Distr.
LIMITADA

LC/CEA.13/DDR/2
20 de noviembre de 2025

ORIGINAL: ESPAÑOL

2500760[S]

Decimotercera Reunión de la Conferencia Estadística de las Américas
de la Comisión Económica para América Latina y el Caribe

Santiago, 25 a 27 de noviembre de 2025

METODOLOGÍAS DE IMPUTACIÓN DE DATOS APLICADAS EN ENCUESTAS DE HOGARES Y CENSOS DE POBLACIÓN

NACIONES UNIDAS

CEPAL

Este documento fue realizado por el Grupo de Trabajo para la revisión de metodologías de imputación de datos aplicadas en encuestas de hogares y censos de población, de la Conferencia Estadística de las Américas (CEA) de la Comisión Económica para América Latina y el Caribe (CEPAL), bienio 2024-2025. El Grupo fue coordinado por Chile (Instituto Nacional de Estadísticas (INE)), y tuvo como secretaría técnica a la División de Estadísticas de la CEPAL. Los países miembros son: Argentina (Instituto Nacional de Estadística y Censos (INDEC)), Brasil (Instituto Brasileiro de Geografia e Estatística (IBGE)), Colombia (Departamento Administrativo Nacional de Estadística (DANE)), Costa Rica (Instituto Nacional de Estadística y Censos (INEC)), Cuba (Oficina Nacional de Estadística e Información (ONEI)), Ecuador (Instituto Nacional de Estadística y Censos (INEC)), Guatemala (Instituto Nacional de Estadística (INE)), Honduras (Instituto Nacional de Estadística (INE)), México (Instituto Nacional de Estadística y Geografía (INEGI)), Panamá (Instituto Nacional de Estadística y Censo (INEC)), Paraguay (Instituto Nacional de Estadística (INE)), Perú (Instituto Nacional de Estadística e Informática (INEI)), Uruguay (Instituto Nacional de Estadística (INE)), Venezuela (Instituto Nacional de Estadística (INE)) y la División de Población (CELADE) de la CEPAL.

ÍNDICE

Página

1. ANTECEDENTES

Las encuestas de hogares y los censos de población constituyen las principales fuentes de información estadística para las Oficinas Nacionales de Estadística (ONE) en América Latina y el Caribe. A través de estos instrumentos, se obtiene información demográfica y socioeconómica esencial para la formulación, monitoreo y evaluación de políticas públicas. Es por ello por lo que garantizar la calidad de estos datos es un desafío clave para las ONE, pues la información recolectada en encuestas y censos puede verse afectada por diversos factores, entre ellos la falta de respuesta total o parcial por parte de la población objetivo. A pesar de contar con una planificación meticulosa en la recolección de la información, la falta de respuesta total o parcial es una problemática recurrente en encuestas y censos y, la presencia de datos faltantes, derivada de la negativa de participación de los hogares o personas, dificultades en la localización de los encuestados o la omisión de respuestas a preguntas específicas, puede impactar la precisión de las estimaciones y generar sesgos en los resultados, afectando así, la validez de los análisis y la toma de decisiones basadas en estos datos.

La imputación de datos se ha consolidado como una herramienta metodológica crucial para abordar la falta de respuesta parcial, permitiendo completar los conjuntos de datos y mejorar la calidad de la información. Mientras que la no respuesta total de hogares suele tratarse mediante técnicas de ajuste de factores de expansión en encuestas de hogares, en censos de población y vivienda la práctica en la región ha combinado, según el contexto, (i) la estimación de la omisión y su ajuste correspondiente sin imputar la unidad completa y (ii) en algunos casos, la imputación acotada de no respuesta total cuando existen indicios de ocupación de la vivienda o riesgos de sesgos de cobertura estructural (p. ej., omisión de menores). En estos casos se han utilizado esquemas de donante cercano (*hot-deck*) dentro de áreas pequeñas para completar rasgos mínimos y coherentes (P. ej., tamaño del hogar y distribución por sexo y edad), manteniendo como “No especificado”, “No aplica”, etc., los atributos no observados; otras oficinas, en cambio, han registrado la no respuesta total como omisión (Omisión Censal). La correcta implementación de métodos de imputación contribuye a minimizar sesgos y aumentar la precisión de las estimaciones, asegurando que los datos resultantes sean representativos y confiables, en la medida que se disponga de información de hogares, viviendas y/o personas suficientes para sostener dichas inferencias.

En los últimos años, se han desarrollado diversas metodologías de imputación en el área de las estadísticas, incluyendo enfoques tradicionales como la eliminación de casos, imputación por la media, *hot-deck*, así como métodos más sofisticados como la imputación múltiple y técnicas basadas en modelos de aprendizaje automático. Sin embargo, la aplicación de estas metodologías no siempre se ha sustentado en marcos teóricos sólidos que consideren diseños de muestreos complejos ni ha considerado la evaluación sistemática de su impacto en las estimaciones finales.

El uso de metodologías de imputación adecuadas es particularmente relevante en América Latina y el Caribe debido a los desafíos estructurales que enfrentan las ONE en términos de cobertura, financiamiento y acceso a registros administrativos de calidad. La heterogeneidad de los sistemas estadísticos en la región requiere un enfoque flexible y adaptativo que permita la implementación de estrategias robustas y replicables en distintos países. Asimismo, la disponibilidad creciente de registros administrativos ofrece una oportunidad única para mejorar las imputaciones en encuestas y censos. Incorporar estos registros como fuentes auxiliares puede contribuir a una mayor precisión de las estimaciones de población, viviendas, ingreso, empleo, etc. Fortaleciendo la confiabilidad de los datos producidos por las instituciones.

En este contexto, este documento busca sistematizar y difundir prácticas metodológicas sobre imputación de datos en censos y encuestas de hogares, con un enfoque específico en América Latina

y el Caribe en el marco de los grupos de trabajo del bienio 2024-2025 de la Conferencia de Estadística de las Américas (CEA CEPAL). Por otra parte, se busca consolidar las mejores prácticas en materia de imputación de datos y proporcionar herramientas metodológicas actualizadas que faciliten su aplicación en censos y encuestas de hogares en la región. A través de un esfuerzo colaborativo entre las ONE de la región y la colaboración de CEPAL, se pretende establecer lineamientos estandarizados que puedan adaptarse a las realidades nacionales y mejorar la calidad de las estadísticas oficiales.

El desarrollo de estas recomendaciones metodológicas se apoya en estudios previos y documentos de referencia clave en la materia, así como en la experiencia de las ONE y sus estrategias implementadas. En última instancia, este trabajo busca fortalecer la capacidad de los institutos nacionales de estadística para producir información confiable y comparable, asegurando que los datos generados no solo sean más precisos y completos, sino que también se constituyan en una base sólida para la formulación de políticas públicas eficaces en América Latina y el Caribe.

2. DEFINICIÓN Y TIPOS DE DATOS FALTANTES

2.1 Concepto de Datos Faltantes

En el contexto de censos y encuestas de hogares, el concepto de datos faltantes hace referencia a valores ausentes en las bases de datos debido a diversos factores; estos pueden deberse a la negativa en hogares y/o personas a responder determinadas preguntas (Puede obedecer a razones tales como desconocimiento, sensibilidad del tema tratado, etc.), errores en la captura o recolección de la información y/o problemas operativos.

Para un adecuado tratamiento de los datos, es fundamental diferenciar entre los valores verdaderamente faltantes y aquellos que no son registrados debido al flujo del cuestionario correspondiente. Los datos faltantes, en el contexto del presente documento, corresponden a aquellos valores ausentes en una variable que debía ser respondida en el cuestionario. En cambio, la no respuesta debido a saltos de pregunta lógicos o estructurales, ocurren cuando una pregunta no aplica a un encuestado debido a la respuesta en la pregunta anterior (o la que corresponda), que provoca el no registro de la misma; por ejemplo, en el caso de las encuestas de empleo, aquellas personas que están desocupadas omitirán las preguntas acerca de ocupación, dado que el sistema debe registrar estos valores como “no aplicables” en lugar de considerarlos como datos faltantes.

Finalmente, los errores en la recolección de datos se refieren a aquellas fallas en el diseño del cuestionario, errores operativos en la captura o problemas técnicos que impiden registrar la información correctamente. La utilización de una codificación estandarizada permite diferenciar estos casos y es clave para la correcta interpretación de los datos y la aplicación de técnicas de imputación cuando sea necesario (Eltinge & Luery, 2003; Donza, 2013).

2.2 Clasificación de Datos Faltantes

La clasificación de los datos faltantes es fundamental para determinar el método más adecuado para la aplicación de técnicas de imputación. La ausencia de datos puede ocurrir por múltiples razones como se ha comentado en la sección 2.1. y su tipificación permite entender su impacto en la calidad de las estimaciones y en la representatividad de los resultados.

Debido a la naturaleza binaria de la respuesta (observada o no observada), es posible definir un proceso aleatorio $P(\cdot)$ para una variable aleatoria **R** que representa si los datos fueron o no

observados (**R** = 1, **R** = 0 respectivamente) el cuál puede ser denominado como "mecanismo de no respuesta"[1]. Dicho mecanismo, puede ser formulado como un modelo estadístico para **R**, dado un conjunto de datos **Y** de la forma:

$$P(\boldsymbol{R}|\boldsymbol{Y}, \psi)$$

Dicha probabilidad condicional, contiene un vector de parámetros desconocidos ψ del modelo formulado para **R**. (Little & Rubin, 2020a). De este modo, se tiene un modelo que establece una relación estrecha entre **R**, completamente observado e **Y**, el cuál puede descomponerse en un vector de datos observados y datos no observados, es decir, $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$.

La descomposición de **Y** permite clasificar el mecanismo de datos faltantes según si los valores faltantes están relacionados o no con los valores de **Y**. Little & Rubin (2013a, sec 1.3), Enders (2022) entre otros autores establecen tres tipos de mecanismos: Ausencia de datos completamente aleatoria (MCAR), ausencia de datos aleatoria (MAR) y ausencia de datos no aleatoria (MNAR). Esta diferenciación es de vital importancia puesto que funcionan como supuestos estadísticos para el modelamiento de los datos.

Ausencia de respuesta completamente aleatoria (MCAR): Cuando la probabilidad de no observar un dato (**R** = 0) es igual para todas las observaciones, entonces, la ausencia de datos se dice que es completamente aleatoria, En otras palabras, se establece que la probabilidad de que un dato esté ausente no está relacionada con los datos (observados y ausentes) (Enders, 2022, pg.6). Este mecanismo se define matemáticamente como:

$$P(\boldsymbol{R} = 0|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \psi) = P(\boldsymbol{R} = 0, \psi)$$

Es decir, la probabilidad de los datos faltantes no está relacionado con los datos **Y**. Si bien este escenario resulta conveniente, Little & Rubin (2020a) y Van Buuren (2012, pg.7) mencionan que este mecanismo no se da en la práctica debido a que resulta poco realista.

Ausencia de respuesta aleatoria (MAR): Cuando la probabilidad de no observar un valor de **Y** es igual dentro de algunos grupos definidos por los datos observados, se dice que la ausencia de datos es aleatoria (Van Buuren, 2012, pg.7); es decir, la probabilidad de que un dato sea faltante está relacionada con los datos observados, pero no con los datos perdidos. Considerando la definición formal de la distribución condicional de **R**, la distribución para un mecanismo MAR (He, Zhang & Hsu, 2021a) viene dada por:

$$P(\boldsymbol{R} = 0 \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \psi) = P(\boldsymbol{R} = 0|\boldsymbol{Y}_{mis}, \psi)$$

En general, los métodos modernos de imputación suponen que la falta de datos es generada por este mecanismo puesto que resulta en un supuesto más realista que el mecanismo anterior.

---

[1] El uso de la letra **R** que será utilizada a lo largo de este documento para describir la ausencia de respuesta. Esta notación también se puede encontrar en textos como Van Bureen (2012), Schafer (1997a); He, Zhang & Hsu (2021a), entre otros. Alternativamente, otros autores como R. J. A Little & Rubin (2020a) o Enders (2022) definen una matriz indicadora de falta de respuesta **M**, la cual caracteriza los datos no observados dentro de un conjunto de datos **Y** de dimensiones $n \times p$. Es decir, **M** será una matriz de $n \times p$ cuyos elementos m_{ij} se definen como $m_{ij} = 1$ cuando y_{ij} no fue observado y 0 en otro caso.

