias sobre los parámetros poblacionales desconocidos. Un IC consiste en un rango de valores, construido alrededor de la estimación puntual del parámetro, dentro del cual se "espera" que se encuentre el verdadero valor con un determinado nivel de confianza o seguridad.

La amplitud de un IC está determinada por dos componentes: la precisión de la estimación (cuantificada por medio del SE) y el nivel de confianza elegido (usualmente 66%, 95% o 99%). Técnicamente, el intervalo se calcula multiplicando el SE por un valor crítico *z* derivado de la distribución de probabilidad asumida (habitualmente la distribución normal).

A modo de ejemplo ilustrativo, consideremos la estimación del porcentaje de hogares con al menos una Necesidad Básica Insatisfecha (NBI). Supongamos que la estimación puntual para una localidad dada se situó en 30%, con un SE de 0,8%.

- Un IC al 66% para la verdadera proporción de hogares con NBI sería:

$$30\% \pm 0,8\%.$$

  es decir, entre 29,2% y 30,8%.

- Un IC al 95% sería:

$$30\% \pm 1,96 \times 0,8\%$$

$$30\% \pm 1,6\%,$$

  lo que sitúa el rango entre 28,4% y 31,6%.

Bajo un enfoque frecuentista, la interpretación correcta de este último intervalo implica que si el proceso de estimación se repitiera bajo las mismas condiciones sobre un gran número de muestras independientes, aproximadamente el 95% de los intervalos así construidos contendrían el verdadero valor poblacional de la incidencia de NBI (el cual es desconocido). Este ejemplo permite visualizar que, cuanto mayor es el nivel de confianza exigido (del 66% al 95%), más amplio debe ser el intervalo para asegurar que el parámetro poblacional se encuentre contenido en él. En este sentido, un intervalo más angosto indica una mayor precisión en la estimación del indicador.

## 6.2. Coeficiente de Variación

El coeficiente de variación (*CV*) es una medida de precisión relativa que permite comparar la calidad de estimaciones que tienen diferentes unidades de medida o magnitudes. El *CV* se calcula como el cociente entre el error estándar (SE) y la estimación puntual, y usualmente se expresa como un porcentaje:

$$CV = \frac{SE}{\text{estimación}} \times 100$$

Mientras más pequeño es el *CV*, mayor es la precisión de la estimación. La principal ventaja del *CV* es que permite entender la magnitud del error en relación con lo que se está midiendo. Por ejemplo, un *SE* de 1,000 personas puede ser insignificante para estimar la población total del país, pero crítico para estimar la población de una localidad pequeña.

Es importante tener en cuenta que el *CV* presenta problemas cuando la estimación es muy cercana a cero (por ejemplo, en indicadores de baja prevalencia). En estos casos, el *CV* tiende a "inflarse artificialmente", perdiendo su utilidad como indicador de calidad. Por este motivo, para indicadores con valores pequeños, se recomienda evaluar la precisión de forma integral, observando tanto el *CV* como el error estándar (SE) y la amplitud del intervalo de confianza.

**Escala de valoración de la calidad**

Siguiendo los estándares institucionales del INE, se utiliza la siguiente escala de valoración para definir si una estimación es lo suficientemente estable para ser publicada:

- Excelente si el $CV < 5\%$. La estimación posee un nivel de precisión óptimo.
- Muy bueno si $5\% \leq CV < 10\%$. La estimación es altamente confiable.
- Bueno si $10\% \leq CV < 15\%$. La estimación es sólida para análisis generales.
- Aceptable: si $15\% \leq CV < 25\%$. La estimación tiene una precisión moderada.
- Utilizar con precaución: si $25\% \leq CV < 35\%$ La estimación presenta una variabilidad importante; se recomienda su uso solo con fines descriptivos y mencionando su inestabilidad.
- No publicar: si $CV \geq 35\%$ La estimación se considera estadísticamente inestable y su publicación podría inducir a conclusiones erróneas.

# 7. Supuestos y limitaciones del enfoque de ponderación

La estrategia de ponderación adoptada para el Censo 2023 se basa en un conjunto de supuestos estadísticos que permiten corregir, al menos parcialmente, el sesgo asociado a la omisión censal diferencial. Como en todo procedimiento de ajuste basado en modelos, la validez de las estimaciones obtenidas depende del grado en que dichos supuestos se aproximen a la realidad.

El objetivo del procedimiento no es eliminar completamente el error asociado a la omisión, sino reducirlo en la medida de lo posible, utilizando la información auxiliar disponible y manteniendo la coherencia interna de los datos censales. Por lo tanto, para realizar inferencias válidas es necesario asumir una serie de condiciones sobre los mecanismos de respuesta y las variables asociadas a la omisión. Estos supuestos se presentan de forma explícita a continuación, con el fin de transparentar el alcance, las limitaciones y las condiciones bajo las cuales deben interpretarse los resultados obtenidos.

## 7.1 Supuestos del modelo de respuesta

El ajuste por no respuesta se basa en la estimación de la probabilidad o propensión de ser censado. Para ello, se asume que dicha probabilidad depende únicamente de las variables auxiliares incorporadas en el modelo.

En particular, se consideró que, una vez tenidas en cuenta las variables disponibles, las probabilidades de ser censado dentro de un mismo segmento censal y entre hogares con características similares pueden considerarse homogéneas. Este supuesto, conocido como ignorabilidad condicional, implica que, dado el conjunto de variables auxiliares utilizadas, la omisión censal puede considerarse independiente de las variables sustantivas o “core” del censo.

En la práctica, este supuesto puede no cumplirse de forma exacta, ya que pueden existir factores no observados que influyan simultáneamente tanto en la probabilidad de ser censado como en las variables de interés. Asimismo, dentro de un mismo segmento censal pueden persistir diferencias no captadas completamente por la información disponible. En estos casos, un modelo de respuesta incorrectamente especificado podría no eliminar totalmente el sesgo asociado a la omisión.

No obstante, el uso del segmento como dominio de corrección permite reducir de forma importante la heterogeneidad territorial en las probabilidades de respuesta. Además, el procedimiento aplicado no depende exclusivamente del modelo de respuesta: los ponderadores iniciales obtenidos a partir de la propensión estimada son posteriormente ajustados mediante una etapa de calibración sobre totales externos, lo que contribuye a reducir posibles errores residuales y mejora la consistencia general de las estimaciones.

## 7.2. Supuestos del modelo superpoblacional implícito en la calibración

La calibración de los ponderadores a totales poblacionales conocidos puede interpretarse como la utilización implícita de un modelo *superpoblacional* o *working model* para las variables de interés. Bajo este enfoque, se asume que dichas variables están relacionadas con las variables auxiliares utilizadas en la calibración.

En particular, se supone que la variable de interés puede explicarse, al menos parcialmente, a partir de variables como la edad, el sexo y el departamento de residencia. Es decir, se asume que estas variables contienen información relevante para explicar parte de la variabilidad observada en las principales variables censales. En la medida en que estas variables auxiliares estén efectivamente asociadas a las variables de interés, la calibración permite reducir el sesgo remanente incluso cuando el modelo de respuesta no está completamente bien especificado. En ese sentido, la calibración actúa como un mecanismo adicional de corrección sobre los ponderadores obtenidos mediante el ajuste por propensión.

Asimismo, se observó que existen grupos demográficos que se encuentran subrepresentados o sobrerrepresentados en la población efectivamente censada. Al forzar la consistencia de los ponderadores con distribuciones conocidas por edad, sexo y territorio, la calibración contribuye a corregir parte de esas diferencias sistemáticas. Sin embargo, si las variables auxiliares utilizadas no capturan adecuadamente las diferencias entre la población censada y la omitida, el ajuste puede resultar incompleto. Por el contrario, cuanto mayor sea la relación entre las variables auxiliares y las variables de interés, mayor será la reducción esperada tanto del sesgo como de la varianza de las estimaciones.

## 7.3. Información auxiliar

El procedimiento de ponderación depende en gran medida de la calidad de la información auxiliar utilizada tanto en la modelización de la respuesta como en la etapa de calibración.

En la modelización de la  propensión de responder, se utilizaron variables asociadas a características demográficas, territoriales y operativas del relevamiento. Se utiliza el segmento censal, la modalidad de relevamiento (CAWI o CAPI) y la tasa de enlace de cuestionarios web en cada segmento. En particular, la tasa de enlace de cuestionarios web se utilizó como un indicador indirecto de la probabilidad de contacto efectivo y de las dificultades operativas observadas durante el operativo censal. Bajo este enfoque, se asume que los segmentos con menores tasas de enlace presentan mayores niveles de omisión. Esta variable permite capturar heterogeneidad territorial en la respuesta que no necesariamente queda reflejada únicamente a través de variables demográficas tradicionales. No obstante, esta aproximación también presenta limitaciones. La tasa de enlace puede estar influida por factores distintos a la omisión censal, como diferencias en el uso de internet, en la disponibilidad de dispositivos o en la preferencia por responder mediante CAWI o CAPI. Asimismo, el propio proceso de enlace entre cuestionarios puede contener errores o no reflejar completamente la intensidad real de las dificultades operativas en todos los segmentos.