Ausencia de respuesta no aleatorias (MNAR): Si la probabilidad de no observar el valor de un dato está relacionada con la variable que se quiere observar, se dice que la ausencia de respuesta es no aleatoria; en otras palabras, el mecanismo MNAR establece que la probabilidad de que haya ausencia de respuesta está relacionada con los datos observados y, también, con los datos no observados (Enders, 2022). A diferencia de los mecanismos anteriores, la distribución condicional de **R**, dado los datos **Y**, no se simplifica.

Cabe destacar que, el conjunto de parámetros ψ del modelo formulado para **R** generalmente son desconocidos, por lo que, obviar o ignorar dicho conjunto de parámetros permitiría simplificar el análisis de los datos. En ese sentido, es fundamental considerar la ignorabilidad de los mecanismos de no respuesta al seleccionar y aplicar métodos de imputación. Como se ha descrito previamente, la naturaleza de los datos faltantes ya sea MAR o MCAR, determina la viabilidad y validez de las técnicas de imputación a emplear. En particular, la ignorabilidad, definida por la condición $P(\boldsymbol{R} = 0|Y_{obs}, Y_{mis}, \psi) = P(\boldsymbol{R} = 0, |Y_{obs}, \psi)$, implica que el mecanismo de no respuesta depende únicamente de los datos observados y no de los datos faltantes, lo que simplifica por tanto el proceso de imputación (Rubin, 1976).

Cuando esta condición se cumple, como en el caso de datos MAR, métodos como la imputación múltiple (Rubin, 1987) o los modelos de regresión (Van Buuren, 2018), pueden aplicarse sin necesidad de modelar explícitamente el mecanismo de no respuesta. En este escenario, los parámetros desconocidos ψ que describen la relación entre el mecanismo de no respuesta y los datos, pueden ignorarse, ya que no aportan información adicional relevante para la imputación. No obstante, en escenarios MNAR, donde la ignorabilidad no se sostiene, se requieren enfoques más complejos que incorporen supuestos adicionales sobre la estructura de los datos faltantes y, por tanto, es necesario estimar o modelar ψ. La comprensión y verificación de la ignorabilidad es un paso crítico para garantizar la robustez y la validez de las imputaciones realizadas tanto en encuestas de hogares como en censos de población, asegurando que los resultados derivados sean confiables y representativos de la población bajo estudio.

3. OBJETIVOS TEÓRICOS Y PRÁCTICOS DE LA IMPUTACIÓN

3.1 Justificación de la Imputación

En el marco del Generic Statistical Business Process Model[2] (GSBPM), el tratamiento de datos faltantes y la imputación se inscriben en los Procesos 2.5. “Design processing and análisis” y principalmente en el 5.4 “Edit and Impute”, cuyo propósito es asegurar la completitud y coherencia de los microdatos previos al análisis y la difusión. De acuerdo al estándar del GSBPM v5.2 (UNECE, 2025) los pasos específicos normalmente incluyen: determinar si se deben agregar o cambiar datos; seleccionar el método de edición e/o imputación a utilizar; Agregar/cambiar valores de datos; Escribir los nuevos valores de datos de vuelta en el conjunto de datos y marcarlos como cambiados; generar metadatos sobre el proceso de edición e imputación.

En este contexto, el análisis de datos faltantes es un paso crucial en el procesamiento de censos y encuestas, ya que permite identificar la magnitud y distribución de los valores ausentes,

[2] El GSBPM es un modelo estándar que describe y define el conjunto de procesos necesarios para producir estadísticas oficiales. Proporciona un marco común y terminología armonizada para ayudar a las ONE a modernizar sus procesos, mejorar la calidad y compartir métodos, lo que facilita la estandarización y comparación de estadísticas https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2016/mtg4/Paper_8_GSBPM_5.0_v1.1.pdf.

facilitando la selección de estrategias de imputación adecuadas. Para ello, se emplean herramientas de análisis estadístico y visualización que permiten calcular la tasa de respuesta por ítems, identificar patrones de omisión y evaluar combinaciones de preguntas con alta incidencia de datos faltantes. Entre las técnicas más utilizadas se encuentran el cálculo de tasas de respuesta por variable, las tablas de contingencia y matrices de ausencia de datos, los gráficos de datos faltantes como diagramas de calor o gráficos de mosaico, y el análisis de correlación entre variables con valores ausentes. Estas herramientas permiten evaluar si la ausencia de datos sigue un patrón sistemático o si puede considerarse aleatoria, lo que resulta determinante para la selección de la metodología de imputación más adecuada (Lin & Tsai, 2019; Sthamer, 2021).

En censos de población y vivienda, la imputación de registros completos (por omisión parcial o total de vivienda, hogar y/o personas) no debe asumirse como obvia ni universal. A diferencia de las encuestas (donde la no respuesta total suele abordarse mediante ajustes de factores de expansión), en los censos se convive con una tasa de omisión ("cobertura neta") considerada razonable y estimada externamente mediante encuestas poscensales u otros estudios de cobertura. En consecuencia, la base micro del censo se entiende, en principio, como la población efectivamente censada, mientras que las correcciones de cobertura se realizan en un plano analítico (p. ej., estimación por sistema dual u otros métodos demográficos), sin que ello implique necesariamente modificar los microdatos. En la práctica regional coexisten dos enfoques: (i) preservar los microdatos como enumeración efectiva y reflejar la omisión solo en estimaciones externas de cobertura; y (ii) aplicar imputación acotada de registros completos únicamente cuando existan indicios verificables de ocupación o no respuesta total, generando registros de contenido mínimo y coherente para resguardar la consistencia de tabulados en áreas pequeñas, manteniendo explícitos los códigos de "no especificado" en atributos no observados.

Desde una perspectiva metodológica, diversos estudios han abordado la clasificación y tratamiento de los datos faltantes en encuestas y censos. De Waal, Pannekoek y Scholtus (2011) detallan en el *Handbook of Statistical Data Editing and Imputation* las categorías de datos faltantes y su diferenciación en procesos de edición y validación de datos. Medina y Galván (2007), en *Imputación de datos: teoría y práctica* de la CEPAL, analizan cómo la presencia de datos faltantes afecta indicadores socioeconómicos y proponen estrategias para mitigar su impacto. Lin y Tsai (2019) presentan una revisión exhaustiva sobre imputación y describen métodos para distinguir datos faltantes de errores en la captura o preguntas no aplicables, mientras que Eltinge y Luery (2003) documentan los procedimientos empleados por agencias federales para clasificar y manejar datos faltantes en encuestas oficiales. Estas referencias proporcionan un marco robusto para comprender las decisiones de imputación bajo el GSBPM 5.4, tanto para no respuesta por ítem como para los escenarios censales en que se discute la imputación de registros completos frente a la medición externa de cobertura.

3.2 Impacto en la calidad de los datos

La imputación puede mejorar la completitud y la coherencia interna de los microdatos, pero su efecto sobre la calidad depende críticamente del mecanismo de no respuesta, la tasa de omisión y la adecuación del modelo utilizado. En contextos donde los datos verifican patrones de no respuesta MAR/MCAR, los métodos modernos de imputación (Como imputación múltiple) permiten inferencias válidas al propagar la incertidumbre de los valores faltantes. Bajo datos con patrones de no respuesta MNAR, la imputación puede introducir sesgos difíciles de cuantificar si no se modelan explícitamente las causas de la no respuesta. Es por ello que, previo a la imputación, es indispensable caracterizar el patrón de datos faltantes y documentar los supuestos que sostienen dicho procedimiento. Incorporar las variables relevantes (Como, por ejemplo, la perspectiva de género y

condiciones de cuidado) ayuda a evitar imputaciones que repliquen inequidades cuando la falta de respuesta no es aleatoria.

En encuestas complejas, los efectos en sesgo y varianza dependen también del diseño muestral (Estatificación, conglomeración, factores de expansión), métodos que ignoran el diseño pueden subestimar errores estándar o alterar las correlaciones entre las variables. Por ello, la imputación debe, en lo posible, respetar las condiciones de estratificación y conglomeración dadas en la fase de diseño, considerar pesos, o al menos, evaluar su impacto posterior en la estimación y en la calibración. Es importante recalcar que los procedimientos de imputación no sustituyen el control de calidad en la fase de levantamiento y que, incluso bien aplicados, añaden incertidumbre que puede verse reflejada en la inferencia.

En ese sentido, la imputación deja de ser aconsejable cuando la tasa de no respuesta compromete la representatividad de la muestra observada o cuando los supuestos MAR resultan poco plausibles. A modo de ejemplo, ajustes simples (Como reponderación) son más defendibles con tasas de no respuestas bajas (<5% por ejemplo), mientras que con niveles altos (> 20% a modo de ejemplo y, considerando que estos umbrales deben ser definidos de acuerdo a cada variable y realidad) se deteriora la efectividad de compensar la no respuesta y aumenta el riesgo de algún tipo de sesgo residual. En tales escenarios es recomendable combinar estrategias para afrontar los problemas de la variable debido a la ausencia de respuesta o bien, transparentar dichas limitaciones en las documentaciones metodológicas respectivas.

Como se describirá en la sección 7 del presente documento, la experiencia de oficinas de la región ofrece referencias operativas útiles. Para el Censo de México 2020 se controló que la imputación de variables distintas de “No especificado” no superara umbrales del 1% por variable, mientras que la imputación total de viviendas/personas se mantuvo en niveles inferiores al 5%, con verificación de que los indicadores con y sin imputación permanecieran estables. En el caso de Brasil, las tasas de imputación de ingresos en la PNAD Continua han estado por debajo del 10% y, desde 2015, bajo 3% con evidencia de que la contribución de la no respuesta a la varianza estimada fue baja.

Para evaluar y transparentar la calidad tras la imputación, es posible reportar indicadores tales como: Tasas de imputación por item y dominio, fracción de información faltante y el incremento relativo de la varianza (Que permite resumir cuanto de la incertidumbre proviene de los datos faltantes), comparaciones “con/sin imputación” de distribuciones, totales e indicadores sensibles (Pobreza, desigualdad, brechas de género) y análisis de sensibilidad a supuestos MNAR (por ejemplo, escenarios pesimistas/optimistas).

Desde el enfoque de género y otras dimensiones de desigualdad, es clave vigilar que la imputación no homogenice en exceso ni invisibilice patrones diferenciales. Métodos simples (medias, modas) o donantes mal definidos pueden reducir la variabilidad y desplazar medias en subgrupos, introduciendo sesgos sistemáticos si no se incluyen covariables pertinentes (Cuidado, informalidad, jefes de hogar mujeres). El uso de modelos o donantes debe, por tanto, incorporar variables que capturen estas heterogeneidades y validar resultados de manera desagregada.

En síntesis, la imputación mejora la usabilidad de los datos cuando: (i) los supuestos sobre el mecanismo de datos faltantes son defendibles, (ii) las tasas de no respuesta son compatibles con el método elegido, (iii) se respeta el diseño muestral, y (iv) se cuantifica y comunica la incertidumbre adicional introducida. Donde estos requisitos no se cumplen (Por altas tasas de no respuesta, MNAR no modelado o fuerte dependencia de pocos donantes) el riesgo de sesgo supera los beneficios de la imputación, por lo que se debe privilegiar mejoras en la recolección, uso de fuentes auxiliares y la transparencia sobre limitaciones y sensibilidad de los resultados.