Por otra parte, la calibración utiliza totales poblacionales obtenidos a partir del censo combinado, el cual integra información censal con registros administrativos. Este enfoque permite mejorar la estimación del tamaño y de la estructura de la población, pero también introduce ciertas limitaciones. La cobertura de los registros administrativos puede variar según el nivel de interacción de las personas con el Estado. Como consecuencia, si bien los totales demográficos utilizados se consideran confiables, no es posible garantizar que dicha información esté completamente libre de errores. En caso de que los totales de referencia utilizados para la calibración presenten errores de cobertura o clasificación, estos pueden trasladarse a los ponderadores y generar sesgos en las estimaciones finales.

En términos generales, errores en las hipótesis del modelo de respuesta *MRTE*, en la medición de las variables auxiliares, en los registros administrativos o en los totales poblacionales de referencia pueden afectar la precisión de los ponderadores y, en consecuencia, de las estimaciones obtenidas.

## 7.4. Conclusiones

La ponderación aplicada al Censo 2023 permite reducir el sesgo asociado a la omisión diferencial y producir estimaciones más consistentes con la estructura real de la población. No obstante, como en cualquier procedimiento basado en modelos, los resultados deben interpretarse teniendo en cuenta los supuestos y limitaciones descritos.

En particular, el enfoque adoptado reconoce explícitamente que la omisión censal no es aleatoria, sino que se concentra en determinados grupos poblacionales y territorios, lo que implica que el uso directo de los datos censales sin ajuste alguno conduce a estimaciones sistemáticamente sesgadas. En este contexto, la ponderación no constituye un complemento opcional, sino un componente esencial para la producción de inferencias válidas a partir de los microdatos censales.

El uso de un enfoque *doblemente robusto* permite reducir el sesgo bajo un conjunto amplio de escenarios, en la medida en que al menos uno de los dos modelos (el modelo de respuesta o el modelo superpoblacional implícito en la calibración) esté correctamente especificado. Esta propiedad otorga mayor robustez al procedimiento en comparación con enfoques tradicionales basados exclusivamente en imputación o en ajustes univariados.

Por otra parte, la integración de información auxiliar proveniente del censo combinado, en particular a través de la calibración a totales poblacionales por edad, sexo y departamento, permite alinear las estimaciones con distribuciones demográficas consideradas confiables. No obstante, la efectividad de este ajuste depende de la calidad de dicha información y de su capacidad para capturar las diferencias entre la población censada y la omitida.

Es importante destacar que la ponderación no elimina completamente todas las fuentes de error. Pueden persistir sesgos asociados a variables no observadas o a factores no capturados por las variables auxiliares utilizadas en los modelos. Del mismo modo, errores en los registros administrativos o en los totales de calibración pueden trasladarse a las estimaciones finales.

Desde el punto de vista del usuario, esto implica que los microdatos censales ponderados deben ser utilizados como la base principal para el análisis, evitando el uso de los datos sin ponderar para la producción de indicadores poblacionales. Asimismo, las estimaciones deben interpretarse en el marco del diseño metodológico adoptado, considerando tanto las mejoras introducidas por la ponderación como las limitaciones inherentes al proceso.

Finalmente, la estrategia implementada en el Censo 2023 refleja una adaptación metodológica a los desafíos crecientes que enfrentan los censos modernos en materia de cobertura. En este sentido, el uso de técnicas avanzadas de inferencia originalmente diseñadas para el tratamiento de muestras no probabilísticas o de voluntarios, junto con la integración de múltiples fuentes de información auxiliar, constituye una línea de trabajo clave para el desarrollo futuro de los sistemas estadísticos oficiales.

# 8. Guía de uso de microdatos y cálculo de varianza en R

## 8.1. Conceptos básicos para la estimación con datos ponderados

El procesamiento de los microdatos del Censo 2023 requiere que el usuario integre dos dimensiones: la estructura operativa de las tablas y el diseño estadístico de la estimación. Para una descripción detallada de la arquitectura de la base (variables, diccionarios e identificadores), se recomienda consultar el documento Aspectos conceptuales y metodológicos para uso de los microdatos censales (INE, 2026). El presente capítulo complementa dicha guía, enfocándose en la lógica de los ponderadores y el cálculo de la precisión de los indicadores.