4. PATRONES Y DISTRIBUCIÓN DE LOS DATOS FALTANTES

El análisis de los datos faltantes en censos y encuestas de hogares debe comenzar con una caracterización sistemática de su magnitud, localización y mecanismo subyacente. Este diagnóstico permite establecer si el faltante puede tratarse como completamente al azar (MCAR), al azar condicional a variables observadas (MAR) o no al azar (MNAR), y define las opciones metodológicas disponibles para su tratamiento. En la práctica regional, esta caracterización se apoya tanto en métricas descriptivas **(**tasas de no respuesta por variable, módulo, dominio geográfico y sociodemográfico) como en evidencia operativa proveniente de la metadata, que incluye intentos de contacto, duración de entrevistas, modo de levantamiento y calendario de trabajo de campo. Asumir explícitamente el mecanismo de faltantes, sustentado en evidencia empírica, es una condición previa para seleccionar métodos de imputación adecuados y para comunicar de forma transparente el alcance y las limitaciones de los resultados. Una implementación robusta integra el diagnóstico al flujo de producción, diferenciando el enfoque entre los Censos y encuestas de hogares. En Censos, el primer paso es distinguir los valores que verdaderamente "No aplican" versus la ausencia de respuesta a nivel de item, verificando concentraciones espaciales de los datos faltantes que puedan revelar potenciales problemas operativos (Cobertura, supervisión, etc.) Por otra parte, en encuestas, la distinción "No aplica" de la no respuesta se combina con el análisis del mecanismo de ausencia de respuesta y con la evaluación por estrado y/o UPM, incorporando además tasas de no respuestas al ítem para valorar sesgos de no observación. En ambos casos, se estima el impacto potencial del dato faltante sobre indicadores clave mediante comparaciones controladas con y sin imputación preliminar: Para Censos, con foco en distribuciones y totales demográficos por área/región y grupos de edad/sexo; para encuestas, incluyendo efectos sobre medias, proporciones muestrales y varianzas de diseño, así como posibles efectos en los factores de expansión al momento de la calibración. El resultado es posible sistematizarlo en un "cuaderno de imputación" con indicadores previos y posteriores por variable y subpoblaciones prioritarias (Sexo, edad, área, región, pertenencia indígena, condición migratoria, etc.) y con las reglas y supuestos que guiaron las decisiones correspondientes. Este registro fortalece la trazabilidad, la reproducibilidad y facilita la convergencia hacia lineamientos compartidos a nivel regional.

4.1 Análisis de patrones comunes

En el contexto de América Latina y el Caribe se observan patrones recurrentes que conviene documentar. La no respuesta de la unidad (a nivel de vivienda u hogar) suele concentrarse en zonas de difícil acceso o con restricciones operativas, mientras que la no respuesta al ítem aparece con mayor frecuencia en módulos o preguntas sensibles, extensos o con saltos complejos, como ingresos, ocupación y fecundidad. También es posible que el dato faltante pueda depender del modo de levantamiento y del calendario de campo, lo que introduce estacionalidad y heterogeneidad espacial. A diferencia del dato faltante "estructural", que deriva de filtros lógicos del cuestionario (razón por la cuál no debe imputarse), los patrones mencionados se asocian a mecanismos MAR o MNAR que exigen un tratamiento cuidadoso con apoyo de covariables y metadata.

La descripción empírica combina medidas sintéticas y visualizaciones. Tasas de no respuesta por variable y dominio, matrices de co-ocurrencia para detectar bloques de preguntas omitidas y gradientes de desigualdad por subgrupos proveen señales tempranas sobre el mecanismo de faltantes.

En operativos censales, los mapas temáticos por sector o manzana permiten aislar fenómenos operativos (P. ej., subcobertura en áreas periurbanas, rezago de supervisión o fallas de captura en módulos específicos) y detectar, por ejemplo, clústeres espaciales de no respuesta que requieran

acciones correctivas operativas o reglas de imputación focalizadas en dicho territorio. En encuestas de hogares, el análisis debe integrarse al diseño muestral: la estimación y comparación de tasas de no respuesta por estrato y UPM ayuda a identificar potenciales sesgos de selección y orientan decisiones de imputación dentro de dominios relativamente homogéneos, considerando también efectos sobre los factores de expansión al momento de la calibración y la varianza del diseño.

El uso de herramientas especializadas facilita este diagnóstico. En R, funciones como md.pattern del paquete mice, vis_miss de naniar o aggr de VIM permiten visualizar patrones y cuantificar co-ocurrencias, mientras que el paquete survey asegura que los resúmenes respeten estratos y conglomerados. En Stata, misstable summarize y misstable patterns ofrecen resúmenes rápidos integrables a flujos svy. En Python, missingno y pandas brindan paneles reproducibles para seguimiento operativo. Cuando el operativo lo amerita, herramientas SIG (QGIS/ArcGIS) apoyan la identificación de focos geográficos de no respuesta, insumo clave para diseñar clases de donación local en censos.

La evaluación del mecanismo debe incorporar pruebas formales cuando sea pertinente. El contraste de Little aporta evidencia sobre la plausibilidad de MCAR; modelos de propensión a faltar (logísticos o probit) con paradata y covariables del cuestionario ayudan a respaldar el supuesto MAR; y la persistencia de patrones de no respuesta luego de controlar por dichas covariables, especialmente en las colas de la distribución, sugiere la necesidad de tratar escenarios MNAR mediante análisis de sensibilidad. Estas pruebas no sustituyen el juicio metodológico, pero acotan la incertidumbre y orientan elecciones de imputación consistentes con los datos.

4.2 Métodos para detectar y manejar patrones

El manejo de los datos faltantes parte de un pipeline reproducible. Primero se ejecuta una auditoría o revisión que separa “No aplica” de “No responde”, calcula tasas y co-ocurrencias por variable y dominio, y describe su distribución espacial o temporal. Luego se aplican pruebas MCAR y modelos de propensión para cada variable crítica, incorporando paradata y variables del diseño muestral. Con este insumo se evalúa el impacto potencial del faltante sobre estimaciones clave, comparando escenarios con y sin imputación preliminar, y se clasifica cada variable en familias de tratamiento: (i) ajustes de ponderación; (ii) métodos de donante (Como hot-deck o k-NN); (iii) modelos estocásticos (GLM, multinivel, bayesianos); (iv) imputaicón múltiple; y (v) modelos de aprendizaje automático o de machine learning que pueden operar como motores predictivos dentro de (iii) y (iv) o como métodos auxiliares de (ii). En particular, algoritmos como random forest, gradient boosting o BART pueden ser utilizados para predecir items faltantes (Mediante imputación múltiple o simple), para estimar probabilidades de no respuesta (Propensity scores) y/o para definir vecindarios de donantes más “similares”.

Cuando el dato faltante es bajo y compatible con MCAR, los ajustes por no respuesta a nivel de ponderadores (a través de clases de respuesta homogénea) y la imputación por donante simple dentro de clases bien definidas suelen ser suficientes, siempre que se mantenga una bandera de imputación y que la estimación de varianzas respete el diseño (p. ej., replicación). En escenarios MAR con auxiliares informativos, la donación con clases finas por estrato y dominio, combinada con métodos como vecino más cercano con restricciones de distancia o predictive mean matching, preserva la forma de la distribución y limita sesgos. La imputación por regresión estocástica (lineal, logística u ordinal, según la naturaleza de la variable) es una alternativa eficaz cuando las relaciones entre covariables son estables y bien especificadas.

Para variables estratégicas o patrones complejos, la imputación múltiple constituye la opción preferente, dado que modela la incertidumbre y permite combinar estimadores y varianzas bajo reglas estándar. Su implementación debe ser compatible con el diseño: idealmente se imputan dentro de estratos o incorporando estratos y conglomerados como efectos o predictores, y el análisis posterior se realiza con los módulos de encuesta correspondientes antes de combinar resultados. Cuando existen registros administrativos de calidad, su integración como predictores o fuentes de donación mejora la coherencia; no obstante, reemplazar respuestas autodeclaradas por RRAA requiere análisis de concordancia y trazabilidad, y solo debería aplicarse bajo reglas explícitas y documentadas.

Si la evidencia sugiere MNAR, se recomienda complementar los métodos anteriores con análisis de sensibilidad. Los enfoques de mezcla por patrones y los ajustes delta permiten explorar cuánto deben desplazarse las imputaciones en segmentos específicos (P. ej., ingresos altos) para modificar conclusiones sustantivas. De manera similar, modelos de selección pueden ser útiles cuando la probabilidad de respuesta está asociada con el valor no observado. En todos los casos, conviene reportar los rangos de variación de las estimaciones y explicitar los supuestos que los generan, particularmente en dominios pequeños o poblaciones prioritarias.

La implementación técnica se apoya en software ampliamente disponible. En R, mice, simputation, Amelia, missForest/missRanger cubren un espectro desde donación hasta métodos basados en árboles o Imputación Múltiple; survey, mitools y miceadds por su parte facilitan la combinación con diseños complejos. En Python, scikit-learn, statsmodels y geopandas permiten construir pipelines que combinan modelos de propensión, donación por proximidad y validación geográfica. Este andamiaje debe culminar en un panel de control que reporte tasas de imputación, fracción de información faltante, medidas de variabilidad relativa y comparaciones con y sin imputación, junto con pruebas de coherencia interna. Documentar reglas, supuestos y banderas de imputación en microdatos y metadatos es una buena práctica esencial para la transparencia y la comparabilidad regional.

5. MÉTODOS TRADICIONALES DE IMPUTACIÓN

5.1 Listwise Deletion

El método de eliminación por lista, conocido también como análisis de casos completos, es una de las técnicas más sencillas para abordar la falta de datos en encuestas de hogares y censos de población. Este método consiste en eliminar todos los registros que presenten uno o más valores faltantes en las variables de interés, conservando únicamente aquellos casos que estén completos. Cuando el mecanismo de datos faltantes es del tipo MCAR, este enfoque produce estimaciones insesgadas para parámetros como medias, varianzas y coeficientes de regresión. Sin embargo, los errores estándar y los niveles de significancia calculados solo son válidos para el subconjunto de casos completos, los cuáles suelen ser más grandes en comparación con los obtenidos si se utilizara toda la información disponible.

No obstante, la aplicación de este método presenta limitaciones significativas sobre todo en encuestas de hogares y censos. Al eliminar casos incompletos, se asume que la submuestra excluida tiene características similares a la de los casos completos y que la falta de respuesta se generó de manera aleatoria, supuestos que rara vez se cumplen en la práctica (Medina y Galván, 2007, p.22). Además, una desventaja evidente es la posible pérdida de una proporción considerable de registros, especialmente cuando el número de variables con datos faltantes es elevado, lo que puede reducir drásticamente el tamaño de la muestra y afectar la representatividad de los resultados (Van Buuren, 2012, p.8). Little y Rubin (2020a, p.47) mencionan que las consecuencias de descartar casos

incompletos incluyen no solo una pérdida de precisión en las estimaciones, sino también en la introducción de sesgos cuando el mecanismo de no respuesta es MCAR. El grado de sesgo y la magnitud de la pérdida de precisión dependen además de las diferencias entre las unidades completas e incompletas en término de sus características observables y no observables.

En la región, las encuestas de hogares suelen exhibir tasas de no respuesta (total y por ítem) y patrones de ausencia más complejos, mientras que en censos de población y vivienda la no respuesta total es, por diseño y control operativo, generalmente baja; no obstante, pueden registrarse omisiones por ítem en preguntas complejas o módulos sensibles. En este contexto, la eliminación por lista (listwise deletion) puede resultar problemática (en especial cuando la ausencia no es MCAR) al introducir sesgos y pérdida de precisión, afectando la representatividad de dominios y subpoblaciones. Por ello, es crucial evaluar la viabilidad de su aplicación (diagnóstico MCAR/MAR/MNAR y magnitud por ítem/dominio) y considerar alternativas de imputación que preserven la integridad de la información y la comparabilidad de los resultados.

5.2 Pairwise Deletion

El método de eliminación por pares, también conocido como análisis de casos disponibles, busca mitigar la pérdida de información que caracteriza al método de eliminación por lista. En este enfoque, las estimaciones para cada variable se calculan utilizando únicamente los casos disponibles en dicha variable. Por ejemplo, las estimaciones para la variable V se basan en los casos completos de V, mientras que las estimaciones para la variable W se derivan de los casos completos de W, y así sucesivamente para el resto de las variables. Este método es relativamente simple, ya que aprovecha toda la información disponible y produce estimaciones consistentes para medias, correlaciones y covarianzas bajo el supuesto de que el mecanismo de datos faltantes es MCAR (Van Buuren, 2012, p. 10).

No obstante, en el contexto de las encuestas de hogares y censos de población en América Latina y el Caribe, este método presenta limitaciones significativas cuando se consideran las estimaciones en conjunto. En primer lugar, las estimaciones pueden estar sesgadas si el mecanismo de datos faltantes no es MCAR, lo cual es frecuente en este tipo de estudios debido a la complejidad de los patrones de no respuesta (Van Buuren, 2012, p. 10). Además, surgen problemas computacionales, como la posibilidad de que la matriz de correlación no sea definida positiva, un requisito esencial para la mayoría de los procedimientos multivariados. También pueden ocurrir correlaciones fuera del rango $[-1,+1]$, un problema que surge al utilizar diferentes subconjuntos de datos para calcular varianzas y covarianzas. Por último, no queda claro qué tamaño de muestra debe utilizarse para calcular los errores estándar, lo que dificulta la interpretación de los resultados (Van Buuren, 2012, p. 10).