### 8.1.1. Interpretación de los ponderadores

A diferencia de un censo tradicional donde cada registro representa exactamente a una persona, en el Censo 2023 cada hogar y persona tiene asignado un ponderador (o peso) único. Es fundamental que las y los usuarios comprendan la naturaleza técnica de estos valores antes de procesar los microdatos.

El ponderador indica a cuántas unidades de la población total representa cada registro de la base de microdatos. Debido a los procesos de ajuste por no respuesta y calibración mediante distancias chi-cuadrado (o estimadores de regresión), estos valores presentan dos propiedades particulares:

1. Valores con decimales: Los ponderadores son el resultado de la aplicación de modelos de propensión y calibración, por lo que no son números enteros. Esto implica que las sumas de personas o características (e.g. total de personas con nivel educativo terciario) resultarán habitualmente en valores decimales (e.g. 1,240.7 personas). Estos decimales no indican *"fracciones de personas"*, sino que son el residuo matemático de distribuir el peso de la omisión entre los hogares efectivamente censados.

2. Ponderadores menores a uno: es frecuente encontrar una proporción importante de registros con pesos menores a la unidad (e.g. 0.85). Si bien esto puede resultar contraintuitivo, ya que implica que una unidad "no se representa íntegramente a sí misma", ello no afecta las propiedades estadísticas de los estimadores. Esta situación ocurre porque ciertos grupos (por ejemplo, determinados tramos de edad y sexo) quedaron sobrerrepresentados en el operativo de campo. Para corregir este sesgo y reproducir fielmente la estructura poblacional, el proceso de calibración debe asignar a estos grupos ponderadores menores a la unidad.

Un fenómeno derivado de lo anterior es la existencia de segmentos censales donde la población ponderada (expandida) resulta inferior a la cantidad de personas efectivamente contadas. Aunque esto puede parecer un error de procesamiento, es un resultado estadísticamente consistente con el método de calibración adoptado.

Este escenario ocurre bajo la combinación de dos factores:

1. **Baja omisión territorial:** El segmento presentó una tasa de respuesta muy alta, por lo que el ajuste por no respuesta es prácticamente despreciable.

2. **Efecto de la calibración demográfica:** El segmento está compuesto mayoritariamente por grupos de edad y sexo que, a nivel general, se encontraban sobrerrepresentados en el operativo de campo.

En estos casos, el ponderador calibrado (menor a 1) tiene un impacto superior al factor de no respuesta, resultando en un ponderador final que "comprime" el conteo observado para que la estructura demográfica sea coherente con los totales poblacionales oficiales.

Es importante destacar que el uso de estos ponderadores es teóricamente preferible y metodológicamente necesario. Las estimaciones expandidas, aun cuando resulten inferiores al conteo directo en ciertos segmentos, poseen un menor sesgo que los datos crudos, ya que han sido ajustadas para reflejar la verdadera estructura de la población. Ignorar estos pesos y optar por el conteo directo no solo mantiene los sesgos de sobrerrepresentación, sino que generaría un problema de benchmarking insalvable: la suma de los datos de estos dominios no coincidiría con los totales poblacionales de referencia. Por lo tanto, el uso del ponderador garantiza que el análisis sea estadísticamente válido y plenamente coherente con el marco demográfico del Censo 2023.

El sistema de ponderación desarrollado provee la base necesaria para la mayoría de los análisis sociodemográficos del país. Sin embargo, en investigaciones que requieran aperturas territoriales muy pequeñas o temáticas de muy baja prevalencia, la ponderación puede no ser suficiente para brindar estimaciones estables. En eos casos, se recomienda a las y los usuarios evaluar la precisión de las estimaciones mediante el coeficiente de Variación (CV) y, de ser necesario, recurrir a niveles de agregación geográfica o temática superiores para garantizar la robustez de las conclusiones.

### 8.1.2. Consecuencias en el procesamiento y redondeo

Estas características exigen un manejo cuidadoso de los datos para garantizar la coherencia interna y la consistencia con los totales oficiales del INE:

- **No redondear antes de tiempo:** Todos los cálculos deben realizarse utilizando el valor completo del ponderador con todos sus decimales. Redondear los pesos individualmente antes de realizar las sumas destruiría la precisión del ajuste y la coherencia demográfica.
- **Gestión de la visibilidad y preservación de totales:** Dado que existen ponderadores pequeños, un redondeo tradicional podría "invisibilizar" celdas con pocas observaciones. Por este motivo, el INE propone un criterio de redondeo específico (detallado en la sección siguiente) que asegura que cualquier categoría observada en la muestra tenga una representación mínima de 1 en las tablas finales, preservando simultáneamente la suma total de la población.

## 8.2. Procedimiento de redondeo y preservación de totales

Como se detalló anteriormente, el uso de ponderadores con decimales genera totales que no son enteros. Para la publicación de resultados y para asegurar la concordancia con los totales de control (*benchmarking*), se adopta un procedimiento de redondeo que garantiza que la suma de las categorías coincida exactamente con el redondeo de la suma total, evitando discrepancias en las tablas finales.

Para ello, el INE utiliza un algoritmo basado en el reparto de residuos mayores, al cual se le añade una restricción de "visibilidad estadística". Este criterio asegura que cualquier grupo observado en los microdatos mantenga su presencia en los resultados finales. La lógica del algoritmo es la siguiente:

1. **Garantía de representación mínima:** El primer paso del algoritmo fuerza que, si una categoría tiene al menos una observación en los microdatos (independientemente de que su peso ponderado sea menor a uno), el valor resultante sea al menos 1. Esto evita que grupos o categorías con muy baja prevalencia desaparezcan de las estimaciones por un efecto puramente matemático del redondeo.
2. **Ajuste por residuos:** Una vez asegurado el mínimo, el algoritmo calcula la diferencia entre la suma de los valores truncados y el total poblacional esperado. Esa diferencia (faltante) se distribuye sumando una unidad a aquellas categorías que presentaban los mayores residuos decimales antes del truncamiento.

Para aplicar este criterio de forma reproducible, se proporciona la siguiente función en R:

```r
redondear_censo <- function(x, digits = 0) {
  up <- 10 ^ digits
  x <- x * up
  y <- ifelse(x > 0 & x < 1, 1, floor(x))

  #Cálculo de la diferencia respecto al total esperado (benchmarking)
  diferencia <- round(sum(x)) - sum(y)

  # Distribución de la diferencia entre los residuos más altos
  if (diferencia > 0) {
    indices <- tail(order(x - y), diferencia)
    y[indices] <- y[indices] + 1
  }
  y / up

}
```

Este procedimiento garantiza que las tablas procesadas por las y los usuarios sean coherentes con los totales demográficos oficiales y que no se pierda información de celdas con baja frecuencia pero presencia real en el territorio.

La aplicación sistemática de este procedimiento de redondeo, junto con el uso de un ponderador único por hogar, garantiza la coherencia interna de los microdatos del Censo 2023. Al asegurar que la suma de las partes sea siempre igual al total y que ningún grupo observado quede excluido de las estimaciones, se logra que la información a nivel de hogares sea plenamente consistente con la información a nivel de personas. Este enfoque permite que se obtengan resultados precisos y validados que respetan las restricciones demográficas oficiales, proporcionando una base sólida y confiable para el análisis de políticas públicas y la investigación social en todo el territorio nacional.

## 8.3. Cómputo de los errores estándar en R

Para facilitar el uso de los microdatos del censo y garantizar que los usuarios y usuarias, puedan computar las precisiones las estimaciones bajo el método descrito en el capítulo 6. Medidas de la calidad de las estimaciones, se recomienda el uso del paquete `srvyr` (o `survey`) en el entorno R.

Dado el tamaño de la base de datos del censo (superior a los tres millones de registros) y la estructura de estratificación por segmento censal, el cómputo de varianza mediante el *método del último conglomerado* requiere una configuración específica para evitar errores de desbordamiento de memoria (*integer overflow*). Al intentar procesar un diseño complejo con tantos estratos y observaciones, `R` puede exceder el límite de elementos indexables ($2^{31}$). Por lo tanto, para procesar correctamente el diseño muestral equivalente del Censo 2023, se debe utilizar el siguiente bloque de código:

Como primer paso carga los paquetes/librerías necesarias:

```
library(srvyr)
library(survey)
```

Luego, se realiza la configuración técnica para grandes bases de datos. Para ello, se activa el método de último conglomerado y se evita la validación exhaustiva de estratos que causa errores de memoria en muestras masivas.

```
options(survey.lonely.psu = "remove",
        survey.ultimate.cluster = TRUE)
```

Se define el diseño muestral equivalente en donde lo SEs son computados utilizando el *método del último conglomerado* (ver 6. Medidas de la calidad de las estimaciones), donde `ID` es el Identificador del hogar (i.e. unidad de conglomeración), `codseg` es el segmento censal (i.e. el estrato), `TR` es es la tasa de muestreo para realizar la corrección por población finita a nivel de estrato al momento de computar los SEs y `w` es el ponderador.

```
censo_diseno <- microdatos |>
  as_survey_design(
    id = ID_HOGAR,
    strata = SEGMENTO,
    fpc = TR,
    weights = w
                  )
```

La opción `survey.ultimate.cluster = TRUE` permite que el cálculo se realice de forma eficiente, capturando la variabilidad entre las unidades primarias de muestreo (hogares) para aproximar la varianza total del diseño sin comprometer la capacidad de procesamiento del sistema.