Estas limitaciones hacen que el método de eliminación por pares sea menos adecuado para el análisis de encuestas de hogares y censos en la región, donde los datos faltantes suelen seguir patrones complejos y no aleatorios. Por lo tanto, es fundamental evaluar cuidadosamente su aplicabilidad y considerar métodos alternativos de imputación que permitan preservar la integridad de la información y garantizar la validez de las estimaciones.

5.3 Reponderación

Los procedimientos de imputación ponderada representan una alternativa para compensar la falta de respuesta en encuestas de hogares y censos de población, ajustando los factores de ponderación de las unidades que sí respondieron. Este enfoque busca que la muestra observada genere

estimaciones compatibles con los valores poblacionales de las variables de interés. En términos generales, cuando se detectan datos faltantes en una característica específica, los ponderadores de las unidades que proporcionaron información se recalibran para reflejar la estructura poblacional, lo que permite corregir parcialmente el sesgo introducido por la ausencia de respuesta.

No obstante, la aplicación de estos métodos enfrenta desafíos considerables. En primer lugar, es necesario generar tantos vectores de ponderación como variables con datos faltantes, lo que puede volverse computacionalmente complejo y poco práctico, especialmente en encuestas con un gran número de variables y patrones de no respuesta heterogéneos. Esta complejidad operativa ha llevado a que la literatura especializada recomiende evitar el uso de estos métodos en la mayoría de los casos, ya que su implementación puede resultar ineficiente y propensa a errores (Medina y Galván, 2007, p. 24). Además, la efectividad de los procedimientos de imputación ponderada depende en gran medida de la calidad y disponibilidad de información auxiliar para realizar los ajustes, un requisito que no siempre se cumple debido a limitaciones en los sistemas de información y en la cobertura de los registros administrativos.

Un aspecto crítico para considerar es la tasa de ausencia de respuesta; no existe un porcentaje universal que determine cuándo es recomendable utilizar estos métodos, pero en la práctica se sugiere que son más adecuados cuando las tasas de no respuesta son moderadas. Por ejemplo, cuando la tasa de no respuesta es baja (menos del 5 %), el impacto de los datos faltantes en las estimaciones puede ser mínimo, y métodos simples como la eliminación por lista o la imputación por la media (Que será detallada en la sección 6.1) podrían ser suficientes. En cambio, cuando la tasa de no respuesta es alta (superior al 20 %), los métodos de imputación ponderada pueden volverse menos efectivos, ya que la submuestra observada podría no ser representativa de la población, incluso después de ajustar los ponderadores. En estos casos, métodos más robustos, como la imputación múltiple o modelos basados en regresión, suelen ser más apropiados.

El mecanismo de no respuesta también juega un papel determinante. Si los datos faltantes siguen un patrón MCAR, los métodos de imputación ponderada pueden funcionar bien incluso con tasas de no respuesta moderadas. Sin embargo, si el mecanismo es MAR o MNAR, la efectividad de estos métodos disminuye, ya que los ajustes de ponderación no corrigen adecuadamente los sesgos introducidos por la no respuesta.

En encuestas de hogares y censos de población, donde las tasas de no respuesta suelen ser variables y, en muchos casos, elevadas, se recomienda realizar un análisis exhaustivo del patrón de datos faltantes antes de optar por métodos de imputación ponderada. Esto incluye evaluar la disponibilidad de información auxiliar de calidad y la capacidad técnica para implementar los ajustes necesarios. En casos de tasas de no respuesta altas o patrones complejos, es preferible considerar métodos de imputación más avanzados que permitan preservar la integridad y representatividad de los datos.

6. MÉTODOS MODERNOS DE IMPUTACIÓN

Los métodos modernos de imputación han surgido como respuesta a las limitaciones de los enfoques tradicionales, ofreciendo soluciones más robustas y flexibles para manejar datos faltantes. Estos métodos se caracterizan por su capacidad para preservar la estructura de los datos, mantener las relaciones entre variables y reducir los sesgos asociados con la no respuesta. A continuación, se describen algunos de los métodos modernos más utilizados en la práctica estadística.

6.1 Imputación Simple

La imputación simple es un conjunto de técnicas que buscan reemplazar los valores faltantes en un conjunto de datos por valores estimados, utilizando métodos sencillos y rápidos de aplicar. Estas técnicas son particularmente útiles cuando se requiere una solución inmediata para manejar datos incompletos, especialmente en contextos donde los recursos computacionales o el tiempo son limitados. Sin embargo, su simplicidad implica que no consideran la variabilidad individual de los datos ni las relaciones entre las variables, lo que puede introducir sesgos y distorsiones en las estimaciones. Esta metodología de imputación incluye métodos como la imputación por la media, la mediana, la moda, el uso de una constante, el método hot-deck y la regresión lineal. La elección del método depende del tipo de variable (continua, categórica o discreta) y del análisis que se desea realizar.

Imputación por medias no condicionadas

El método de imputación por la media no condicionada, también conocido simplemente como imputación por la media, es un enfoque de imputación que completa los datos faltantes en una variable continua con el promedio de los datos observados en dicha variable[3]. En el caso de variables categóricas, la imputación se realiza utilizando la moda de los datos observados, y el mismo criterio puede aplicarse a variables numéricas discretas (Van Buuren, 2012, p. 10). Este método es una solución rápida y sencilla para abordar la falta de datos, lo que lo hace atractivo en situaciones donde se requiere una respuesta inmediata. Sin embargo, este método carece de justificación teórica y distorsiona las estimaciones de parámetros, independientemente del mecanismo que genera la falta de datos (Enders, 2022, p. 25). La imputación por la media (No condicionada) subestima la varianza, altera las relaciones entre las variables y sesga casi cualquier estimación que no sea la media. Incluso la estimación de la media puede verse sesgada cuando el mecanismo de no respuesta no es completamente al azar (distinto de MCAR), es decir, cuando la ausencia depende de covariables observadas (MAR) o de valores no observados (MNAR). Solo bajo el supuesto fuerte de MCAR (y dentro de dominios correctamente definidos) la media imputada coincide, en expectativa, con la verdadera; incluso en ese caso, las varianzas quedan artificialmente reducidas y los errores estándar subestimados (Van Buuren, 2012, p. 11). Por estas razones, el uso de este método debe evitarse en la mayoría de los casos. Como señala Graham (2012, p. 51), “Recomiendo que las personas NUNCA utilicen este procedimiento”.

Imputación por medias condicionadas

El método de imputación por medias condicionadas es una variante refinada del método anterior. En este caso, la imputación se realiza utilizando medias condicionadas a los datos observados, lo que implica formar categorías basadas en covariables correlacionadas con la variable de interés. Los datos faltantes se imputan con la media de las observaciones provenientes de la submuestra que comparte características comunes dentro de cada categoría formada (Little & Rubin, 2020a, p. 70). Aun cuando este método atenúa algunos de los sesgos asociados con la imputación por la media no condicionada, sigue asumiendo que los datos faltantes siguen un mecanismo MCAR. Por otra parte, la formación de categorías puede introducir sesgos si las covariables utilizadas no capturan adecuadamente la estructura de los datos faltantes. Por lo tanto, si bien este enfoque es una mejora respecto al método

[3] En el caso de variables categóricas, la imputación de los datos faltantes es realizada utilizando la moda de los datos observados (Van Buuren, 2012, p.10). El mismo criterio puede utilizarse en el caso de variables numéricas discretas.

anterior, no se recomienda su uso cuando se dispone de alternativas más robustas para la imputación de datos faltantes (Medina & Galván, 2007, p. 27).

Imputación por Regresión

El método de imputación por regresión busca mejorar la precisión de las imputaciones al incorporar la información contenida en otras variables del conjunto de datos. Este enfoque comienza ajustando un modelo de regresión utilizando los datos observados, donde la variable con datos faltantes se modela en función de las variables observadas. Luego, los valores faltantes se reemplazan por los valores predichos (o estimados) bajo el modelo ajustado. De esta manera, los valores imputados corresponden a los valores más verosímiles según el modelo utilizado (Van Buuren, 2012, p. 12). Aunque este método aprovecha las relaciones entre variables para generar imputaciones más precisas[4], presenta limitaciones importantes. En primer lugar, los valores imputados tienden a tener una variabilidad menor que los valores observados, lo que puede subestimar la dispersión real de los datos. Además, la correlación entre los valores imputados y los valores de las variables predictoras es perfecta (igual a 1), lo que incrementa artificialmente las correlaciones en el conjunto de datos completos. Esto introduce sesgos en las estimaciones de varianzas y correlaciones, afectando la validez de los análisis posteriores (Enders, 2022, p. 27).

Bajo el supuesto de que el mecanismo de no respuesta es MCAR, la imputación por regresión produce estimaciones insesgadas para las medias y los coeficientes del modelo de regresión utilizado, siempre que las variables explicativas estén completas. Sin embargo, la variabilidad de los datos imputados se subestima sistemáticamente, y el grado de subestimación depende de la varianza explicada por el modelo y de la proporción de datos faltantes (Van Buuren, 2012, p. 12).

La idea subyacente a este método resulta intuitiva, dado que las variables en un conjunto de datos suelen estar correlacionadas, por lo que los valores faltantes pueden estimarse utilizando información de otras variables observadas. No obstante, como se ha mencionado, las imputaciones resultantes pueden introducir sesgos cuya naturaleza y magnitud dependen del mecanismo de no respuesta y del tipo de estimaciones realizadas (Enders, 2022, p. 27). Por lo tanto, aunque la imputación por regresión es una mejora respecto a métodos más simples como la imputación por la media, su uso debe ser cuidadosamente evaluado, especialmente en contextos donde la precisión y la validez de las estimaciones son críticas.

Imputación por Regresión Estocástica

El método de imputación por regresión estocástica es una mejora del método de imputación por regresión, diseñado para abordar algunas de sus limitaciones. Al igual que en la imputación por regresión, este método ajusta un modelo de regresión utilizando los datos observados y predice los valores faltantes basándose en las variables observadas. Sin embargo, a diferencia del método tradicional, la imputación por regresión estocástica agrega un término de ruido aleatorio a cada valor predicho. Este paso adicional introduce variabilidad en las imputaciones, lo que ayuda a restaurar la dispersión natural de los datos y a reducir los sesgos asociados con la imputación por regresión simple (Van Buuren, 2012, p. 13).

La incorporación de este término de ruido aleatorio tiene dos efectos principales. En primer lugar, reduce la correlación perfecta entre los valores imputados y las variables predictoras, lo que

[4] La imputación por la media puede considerarse como un caso especial del método de imputación por regresión donde las variables explicativas (predictores) son variables indicadoras para las celdas dentro de las cuales se imputa el promedio (R.J.A. Little and Rubin 2020a, p.68)

mitiga el sesgo en las estimaciones de correlaciones y varianzas. En segundo lugar, restaura la variabilidad de los datos, que suele subestimarse en la imputación por regresión simple. Como resultado, este método es capaz de producir estimaciones más insesgadas y realistas, especialmente cuando el mecanismo de no respuesta es MAR (Enders, 2022, p. 28).

A pesar de sus ventajas, el método de imputación por regresión estocástica no resuelve todos los problemas asociados con la imputación de datos faltantes[5]. Por ejemplo, la elección del término de ruido y su distribución pueden influir en los resultados, y el método sigue dependiendo de la calidad del modelo de regresión ajustado. No obstante, este enfoque es una de las pocas técnicas tradicionales que puede producir estimaciones insesgadas bajo el supuesto MAR, convirtiéndolo en una herramienta valiosa en el manejo de datos incompletos (Enders, 2022, p. 29). Además, la idea central detrás de la imputación por regresión constituye la base de métodos más avanzados, como los enfoques bayesianos y la imputación múltiple.

Imputación por Donante o *Hot-Deck*

El método de imputación por donante, también conocido como *Hot-Deck*, imputa los valores faltantes utilizando los valores observados en casos "similares" dentro del conjunto de datos, denominados "donantes"[6]. Por ejemplo, en una encuesta de hogares, un donante podría ser una persona del mismo género, grupo de edad, sector residencial y rama de industria que el receptor. La premisa subyacente es que, si dos individuos tienen características similares, es probable que los valores de la variable objetivo a imputar también sean similares. Este método es ampliamente utilizado en la práctica de encuestas y puede involucrar esquemas elaborados para seleccionar los casos donantes.

Una de las principales ventajas del método *Hot-Deck* es que, a diferencia de la imputación por la media, no distorsiona la distribución de los valores de la variable a imputar. Sin embargo, este método puede incrementar la varianza de las estimaciones, lo que puede ser significativo en algunos casos. Aunque es posible reducir este incremento en la varianza mediante estrategias como una selección más eficiente del esquema de muestreo, restricciones en el número de veces que un caso actúa como donante, la formación de estratos basados en variables observadas o el uso de un Hot-Deck secuencial, los métodos de imputación múltiple suelen preferirse sobre este enfoque, puesto que, no solo evitan el incremento en la varianza, sino que también proporcionan errores estándar válidos que consideran la incertidumbre asociada con el proceso de imputación. Cabe destacar que las estimaciones derivadas del método *Hot-Deck* son insesgadas solo bajo el supuesto de que el mecanismo de no respuesta es MCAR, un supuesto que rara vez se cumple en la práctica (Little & Rubin, 2020a, p. 78).

Cuando se aplica la imputación por donante (Hot-Deck), para cada elemento que no responde i, se busca un registro donante d en el conjunto de datos que tenga características lo más similares posible a las del elemento i, en la medida en que estas características estén correlacionadas con la variable objetivo y. Del donante seleccionado, el valor y_d se utiliza para imputar el valor faltante para el elemento i:

$$\tilde{y}_i = y_d$$

[5] Al añadir un ruido aleatorio a los valores ajustados bajo el modelo, es posible que, para valores localizados en los extremos de la distribución, el valor a imputar queda fuera del rango factible de la variable a imputar. Por ejemplo, pueden existir valores imputados cuyos valores son negativos aun cuando la variable a imputar solo toma valores mayores o iguales a cero.

[6] En este método, la imputación de los valores faltantes de un caso es realizada con los valores observados en algún otro caso similar al que se busca imputar. Sin embargo, cuando existen dos o más casos similares, pero con valores observados diferentes en las variables a imputar, la decisión sobre cuál caso tomar como donante, queda al arbitrio de quien realiza la imputación.

El elemento que no responde se denomina “receptor”. Este método puede aplicarse tanto para variables numéricas como categóricas. Si faltan varios valores en un registro, en principio, se utiliza el mismo donante para imputar todos estos valores faltantes, aunque pueden existir excepciones cuando los datos deben satisfacer restricciones específicas.

Existen diferentes enfoques para seleccionar donantes, los cuales pueden clasificarse en dos categorías principales:

Métodos que utilizan clases de imputación:

- Hot-Deck Aleatorio: Las clases de imputación se forman en función de variables auxiliares, y se selecciona aleatoriamente un donante dentro de la misma clase que el receptor.
- Hot-Deck Secuencial: No se construyen clases de imputación explícitas; en su lugar, se utiliza el primer registro posterior con valores similares en las variables auxiliares para imputar el valor faltante.

Métodos que buscan un donante minimizando una función de distancia:

- Hot-Deck del Vecino Más Cercano: Se selecciona el donante que minimiza una función de distancia entre las variables auxiliares del receptor y los donantes potenciales.

Además del Hot-Deck, existe el método de imputación Cold-Deck, donde los valores faltantes se imputan utilizando valores provenientes de otro conjunto de datos, como valores históricos de la misma unidad. Sin embargo, este enfoque rara vez es recomendable, ya que no considera cambios temporales o tendencias. Una mejora de este método es la imputación de razón, donde se ajusta un factor de tendencia para mejorar la precisión de las imputaciones.

La imputación por donante también se utiliza cuando faltan varios valores en un registro en variables fuertemente correlacionadas. Al elegir un solo donante para todos estos valores faltantes, se evita la inconsistencia entre las imputaciones. En tales casos, es necesario crear clases de imputación que sean homogéneas para varias variables objetivo simultáneamente. Este enfoque, conocido como imputación multivariante de donantes, es una solución específica para el problema de la imputación multivariante.

Imputación Aleatoria y secuencial

En la imputación Hot-Deck aleatoria, las clases de imputación se forman en función de variables auxiliares (características de fondo). Dentro de cada clase, se selecciona aleatoriamente un donante entre las unidades que comparten las mismas características que el receptor. Si no se encuentra un donante con todas las características requeridas, es necesario ampliar la clase de imputación descartando o combinando algunas de las variables auxiliares. Para evitar que una unidad sea donante para muchos receptores, lo que podría aumentar los errores estándar de las estimaciones y magnificar el impacto de valores atípicos, se pueden implementar restricciones. Por ejemplo, se puede limitar el número de veces que una unidad actúa como donante dentro de una misma clase de imputación.

En la imputación Hot-Deck secuencial, no se forman clases de imputación explícitas. En su lugar, para cada caso con datos faltantes, se utiliza el primer registro posterior en el conjunto de datos que comparte las mismas características de fondo. Sin embargo, si varios receptores de la misma clase de imputación aparecen en rápida sucesión, es posible que todos obtengan su valor imputado del

mismo donante, lo que puede introducir sesgos. Para mitigar este problema, se puede ajustar el método seleccionando aleatoriamente un donante entre los primeros k registros posteriores con las mismas características, excluyendo donantes ya utilizados.

La imputación secuencial puede aplicarse después de ordenar los registros de manera aleatoria, lo que convierte al método en estocástico. Sin embargo, si no se realiza este ordenamiento aleatorio, los valores imputados dependerán del orden de los registros en el conjunto de datos, lo que podría generar estimaciones sesgadas. Kalton (1983) propone métodos de imputación donde la probabilidad de selección de un donante es proporcional a su peso. Para evitar que una unidad con un peso pequeño sea seleccionada como donante para un receptor con un peso grande (o viceversa), se pueden utilizar variables de ponderación o las variables utilizadas para calcular los pesos como auxiliares en la selección del donante. Esto asegura que los pesos del donante y del receptor no difieran significativamente, mejorando la calidad de las imputaciones.

Imputación por vecinos más cercano

En esta metodología, se define una función de distancia $D(i,k)$ que permita medie la distancia entre dos unidades i,k, donde i corresponde al ítem sin respuesta y k un ítem que si fue respondido. Dicha función de distancia puede ser definida cumpliendo ciertas condiciones. Una función de distancia general de uso frecuente es la llamada distancia de Minkoski definida como:

$$D(i,k) = \left(\sum_j \left| x_{ij} - x_{kj} \right|^z \right)^{1/z}$$

Donde las variables x son numéricas y la suma se toma sobre todas las variables auxiliares a considerar, por otra parte x_{ij} y x_{kj} denotan el valor de la variable x_j en los registros i,k respectivamente. En ese sentido, se dice que $d = \arg\min_k D(i,k)$ es el vecino más cercano al item que no responde i convirtiéndose en su donante.

Cuando z tiende a infinito[7], se obtiene la llamada "distancia minimax"[8] definida como

$$D(i,k) = \max_j |x_{ij} - x_{kj}|$$

Donde el máximo se toma sobre todas las variables auxiliares x_j. Esta medida de distancia es utilizada a menudo en las oficinas de estadística, por ejemplo, en el software de edición e imputación GEIS de Statistics Canada (Cotton, 1991). En esta distancia, se escoge un registro donante tal que la diferencia absoluta máxima entre los valores de las variables auxiliares del donante y el receptor sea mínima. Esto asegura que incluso el valor de la variable auxiliar más diferente del registro del donante se acerque al valor correspondiente del receptor, lo que permite que el método sea robusto frente a la presencia de valores atípicos.

Es posible utilizar una función de distancia más general dada por la formula

$$D_\gamma(i,k) = \left(\sum_j \gamma_j \left| x_{ij} - x_{kj} \right|^z \right)^{1/z}$$

---

[7] Cuando z = 1 se obtiene la distancia "manzana", para z = 2 se obtiene la distancia euclidiana. Para z finitos grandes, las grandes diferencias entre x_{ij} y x_{kj} son penalizadas con más fuerza.

[8] Ver, por ejemplo, Sande 1982; R. J. Little and Rubin 2002b.

Donde γ_j es un factor o peso asociado a la importancia de la variable x_j. Dado que es relevante solo el peso relativo, se puede suponer que $\sum_j \gamma_j = 1$. El peso de la variable x_j debe estar relacionado con la importancia, para una imputación precisa, de encontrar un donante con un valor similar en esta variable. En la práctica, las ponderaciones adecuadas suelen ser más fáciles de determinar cuándo las variables x se normalizan por primera vez para que su varianza sea igual a 1, evitando una ponderación implícita cuando las variables se miden en unidades diferentes. Es posible tener en cuenta las covarianzas entre las variables al definir la distancia $D(i,k)$, pero esto generalmente complica la determinación de pesos adecuados.

La versión ponderada de la función de distancia $minmax$ viene dada por la ecuación

$$D_\gamma(i,k) = \max_j \gamma_j |x_{ij} - x_{kj}|$$

Otro caso especial del vecino más cercano es el método predictivo de emparejamiento de medias (Little, 1988). Cuando se utiliza este método de imputación, primero se lleva a cabo una regresión lineal de la variable objetivo y sobre varias variables predictoras numéricas x , utilizando los registros que tengan respuesta. Luego, el modelo de regresión resultante se utiliza para predecir (para cada registro) un valor para la variable objetivo por medio de la fórmula de regresión simple. El registro del donante para el elemento que no responde i viene dado por el elemento que responde d para el cual el valor predicho $\hat{y}_d$ es más cercano al valor predicho $\hat{y}_i$ para el elemento que no responde. Finalmente, se imputa el valor observado y_d del donante d.

La coincidencia predictiva de medias es otro caso especial donde se busca minimizar la función de distancia definida por

$$D(i,k) = |\overline{y}(\boldsymbol{x}_i) - \hat{y}(\boldsymbol{x}_k)|$$

Siendo $\boldsymbol{x}_i$ el vector con variables predictoras para la unidad i que no responde y $\boldsymbol{x}_k$ el vector que contiene las mismas variables para una unidad k. Esta función puede ser reescrita como

$$D(i,k) = \left|\boldsymbol{x}_i^T\beta - \boldsymbol{x}_k^T\beta\right| = \left|\sum_j (x_{ij} - x_{kj})\,\beta\right|$$

Donde β es el vector de coeficientes de regresión. La ecuación anterior, muestra que no es necesario calcular realmente los valores predichos para aplicar este método.

Imputación Cold-Deck

El método de imputación Cold-Deck consiste en reemplazar los valores faltantes de una variable con un valor constante obtenido de una fuente externa, como los datos de una encuesta anterior o un conjunto de datos histórico. Este enfoque es sencillo de implementar, ya que no requiere modelar relaciones entre variables ni realizar cálculos complejos. Sin embargo, su aplicación práctica suele tratar los datos imputados como si fueran una muestra completa, ignorando las implicaciones estadísticas de la imputación.

Una limitación importante del método Cold-Deck es la falta de una teoría sólida que respalde el análisis de los datos imputados. Como señalan Little y Rubin (2020a, p. 69) y Van Buuren (2012, p. 7), no existe un marco teórico satisfactorio para evaluar la validez de las estimaciones obtenidas mediante este método. Esto se debe a que el valor constante utilizado para la imputación no considera

la variabilidad natural de los datos ni las relaciones entre las variables, lo que puede introducir sesgos significativos en las estimaciones.

Además, el uso de valores históricos o externos puede no ser apropiado si las condiciones o tendencias han cambiado con el tiempo. Por ejemplo, si se utiliza un valor de una encuesta anterior sin ajustar por cambios en la población o en las variables de interés, las imputaciones pueden no reflejar la realidad actual. En la mayoría de los casos, este método no se recomienda, ya que existen alternativas más robustas, como la imputación por regresión o los métodos de imputación múltiple, que permiten incorporar la incertidumbre asociada con los datos faltantes y preservar la estructura de los datos.

A pesar de su facilidad de implementación, las técnicas de imputación simple tienen limitaciones importantes. No tienen en cuenta la incertidumbre asociada con los valores faltantes, lo que puede llevar a subestimar errores estándar y distorsionar inferencias estadísticas. Además, su efectividad depende en gran medida del mecanismo de no respuesta y del tipo de variable analizada por lo cual, su uso debe ser cuidadosamente evaluado, especialmente en contextos donde la precisión y la validez de las estimaciones son críticas.