Una vez cargado el diseño (`censo_diseno`) se realizan las estimaciones puntuales y se le pueden agregar medidas de calidad de las estimaciones. Por ejemplo, se puede calcular el total de una variable de interés y computar el error estándar (`se`), el coeficiente de variación (`cv`) y el intervalo de confianza (`ci`), el cual, por defecto es realizado al 95%.

```
estimacion <- censo_diseno |>
  summarise(
    total_indicador = survey_total(variable_interes,
                                   vartype = c("se",
                                               "cv",
                                               "ci"
                                               )
                                   )
          )
```

En el caso de que la variable sea categórica (e.g. sexo), se debe indicar que la variable es un factor, por ejemplo:

```
estimacion <- censo_diseno |>
  summarise(
    conteo_por_cateogoria = survey_total(as_factor(variable_interes),
                                         vartype = c("se",
                                                     "cv",
                                                     "ci"
                                                     )
                                         )
           )
```

Para obtener detalles adicionales sobre las distintas funciones de `srvyr`, el manejo de diseños complejos y la visualización de resultados bajo un enfoque reproducible, se recomienda consultar el libro (Zimmer, et. al., 2024), el cual proporciona una guía detallada para el procesamiento de microdatos utilizando la gramática de `tidyverse`, facilitando la transición desde el análisis estadístico tradicional hacia flujos de trabajo más eficientes en R.

# 9. Glosario de términos y conceptos metodológicos

Con el fin de unificar criterios sobre esta metodología y facilitar el manejo de los microdatos, a continuación se detallan los conceptos fundamentales utilizados en el proceso de ponderación

### Conceptos del Censo y Registros Administrativos

**Censo Combinado:** Enfoque censal que integra la enumeración tradicional de campo (entrevistas presenciales y web) con datos provenientes de registros administrativos para alcanzar una estimación del conteo total de la población.

**Enumeración Administrativa:** Proceso de incorporar personas a la base de datos censal a partir de registros oficiales (BPS, Salud, Educación, etc.) cuando estas no pudieron ser contactadas u omitieron su participación durante el operativo de campo.

**Señales de Vida (SDV):** Evidencia de actividad reciente de un individuo en los registros administrativos del Estado (empleo, salud, educación, etc.). Se utiliza como criterio técnico para validar la residencia habitual y decidir su inclusión en el conteo final del censo combinado.

**Falla de Vínculo (*Linkage Failure*):** Incapacidad técnica de asociar un registro administrativo de una persona con una vivienda o un hogar específico. Esto genera registros que "flotan" geográficamente a niveles agregados (e.g. localidad o departamento) y carecen de estructura de hogar definida.

**Sesgo de Visibilidad Institucional:** Distorsión en la captura de datos donde los estratos socioeconómicos más altos y con mayor interacción con el Estado son más propensos a figurar en los registros administrativos que los sectores vulnerables o informales.

**Sesgo de Urbanicidad:** Error sistemático de asignación geográfica que tiende a subrepresentar a las localidades pequeñas y áreas rurales en favor de los centros urbanos más densos debido a la mayor disponibilidad de señales o interacciones administrativas en estos últimos.

**Imputación por Donante (*Hot-Deck*):** Método de tratamiento de la no respuesta que consiste en completar datos faltantes asignando los valores de una unidad observada ("donante") con características similares a la unidad omitida.

### Inferencia y Modelización Estadística

**Ponderador (peso):** Valor numérico asignado a cada unidad observada (hogar o persona) que indica a cuántas unidades de la población objetivo representa.

**Propensión de Respuesta:** Probabilidad estimada de que un hogar participe efectivamente en el censo, condicionada por sus características territoriales y operativas.

**Estimador de probabilidad Inversa (IPW):** estimador que consiste en asignar a cada unidad respondente un ponderador igual al inverso de su propensión de respuesta , permitiendo que las unidades observadas representan a los omitidas.

**Estimador Doblemente Robusto (DR):** Método de estimación que combina un modelo de respuesta (propensiones) con un modelo de predicción (superpoblacional). Su principal propiedad es la consistencia: el estimador es válido si al menos uno de los dos modelos está correctamente especificado.