6.2 Imputación múltiple

La imputación múltiple (MI, por sus siglas en inglés), introducida por (Rubin, 1988), es un enfoque para manejar datos faltantes en estudios estadísticos. El enfoque de Donald B. Rubin para la imputación múltiple, tal como se describe en (Rubin, 2004), es un método para tratar los datos faltantes en los análisis estadísticos donde se asume que los datos son MAR, lo que significa que la probabilidad de que un valor sea faltante puede depender de los datos observados, pero no de los datos faltantes.

Esta técnica permite generar valores razonables para datos que faltan, basándose en la distribución de los datos observados. El principio básico es que la imputación debería reflejar la incertidumbre acerca de los valores faltantes, generando varias versiones imputadas diferentes, lo que lleva a la "multiplicidad" en la imputación (Van Buuren, 2018). Un manejo inadecuado de los datos faltantes en un análisis estadístico puede conducir a estimaciones sesgadas y/o ineficientes de parámetros como las medias o los coeficientes de regresión, y errores estándar sesgados que resultan en intervalos de confianza y pruebas de significancia incorrectas. En todos los análisis estadísticos, se hacen algunas suposiciones sobre los datos faltantes.

Bajo el paradigma de imputación múltiple, la idea es generar múltiples conjuntos de datos donde cada valor faltante para un conjunto de datos Y_{mis} es reemplazado con un conjunto de valores plausibles, creando así múltiples versiones completas del conjunto de datos. Supongamos se generan M conjuntos de datos posibles, los resultados de estos M análisis se combinan en una única estimación y una única medida de incertidumbre. Este enfoque tiene la ventaja de reflejar adecuadamente la incertidumbre sobre los valores faltantes en las estimaciones finales, lo que puede dar lugar a inferencias más precisas y confiables en presencia de datos faltantes. En este método, la incertidumbre de la imputación se tiene en cuenta mediante la creación de estos múltiples conjuntos de datos. El proceso de imputación múltiple puede dividirse en tres fases:

- Imputación: Durante la fase de imputación, se generan M conjuntos de datos completos, donde M es el número de imputaciones. Cada conjunto de datos se crea reemplazando los valores faltantes con estimaciones basadas en un modelo de imputación. Este modelo se ajusta a los datos observados y también incorpora la variabilidad aleatoria,

lo que significa que las imputaciones son diferentes en cada uno de los M conjuntos de datos. Es decir, para un conjunto de datos con valores faltantes, se generan M imputaciones para cada valor faltante. Por lo tanto, a partir de un conjunto de datos original con datos faltantes, generamos M conjuntos de datos completos. Si denotamos la m-ésima imputación para el i-ésimo valor faltante como $y_{i,m}$, entonces, para cada i, generamos $y_{i,1}, y_{i,2}, \dots, y_{i,M}$.

- Análisis: En la fase de análisis, se lleva a cabo el análisis estadístico de interés en cada uno de los M conjuntos de datos completos como si fueran datos completos sin faltantes. Cada uno de estos M conjuntos de datos se analiza por separado utilizando el análisis estadístico completo de los datos. Si denotamos el estimador de interés como θ, entonces para cada conjunto de datos completado obtenemos un estimado $\hat{\theta}_m$ para $m = 1,2, \dots, M$. Esto resulta en M conjuntos de estimaciones y estadísticas de prueba.
- Combinación: En la fase de combinación, las M estimaciones y estadísticas de prueba de los conjuntos de datos imputados se combinan para producir una única estimación y estadística de prueba. La combinación tiene en cuenta tanto la variabilidad dentro de cada conjunto de datos imputados (debido a la variabilidad de muestreo) como la variabilidad entre los conjuntos de datos imputados (debido a la incertidumbre en el proceso de imputación). La estimación final de θ se calcula como el promedio de las M estimaciones, es decir, $\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}_j$.

Implementación general de los métodos de imputación múltiple

Supongamos θ es una cantidad de interés a calcular de una población estadística, ya sea una media, total poblacional, coeficiente de regresión, etc. Note que θ es una característica de la población estadística y no depende de características de un determinado diseño. Dado que esta cantidad θ solo es posible calcularla con la población completa, se suele calcular un estimador $\hat{\theta}$ del parámetro poblacional.

El objetivo es encontrar un estimador insesgado de θ tal que la esperanza de $\hat{\theta}$ sobre todas las muestras posibles de los datos completos Y sea igual al parámetro poblacional deseado, es decir, se busca que $E(\hat{\theta}|Y) = \theta$. Note que la incertidumbre acerca de la estimación $\hat{\theta}$ depende acerca del conocimiento que se tiene acerca del vector Y_{mis}. En ese sentido, si fuese posible generar valores para Y_{mis} de manera exacta, entonces la incertidumbre acerca de la estimación $\hat{Q}$ se reduciría o bien no existiría incertidumbre acerca de la estimación generada para el parámetro poblacional.

Sea $P(\theta|Y_{\text{obs}})$ la distribución a posteriori de θ, esta distribución puede ser descompuesta integrando sobre la distribución conjunta del vector $(Y_{\text{obs}}, Y_{\text{mis}})$, es decir:

$$P(\theta|Y_{\text{obs}}) = \int P(\theta|Y_{\text{obs}}, Y_{\text{mis}})P(Y_{\text{mis}}|Y_{\text{obs}})dY_{\text{mis}}$$

Dado que se desea hacer inferencia sobre el parámetro θ es de interés conocer la distribución de $P(\theta|Y_{\text{obs}})$ pues utiliza la información que se tiene, por otra parte $P(\theta|Y_{\text{obs}}, Y_{\text{mis}})$ es la distribución hipotética del parámetro sobre los datos completos y $P(Y_{\text{mis}}|Y_{\text{obs}})$ es la distribución de los valores perdidos dados los valores observados.

Notar que sería posible obtener M imputaciones $\dot{Y}_{\text{mis}}$ a partir de la distribución $P(Y_{\text{mis}}|Y_{\text{obs}})$, con ello, se podría calcular la cantidad θ a partir de la distribución de $P(\theta|Y_{\text{obs}}, \dot{Y}_{\text{mis}})$. (Van Buuren, 2018) muestran que la media posteriori de $P(\theta|Y_{\text{obs}})$ es igual a

$$E(\theta|Y_{\text{obs}}) = E(E[\theta|Y_{\text{obs}}, Y_{\text{mis}}]|Y_{\text{obs}})$$

En otras palabras, la media posteriori de θ bajo repetidas imputaciones de los datos.

Suponga que $\hat{\theta}_m$ es la estimación usando la m-ésima imputación, la estimación de las M estimaciones combinadas es igual a

$$\bar{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$$

En un caso multivariado, es posible que $\bar{\theta}_m$ contenga k parámetros y por tanto sea un vector de dimensión $k \times 1$. La varianza de la distribución a posteriori $P(\theta|Y_{\text{obs}})$ se puede escribir como la suma de dos componentes, esto es:

$$V(\theta|Y_{\text{obs}}) = E(V(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}) + V(E(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}})$$

La primera componente de la ecuación anterior puede interpretarse como la media de las repetidas imputaciones a posteriori de la varianza de θ (La cuál será denominada como intra-varianza) mientras que la segunda componente es la varianza entre las medias de θ estimadas con la distribución a posteriori (la cuál será llamada entre-varianza).

Si denotamos $\bar{U}_\infty$ y B_∞ como la intra y entre varianzas cuando $M \to \infty$ entonces se tiene que $T_\infty = \bar{U}_\infty + B_\infty$ corresponde a la varianza posteriori de θ. Cuando M es finito, podemos calcular la media de las varianzas de las imputaciones como

$$\bar{U} = \frac{1}{M}\sum_{m=1}^{M}\bar{U}_m$$

donde $\bar{U}_m$ corresponde a la matriz de varianzas covarianzas de $\hat{\theta}_m$ obtenida de la m-ésima imputación. La estimación insesgada de las varianzas entre las M estimaciones realizadas está dadada por

$$B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \bar{\theta}\right)\left(\hat{\theta}_m - \bar{\theta}\right)'$$

Para calcular la varianza total T cuando M es finito, es necesario incorporar el hecho de que $\bar{\theta}$ es estimado usando un número de imputaciones finita. (Rubin, 2004) muestra que dicho factor corresponde a $\frac{B}{M}$. Por tanto, la varianza total T de la estimación $\bar{\theta}$ a travéz de las M imputaciones puede ser escrita como

$$T = \bar{U} + B + \frac{B}{M}$$

$$= \bar{U} + \left(1 + \frac{1}{M}\right)B$$

(Steele, Wang, & Raftery, 2010) investigaron alternativas para obtener estimaciones de T utilizando mezclas de distribuciones normales. En este escenario, cuando existe normalidad multivariante y M no es grande, estos métodos producen estimaciones ligeramente más eficientes de T.

Consideraciones técnicas en la aplicación de los métodos de imputación

Considerando que es necesario realizar inferencia sobre la estimación puntual $\bar{\theta}$ y la varianza estimada definida como T, diversos autores [(Rubin, 1988); (Van Buuren, 2018); (Molenberghs et al., 2014)] proponen utilizar la distribución t.

(Van Buuren, 2018) menciona que la inferencia de un solo parámetro se aplica cuando $k = 1$, o bien si $k > 1$ pero además la prueba se repite para cada uno de los k componentes en el parámetro. Dado que la varianza total T es desconocida, $\bar{\theta}$ sigue una distribución t en lugar de la normal. Las pruebas univariadas para la imputación se basan en la aproximación:

$$\frac{\theta - \bar{\theta}}{\sqrt{T}} \sim t_{\nu}$$

donde t_{ν} es una distribución t-student con ν grados de libertad. Con lo anterior podemos por tanto construir un intervalo de $(1-\alpha)100\%$ para $\bar{\theta}$ definido en la siguiente ecuación

$$\bar{\theta} \pm t_{\nu,1-\alpha/2}\sqrt{T}$$

donde $t_{\nu,1-\alpha/2}$ corresponde al cuantil de probabilidad $1-\alpha/2$ de t_{ν}. Supongamos se desea testear la hipotesis nula $\theta = \theta_0$ para un valor en específico de θ_0. El valor-p del test se puede calcular como

$$P_s = \Pr\left[F_{1,\nu} > \frac{(\theta_0 - \bar{\theta})^2}{T}\right]$$

donde $F_{1,\nu}$ es una distribución (F) Fisher-Snedecor con 1 y ν grados de libertad.

Utilizando una aproximación estándar de tipo Satterthwaite, (Rubin, 1988) calculó los grados de libertad de la distribución de $\bar{\theta}$ dado los M conjunto de datos imputados como:

$$\nu = (M-1)\left[1 + \frac{\bar{U}}{\left(1+\frac{1}{M}\right)B}\right]^2$$

La ecuación anterior puede ser reescrita como

$$\nu = (M-1)\left[1 + \frac{1}{r_M}\right]^2$$

donde $r_M = \frac{(1+m^{-1})B}{\bar{U}}$ es conocida como el incremento de varianza relativa (RVI por sus siglas en inglés) debido a los valores faltantes, considerando que $\bar{U}$ representa la varianza de la estimación $\bar{\theta}$ cuando no existe variación entre los valores estimados $\hat{\theta}_m$, en cuyo caso $B = 0$.

Por otra parte, para θ podemos definir el ratio

$$\lambda_M = \frac{(1+m^{-1})B}{T}$$

el cuál puede ser interpretado como la proporción de varianza que se puede atribuir a la información perdida.

Si $\lambda_M = 0$, la información perdida no añade variación extra a la variación del muestreo, lo cuál ocurre excepcionalmente solo si se recrea de manera perfecta dicha información perdida. Por contraparte si $\lambda_M = 1$ toda la variabilidad es causada por la información faltante. Si $\lambda_M > 0.5$ la influencia del modelo de imputación en el resultado final es mayor que el modelo considerando los datos completos (Van Buuren, 2018). Notar que $r_M = \lambda_M/(1-\lambda_M)$.

Una cantidad estrechamente relacionada con λ_M se denomina "fracción de información faltante" (FMI, por sus siglas en inglés), puede ser calculada comparando la "información" en la densidad posteriori (t) aproximada, definida como el negativo de la segunda derivada de la densidad log-posterior, con la de la densidad posteriori hipotética de los datos completos, dando como resultado [(Rubin, 1988)]:

$$\gamma_M = \frac{r_M + \frac{2}{\nu+3}}{1+r_M}$$

Es fácil ver que $\gamma_M \to r_M/(1+r_M) = \lambda_M$ cuando $M \to \infty$. Esto permite observar que el efecto de los datos faltantes es una combinación de la actual cantidad de información perdida y el grado con el cuál aquella información de los datos incompletos contribuye a la estimación de interés mediante el modelo de imputación.

(Barnard & Rubin, 1999) muestran que la ecuación la ecuación descrita para γ_M puede producir valores en los grados de libertad que son mayores al tamaño muestral en los datos completos cuando la muestra es pequeña. Debido a esto, desarrollaron una adaptación para tamaños de muestras pequeñas teniendo en cuenta dicho problema. Se define ν_{old} como los grados de libertad de la ecuación descrita para γ_M y ν_{com} los grados de libertad de $\bar{\theta}$ cuando se tiene los datos completos sin valores perdidos. En este caso, si se tienen k parámetros para un tamaño muestral de n, entonces $\nu_{com} = n - k$. Los grados de libertad de los datos observados que tienen en cuenta la información faltante es

$$\nu_{obs} = \frac{\nu_{com}+1}{\nu_{com}+3}\nu_{com}(1-\lambda)$$

Los grados de libertad ajustados que se utilizarán para las pruebas en imputación múltiples puede escribir de manera concisa como

$$\nu = \frac{\nu_{old}\nu_{obs}}{\nu_{old}+\nu_{obs}}$$

(Van Buuren, 2018) señala que para la cantidad ν de la ecuación anterior siempre se tiene que $\nu \leq \nu_{com}$.

Número de imputaciones a realizar

La imputación múltiple es una técnica de simulación por lo que $\bar{\theta}$ y su varianza total estimada T están sujetas a errores de simulación. En ese sentido, la fórmula dada por

$$T_m = \left(1 + \frac{\gamma_0}{m}\right) T_\infty$$

es la relación entre la varianza del parámetro estimado en un escenario con un número finito de imputaciones (T_m) y la varianza del parámetro estimado en un escenario con un número infinito de imputaciones (T_∞).

Aquí, m representa el número de imputaciones múltiples y γ_0 es la fracción de información perdida. Esta cantidad es equivalente a la proporción esperada de observaciones que faltan en el caso de que Y sea una variable que no tenga covariab-les asociadas. Sin embargo, esta proporción suele ser menor si existen covariables que pueden predecir el valor de Y. Cuando m tiende a infinito, la varianza del estimador tiende a T_∞, es decir, se reduce la varianza debido al error de simulación. Sin embargo, en la práctica, rara vez se alcanza el límite de $m = \infty$ y se usa un número finito de imputaciones.

La cercanía de T_m a T_∞ es una medida de qué tan bien se ha estimado la varianza del parámetro. En teoría, cuanto mayor sea m, más cercano será T_m a T_∞, lo que significa que la varianza estimada es más precisa. Sin embargo, aumentar el número de imputaciones también aumenta la carga computacional, por lo que se debe encontrar un equilibrio. Según (Bodner, 2008), en la mayoría de los escenarios prácticos, se pueden obtener buenos resultados con solo 20-40 imputaciones múltiples.

El intervalo de confianza para la estimación depende tanto de ν como de m. (Royston, 2004) sugiere un criterio para determinar m basado en el coeficiente de confianza $t_\nu\sqrt{T}$, y propone que el coeficiente de variación de $\log(t_\nu\sqrt{T})$ debería ser inferior a 0.05. Este criterio tiene el efecto de reducir el intervalo de confianza en un 10%, lo que implica que se necesitarían al menos $m > 20$ imputaciones.

En su estudio, (Bodner, 2008) examinó la variabilidad de tres medidas específicas con diferentes números de imputaciones múltiples (m): el ancho del intervalo de confianza del 95%, el valor p y γ_0 (la proporción real de información perdida). En este contexto, Bodner estableció un criterio para seleccionar m, de tal forma que el ancho del intervalo de confianza del 95% no excediera en más del 10% su valor verdadero. Bajo este criterio, Bodner propuso recomendaciones específicas para m en relación con diferentes valores de γ_0. Para $\gamma_0 = (0.05,0.1,0.2,0.3,0.5,0.7,0.9)$, Bodner recomendó los siguientes valores de m: $m = (3,6,12,24,59,114,258)$ respectivamente.

Métodos de imputación

Imputación basada en Regresión lineal

La regresión lineal es frecuentemente el modelo preferido para la imputación de variables continuas que siguen una distribución normal.

$$Y_{\text{obs}}|X;\boldsymbol{\beta} \sim \text{N}(X\boldsymbol{\beta},\sigma^2)$$

Donde, $\hat{\beta}$ es el estimador del parámetro (o un vector de tamaño k) del modelo que se ajusta a las observaciones de datos Y_{obs}. Asimismo, $\mathbf{V}$ representa la matriz de varianzas-covarianzas de $\hat{\beta}$, y $\hat{\sigma}$ es la estimación de la varianza del modelo ajustado.

valores imputados que respeten la incertidumbre en las estimaciones de los parámetros de la regresión.

En la imputación múltiple, estos valores imputados se generan para cada conjunto de datos, y luego los resultados de cada conjunto de datos imputado se combinan para generar una estimación final que tiene en cuenta tanto la variabilidad dentro de los conjuntos de datos imputados como la variabilidad entre ellos.

Esto asegura que la estimación final refleje tanto la incertidumbre en la estimación de los parámetros de la regresión como la incertidumbre debido a los valores faltantes.

1) Se genera σ^* como

$$\sigma^* = \hat{\sigma}\sqrt{(n_{\text{obs}} - k)/\text{g}}$$

donde g es una realización aleatoria de una distribución $\chi^2_{n_{\text{obs}}-k}$.

2) se genera $\boldsymbol{\beta}^*$ como

$$\boldsymbol{\beta}^* = \widehat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}}\mathbf{u_1}V^{\frac{1}{2}}$$

donde $\mathbf{u_1}$ es una fila de k realizaciones independientes de una distribución normal estandar y $V^{\frac{1}{2}}$ es la descomposición de cholesky de $\mathbf{V}$

El valor imputado y_i^* para cada observación faltante y_i se obtiene como

$$y_i^* = \boldsymbol{\beta}\mathbf{x}_i + u_{2i}\sigma^*$$

donde u_{2i} es una realización aleatoria proveniente de una distribución normal estándar.

Imputación de variables binarias

En el caso de variables binarias, es posible utilizar un modelo logístico de la forma

$$\text{logit}P(Y_{\text{obs}} = 1|\mathbf{x}\boldsymbol{\beta}) = \boldsymbol{\beta}\mathbf{x}$$

Sea $\hat{\beta}$ la estimación del parametro del modelo ajustado a los individuos con la información observada Y_{obs} y su matriz de varianzas covarianzas $\mathbf{V}$. Sea $\boldsymbol{\beta}^*$ una realización de la distribución aposteriori de $\boldsymbol{\beta}$ aproximada por $\text{MVN}(\widehat{\boldsymbol{\beta}}, \mathbf{V})$

Para cada observación perdida y_{miss} tomamos $p^* = [1 + \exp(-\boldsymbol{\beta}^*\mathbf{x}_i)]^{-1}$ y generamos la imputación y_i^* como

$$y_i^* = \begin{cases} 1 & \text{si } u_i < p_i^* \\ 0 & \text{En otro caso.} \end{cases}$$

Donde u_i es una realización aleatoria de una distribución uniforme en $(0,1)$

Imputación de variables categóricas no ordenadas

En el caso de variables categóricas no ordenadas con $L > 2$ categorías pueden ser modeladas usando una regresión logística multinumal, donde cada categoría tiene una regresión logistica que se compara con otra categoría determinada (digamos $l = 1$)

$$P(y_{\text{obs}} = l|\mathbf{x}, \boldsymbol{\beta}) = \left[\sum_{l\prime=1}^{L} \exp(\beta_{l\prime}\mathbf{x})\right]^{-1} \exp(\boldsymbol{\beta}_l\mathbf{x})$$

donde β_l es un vector de dimension $k = \dim(\text{x})$ y $\boldsymbol{\beta_1} = 0$.

Sea $\boldsymbol{\beta}^*$ una realización proveniente de una distribución normal aproximada a la distribución aposteriori de $\boldsymbol{\beta} = (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_L)$ vector de $k(L-1)$. Para cada observación perdida y_{miss}, sea $p_{il}^* =$

$P(y_i = l|\mathbf{x_i}, \boldsymbol{\beta}^*)$ la probabilidad de estar en cada categoría y $c_{il} = \sum_{l\prime=1}^{l} p^*_{il\prime}$. Se define cada valor imputado y_i^* como

$$y_i^* = 1 + \sum_{l\prime=1}^{L} I\,(u_i > c_{il})$$

donde u_i es una realización aleatoria proveniente de una distribución uniforme en $(0,1)$ y $I(u_i > c_{il}) = 1$ si $u_i > c_{il}$. 0 en otro caso.

6.3 Consideraciones en la aplicación en encuestas de hogares

Según Medina and Galván (2007), es importante recordar que los métodos de imputación presuponen ciertas características en la distribución de los datos faltantes, pero no abordan explícitamente el mecanismo que condujo a la selección de las unidades de observación. De forma incorrecta, estos métodos suelen asumir que los datos provienen de una muestra aleatoria y que todas las unidades tienen igual probabilidad de ser seleccionadas. Las encuestas de hogares se realizan bajo diseños de muestreo complejos. Algunos autores, como Binder (1996) y Binder and Sun (1996), han planteado dudas sobre la validez de los métodos de imputación múltiple en tales contextos. Se reconoce que la ausencia de datos es un problema inherente en todas las encuestas, y es habitual buscar procedimientos para completar la información. A pesar de ello, los procedimientos existentes se enfocan principalmente en analizar el patrón de datos faltantes, sin considerar que las unidades de observación pueden tener diferentes probabilidades de selección.

Por otra parte, es importante considerar el hecho de que los métodos de imputación múltiple asumen que los datos observados y/o completos siguen cierta distribución, pero bajo el paradigma de las encuestas, Kish (1965) es conocido por enfatizar que en el muestreo de poblaciones no son las observaciones individuales las que siguen una distribución, sino más bien las probabilidades de selección asignadas a cada elemento en la muestra. Esto es un principio fundamental en el diseño de muestreo y análisis de encuestas, que ayuda a garantizar que la muestra sea representativa de la población en su conjunto. Incluso en casos donde la falta de respuesta es baja, Medina and Galván (2007) aconsejan analizar los ponderadores asociados a los datos faltantes. Puede suceder que un pequeño número de hogares en la muestra representen una porción importante de la población total, y un criterio de imputación inadecuado podría introducir sesgos difíciles de identificar y evaluar.

Los autores enfatizan que, en los diseños de muestreo complejos, la selección de observaciones depende del método de estratificación y conglomeración del marco de muestreo, así como del vector de ponderaciones asociado a las diferentes unidades en la muestra. Además, Medina and Galván (2007) señalan que la estratificación, la conglomeración y las ponderaciones deben tenerse en cuenta a la hora de imputar datos. En el contexto de una encuesta de hogares, la imputación no solo debe considerar el patrón de datos faltantes, sino también las probabilidades de selección de las unidades de observación. Este enfoque aborda algunas de las limitaciones de los métodos de imputación tradicionales. Por ejemplo, al considerar las probabilidades de selección, se puede mitigar el riesgo de introducir sesgos en las estimaciones debido a la sobre o subrepresentación de ciertos grupos en la muestra.

Al tener en cuenta la estratificación y la conglomeración, se pueden preservar las correlaciones entre las unidades dentro de cada estrato o conglomerado, que a menudo se pierden en los métodos de imputación que tratan cada unidad de observación de forma independiente. Sin embargo, también se debe tener en cuenta que, sin importar el método de imputación utilizado,

siempre habrá cierta incertidumbre asociada con la imputación de datos faltantes. Por lo tanto, es importante manejar con cuidado los datos imputados y tener en cuenta esta incertidumbre al hacer inferencias a partir de los datos.