**Estimador Basado en Modelos (MB):** Enfoque que utiliza un modelo predictivo para estimar los totales poblacionales basándose en la relación entre variables auxiliares y la variable de interés observada en la muestra.

**Modelo de Trabajo (*Working Model*):** Especificación matemática (lineal, logística, etc.) elegida para representar un fenómeno. Se denomina "de trabajo" porque es la herramienta operativa que se asume como válida para realizar la inferencia.

**Modelo de Respuesta basado en la Tasa de Enlace (MRTE):** Modelo predictivo que utiliza la efectividad del censo web (CAWI) en un segmento como indicador de la dificultad de contacto territorial para corregir la omisión en dicho estrato.

**Calibración (*Benchmarking*):** Ajuste final de los ponderadores para asegurar que la suma de las estimaciones coincidan exactamente con totales poblacionales conocidos.

**Missing at Random (MAR):** Supuesto que establece que la probabilidad de que un dato falte (omisión) depende de variables observadas en la base de datos y no de los valores no observados de la propia variable de interés.

## Métricas de Calidad y Muestreo Complejo

**Error Cuadrático Medio (ECM):** Métrica de calidad total que suma la varianza y el cuadrado del sesgo.

**Error Estándar (SE):** Medida de la variabilidad de la estimación. Cuantifica cuánto varían las estimaciones entre muestra y muestra

**Sesgo de un Estimador:** Diferencia sistemática entre el valor esperado del estimador y el verdadero valor del parámetro poblacional.

**Coeficiente de Variación (CV)**: Medida de precisión relativa expresada como porcentaje.

**Efecto de Diseño de Kish** : Factor que mide cuánto aumenta el error estándar de una estimación debido al uso de ponderadores desiguales en comparación con un muestreo aleatorio simple.

**Factor de Corrección por Población Finita (FPC):** Ajuste que reduce la varianza estimada cuando la muestra representa una fracción muy significativa del universo total.

**Unidad Primaria de Muestreo (UPM):** Es la unidad básica que se selecciona en la primera etapa de un diseño de muestreo complejo. En el marco de esta metodología (diseño equivalente), la UPM se define a nivel de hogar. Es el componente fundamental para el cálculo de los errores estándar, ya que la variabilidad observada entre las UPM es la que permite aproximar la varianza total de las estimaciones finales.

**Método del Último Conglomerado**: Procedimiento para simplificar el cálculo de varianza en diseños complejos, utilizando la variabilidad entre las UPM (hogares) dentro de cada estrato (segmento) para capturar la varianza total.

# Abreviaturas

B Muestra no probabilistica
CAPI Entrevistas personales por computadora
CAWI Entrevistas web por computadora
CNV Certificado de Nacido Vivo
CV Coeficiente de variación
DR Estimador doblemente robusto
ECH Encuesta Continua de Hogares
ECM Error cuadrático medio
ENEC Encuesta Nacional de Evaluación de Cobertura
FPC Factor de Corrección por población finita
IC Intervalo de confianza
IPW Estimador inverso de probabilidad
MAR Missing at random
MB Estimador basado en modelo
MRTE Modelo de Respuesta Censal Basado en la Tasa de Enlace
ONE Oficina Nacional de Estadística
RRAA Registros administrativos
SE Error estándar
TNR Tasa de no respuesta
TR Tasa de respuesta
U Universo de estudio
UPM Unidad Primaria de Muestreo

# Bibliografía

Aboal, D. & Segui, F. (2025). *Uruguay's 2023 Combined Census: Integrating Administrative Registers and Traditional Enumeration*. Documento de Trabajo DT 03/2025, CINVE.

Alexander, C. H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, *13*(2), 183-198.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Report of the AAPOR task force on non-probability sampling.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–973.

Bates, N., & Mulry, M. H. (2011). *Nonresponse in the Census*. En J. Madans et al. (Eds.), Vital and Health Statistics. National Center for Health Statistics.

Bethlehem, J. (2002). *Weighting Nonresponse: Adjusting for Missing Data in Surveys*.

Bethlehem, J. (2020). Sampling and estimation in the pandemic era. *Journal of Survey Statistics and Methodology*, 8(1), 1–19.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. John Wiley & Sons.

Brackstone, R. J. (1987). Small area data from statistical surveys. En R. Platek, J. N. K. Rao, C. E. Särndal y M. P. Singh (Eds.), *Small Area Statistics: An International Symposium* (pp. 3-31). John Wiley & Sons.

Chambers, R., & Clark, R. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford: Oxford University Press.

Chen, Y., Li, P., & Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011–2021.

Chen, Y., Li, P., & Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48(2), 283–311.

Cobben, F. (2009). *Nonresponse in Sample Surveys: Methods for Analysis and Adjustment*. Heerlen: Statistics Netherlands.

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376–382.

Deville, J.-C., Särndal, C.-E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423), 1013–1020.

Elliott, M. R., & Valliant, R. (2017). *Inference for Nonprobability Samples*. Statistical Science.

Fellegi, I. P., & Holt, D. (1976). *"A systematic approach to automatic edit and imputation"*. Publicado en el *Journal of the American Statistical Association*.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. John Wiley & Sons.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. (2002). *Survey Nonresponse*. Wiley.

Groves, R. M. (2006). *Nonresponse rates and nonresponse bias in household surveys*. Public Opinion Quarterly, 70(5), 646–675.

Hand, D. J. (2006). *Classifier Technology and the Illusion of Progress*. Statistical Science.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample Survey Methods and Theory* (Vol. 1). John Wiley & Sons. (Es la fuente original que define el "Ultimate Cluster Method").

Hajek, J. (1971): *Comment on "An essay on the logical foundations of survey sampling"*.

Haziza, D., & Beaumont, J. F. (2017). *Construction of weights in surveys: A review*. Statistical Science

Instituto Nacional de Estadística. (2024e). *Metodología del Censo combinado 2023*. Documento no vigente; disponible para consulta como antecedente técnico.

Instituto Nacional de Estadística. (2026e). Aspectos conceptuales y metodológicos para uso de los microdatos censales. Documento vigente.

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523-539. doi.org

Kim, J. K., & Haziza, D. (2014). *"Doubly robust inference with missing data in survey sampling"*. *Statistica Sinica*.

Kim, J. K., & Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. CRC Press.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Lemaitre, G., & Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, *13*(2), 199-207.

Little, R. J. A. (1993). Post-stratification: A modeler’s perspective. *Journal of the American Statistical Association*, 88(423), 1001–1012.

Little, R. J., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3ra Ed.). Wiley.

Little, R. J., & Vartivarian, S. (2005). *Does weighting for nonresponse increase variance? Methods for weighting and covariate selection*.

Lundström, S., & Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15(2), 305–327.

Meng, X. L. (2018). *Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and Self-Correction Threshold*. The Annals of Applied Statistics.

Office for National Statistics. (2021). *Maximising the quality of the 2021 Census: The Census Coverage Survey*. UK Government.

Pfeffermann, D. (2013). *New important developments in small area estimation*. Statistical Science, 28(1), 40–68.

Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2nd ed.). John Wiley & Sons.

Reiter, J. P. (2007). *Small area estimation and correcting for nonresponse in the Census*. Journal of the American Statistical Association.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.

Särndal, C.-E., & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester: Wiley.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Seaman, S. R., & White, I. R. (2013). *Review of inverse probability weighting for dealing with missing data*. Statistical Methods in Medical Research.

Si, Y. (2020). On the use of auxiliary variables in multilevel regression and poststratification. *Journal of Survey Statistics and Methodology*, 10(2), 271–294.

Stats NZ. (2019). *2018 Census coverage estimation and methodology*. Statistics New Zealand.

United Nations. (2010). *Post Enumeration Surveys: Operational Guide*.

United Nations. (2017). Principles and recommendations for population and housing censuses (Revision 3). United Nations.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite Population Sampling and Estimation: A Prediction Approach*. New York: John Wiley & Sons.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2), 231–263.

Valliant, R., & Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1), 105–137.

Valliant, R., Dever, J. A., & Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples* (2nd ed.). New York: Springer.

Valliant, R., Dever, J. A., Kreuter, F., & Olson, K. (2018). General calibration and weighting methods for volunteer web samples. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, C. Tucker, & B. West (Eds.), *Total Survey Error in Practice* (pp. 375–398). Hoboken: Wiley.

Wallgren, A., & Wallgren, B. (2014). *Register-based Statistics: Administrative Data for Statistical Purposes*. Wiley.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3), 980–991.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. Springer Science & Business Media.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48(2), 283–311.

Zimmer, S. A., Powell, R. J., & Velásquez, I. C. (2024). Exploring Complex Survey Data Analysis Using R: A Tidy Introduction with {srvyr} and {survey}. Chapman & Hall: CRC Press.

Zhang, L. C. (2012). *Topics of analytical error in mixed-source statistics*.

Esta página ha sido dejada intencionalmente en blanco