A pesar de que Binder and Sun (1996) demuestran que bajo el supuesto de un diseño de muestreo aleatorio simple y sin remplazo se pueden generar estimaciones precisas para medias y totales, siempre que se utilicen métodos bayesianos (bootstrap), Binder (1996) conjetura que la imputación múltiple no es adecuada en diseños complejos en los que existen más de una etapa de selección, conglomeración y probabilidades de selección desiguales puesto que las expresiones que se deben aplicar para la estimación de la varianza se complejizan. A pesar de ello, para los autores que proponen la imputación múltiple no resulta una preocupación cómo dichas imputaciones afectan las estimaciones finales ante un muestreo multietápico.

Medina y Galván (2007) también enfatizan que los procedimientos de imputación no deben ser considerados como una solución definitiva para la falta de datos, sino como una herramienta que permite el manejo de los datos faltantes de una manera más rigurosa y estructurada. Los investigadores deben ser conscientes de las suposiciones subyacentes en cada método de imputación y su posible impacto en las conclusiones derivadas de los datos imputados. El autor señala además que aún "persiste el desafío de desarrollar algoritmos de imputaciones robustos que tengan en cuenta el diseño de la muestra y las probabilidades de selección de las observaciones".

La aplicación de métodos de imputación en diseños de muestreo complejos requiere un cuidado adicional. Es importante recordar que estos métodos deben adaptarse al diseño de muestreo particular y a la estructura de los datos faltantes. No todos los métodos de imputación son adecuados para todos los tipos de datos ni para todos los diseños muestrales: la elección debe alinearse con la escala de la variable (continua, categórica ordenada/nominal), el mecanismo de ausencia (MCAR/MAR/MNAR), y los rasgos del diseño (estratos, conglomerados/UPM, pesos), diferenciando además entre operativos censales y encuestas de hogares. Finalmente, los investigadores deben ser conscientes de que incluso los métodos de imputación más sofisticados no pueden reemplazar completamente los datos faltantes. A pesar de las técnicas de imputación, siempre existe el riesgo de sesgo debido a la falta de datos. Por lo tanto, es fundamental minimizar la cantidad de datos faltantes en la etapa de diseño y recolección de datos, y tratar los datos faltantes de manera adecuada en la etapa de análisis.

7. PROCESOS DE IMPUTACIÓN EN LA RONDA DE CENSOS 2020

7.1 Censo México 2020

El Censo de Población y Vivienda 2020 se planeó para ser realizado del 2 al 27 de marzo de 2020 bajo la metdologia de derecho o de jure por lo que las unidades de observación fueron las y los residentes habituales del territorio nacional, así como las viviendas particulares y colectivas. Se captó información con dos cuestionarios: uno básico aplicado a toda la población, y uno ampliado, que incluía todas las preguntas del básico más otras preguntas adicionales, aplicado a una muestra. Se utilizaron diferentes métodos de recolección de la información: entrevista directa, que se realizó principalmente por medio de dispositivos moviles de captura, en cuestionarios impresos en casos excepcionales y posteriormene se habilitó autoenumeración por internet y entrevista asistida por teléfono.

El inicio del operativo de campo practicamente coincidio con la aparición del primer caso confirmado de COVID-19 en el país, el 27 de febrero, lo que impactó en la realización de la

capacitación, tareas operativas y de gabinete. Esto provocó un incremento en el número de viviendas habitadas en las que se negaron a proporcionar información por temor al contagio y la habilitación de medios alternativos para que la población contestara el censo.

La imputación consideró los casos de no respuesta total en las viviendas, omisión de menores y la ausencia de respuesta en cada una de las variables de los instrumentos de captura. En el caso de las encuestas regulares y especiales, los ítems con incongruencias se trataron mediante reglas de edición e imputación. Para los casos de no respuesta, se ajustaron los factores de expansión conforme a la metodología de la ENOE: primero por no respuesta (a nivel de UPM/estrato y por dominios de estudio) y luego por proyección de población al punto medio del levantamiento, asegurando la coherencia con los totales de control (INEGI, 2023)[9].

La unidades y variables consideradas para la imputación fueron:

- Características de las viviendas habitadas donde no se obtuvo respuesta y de su población residentes.
- Las viviendas en áreas que no fue posible visitar y sus residentes.
- Personas menores de 0 a 6 años en viviendas entrevistadas.
- Todas las variables con omisión de respuesta o cuya información fuera incongruente con los criterios de validación establecidos para el cuestionario.

Metodología de imputación

En el Censo 2020 se utilizaron métodos de imputación determinísticos, basados en modelos considerando donadores y métodos mixtos para la no respuesta total. Se utilizó programación en SAS, SPSS y SQL para la implementación de la imputación con estos métodos y realizar el análisis de los resultados.

En el caso de la omisión de respuesta en variables específicas, se aplicó la metodología de vectores teóricos implementada en el Instituto y aplicada desde el Censo 2000, que permite realizar un análisis exhaustivo de la información y dar seguimiento a los cambios realizados. Esta metodología también es utilizada en algunos de los proyectos del programa de encuestas del INEGI, mientras que en otros se utilizan asignaciones directas acordes al diseño conceptual de la encuesta.

Viviendas habitadas sin información

En un censo algunas viviendas habitadas quedan sin información debido a causas como la ausencia de informantes, negativas a participar, o dificultades de acceso. Para el Censo 2020, se implementó una imputación más compleja que en eventos anteriores[10], asignando características de las viviendas y sus habitantes por medio de donantes de la información cuando fue posible.

---

[9] INEGI Encuesta Nacional de Ocupación y Empleo. Cómo se hace la ENOE: métodos y procedimientos. / Instituto Nacional de Estadística y Geografía. -- 3ra. ed. -- México: INEGI, c2023. https://www.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/889463909743.pdf

[10] Para 2010, por ejemplo, se imputaron 3 personas por vivienda de manera homogénea y a todas las variables de vivienda y población se les asignó el valor “No especificado”.

Viviendas pendientes en áreas no levantadas

La cantidad de viviendas habitadas a imputar se estimó utilizando datos de planeación, operativos previos, conteos en campo, otras fuentes de información como el Marco de Muestreo Maestro que utiliza el INEGI para encuestas en hogares, o imágenes satelitales. Toda esta información proporcionó datos más actualizados al período censal, lo que mejoraba la estimación. Las características de estas viviendas se dejaron como datos no especificados para evitar sesgos.

Para estas viviendas se imputó el número de ocupantes utilizando un modelo Poisson con valor esperado igual al promedio de ocupantes en la vivienda, que tuvo en cuenta la variabilidad en el número de habitantes por vivienda en el municipio. A estas personas se les asignó el sexo mujer u hombre de manera alternada dentro de cada municipio y al resto de las características se le asigaron códigos de “No especificado”.

Viviendas pendientes de entrevista en áreas levantadas

En estas áreas, la imputación se realizó mediante el método "*hot-deck*", donde se asignaron a las viviendas sin información las características de viviendas vecinas dentro de la misma AGEB[11]. Este método aseguró que la distribución de las variables imputadas se asemejara a la distribución de las variables observadas, replicando la estructura por edad y sexo de las viviendas captadas. Cuando no fue posible encontrar donantes para estas viviendas, se utilizó la imputación como si se tratara de viviendas pendientes en áreas no levantadas.

Personas de 0 a 6 años omitidas

Habitualmente los censos de población, incluido México, presentan omisión de personas menores, por lo anterior, se realizó un análisis de la población de cero a seis años contabilizada por el Censo 2020 en conjunto con la información de registros administrativos, a partir del cual se implementó una imputación de personas menores de siete años en viviendas en las que reside alguna mujer en edad reproductiva, con hijos nacidos sobrevivientes y los mismos no fueron declarados como residentes habituales de la vivienda.

Variables con omisión

Las variables con omisión son tratadas durante el proceso de edición de datos, el cual se realizó mediante la aplicación de los criterios y tratamientos diseñados con apego al marco conceptual del Censo 2020 (INEGI 2021) y de acuerdo con la estructura y el contenido de los cuestionarios básico y ampliado. Para este proceso se implementó la metodología de vectores teóricos para la validación automática de las características de la población y de las viviendas.

Cada criterio ofreció una solución derivada de la lógica de las preguntas y del flujo del cuestionario respetando al máximo las respuestas de las y los informantes. En los casos en los que no se contó con suficientes datos para dar consistencia a la información, de acuerdo con los códigos válidos de respuesta predefinidos, se asignaron códigos especiales de la categoría “No especificado”.

[11] Las AGEB son unidades geográficas relativamente pequeñas, con una superficie que oscila entre 1 y 10 hectáreas, además, están delimitadas por características físicas o culturales distintivas.

Resultado y evaluación

Se imputaron 0.7% de menores en 2.7% de las viviendas particulares. Con el método del vecino más cercano se imputaron 3.5% de viviendas y el 3.3% de la población en las viviendas pendientes de entrevista; 0.1% de viviendas en esta situación se imputaron con la metodología utilizada para las áreas no levantadas por no encontrarse donantes para ellas. Finalmente, 0.1% de viviendas fueron imputadas en áreas no levantadas con el mismo porcentaje de población. Se generaron indicadores de vivienda y población para sus principales características y esta última por grupos de edad, comparando los indicadores con y sin imputación a nivel municipal, con lo que pudo constatarse que se mantuvieron los indicadores prácticamente iguales y la estructura de la población se mantuvo consistente.

En la imputación de variables, se revisó que el porcentaje de datos imputados a valores distintos a "No especificado" no superara el umbral del 1% sobre el universo de la variable para asegurar que la distribución de los datos no sufriera modificaciones que pudieran sesgar los resultados. Se buscó que la mayor parte de la no respuesta y su imputación con códigos "No especificado" se mantuvieran cerca del límite establecido para cada variable, fijado de acuerdo con su comportamiento histórico.

Es importante que la imputación realizada mantenga la estructura de los datos y permita hacer inferencias sobre la población. Aunque lo ideal es evitar la no respuesta total, esta sigue siendo un reto debido a factores como la negativa de los residentes a participar, la inseguridad o ausencia de informantes en las viviendas habitadas. Por ello es importante seguir explorando estrategias de captura, validación en los dispositivos móviles o de auto enumeración, seguimiento constante y así como fomentar la auto enumeración en busca de disminuir la no respuesta.

Asimismo, es importante considerar en la planificación censal mecanismos para mitigar el rechazo o la no respuesta total y parcial a la vez que también se debe incorporar desde el inicio cuáles serán los métodos por utilizar y la información complementaria o de apoyo que se va a utilizar en los procesos de imputación. Contar con métodos de imputación de información para futuros censos con la finalidad de resolver los casos de viviendas que no responden es crucial. El método del vecino más cercano tiene la ventaja de imputar todas las variables provenientes de viviendas presumiblemente similares, lo que permite mantener congruencia en los resultados y su interpretación. La imputación de menores es una práctica adoptada en varios países, es importante contar con mecanismos para realizarla.

Dentro de las acciones consideradas para futuros eventos, se encuentra la campaña de comunicación, es importante sensibilizar a la población sobre la realización de los Censos de población y la importancia de la información que a través de ellos se obtiene. Por ello, dentro del programa censal se encuentra inmerso el diseño e implementación de una campaña de comunicación antes, durante y después del operativo censal.

El uso de la imputación del vecino más cercano y la inclusión de menores son procedimientos que deberán implementarse desde la planeación de futuros eventos, para contar con una estrategia eficiente para la atención de la no respuesta total.

Considerando que se espera que el siguiente evento censal no se vea afectada por una pandemia o alguna situación similar, será importante hacer una revisión de cómo se obtendrán los vecinos para la imputación considerando las posibles características de las viviendas en las que no se obtiene respuesta.

7.2 Censo Chile 2024

El Censo de Población y Vivienda 2024 de Chile se levantó durante los meses de marzo y julio bajo metodología de derecho, empadronando a las personas de acuerdo con su lugar de residencia habitual. La operación estuvo basada en un diseño mixto secuencial, en donde el método principal de recolección fue la entrevista presencial con dispositivos móviles de captura (CAPI[12]), con métodos adicionales de recolección, como el Censo en línea (CAWI[13]) y cuestionarios en papel para contingencias. Todas las modalidades de recolección estaban integradas a un sistema de seguimiento en tiempo real. La transición desde instrumentos en papel, utilizados en 2017, permitió controles de consistencia en el punto de captura y una gestión operativa oportuna durante los cinco meses de recolección.

El procesamiento estadístico se estructuró en un ciclo iterativo de validación, edición e imputación, guiado por el principio de modificación mínima para preservar la distribución observada. Las reglas aplicadas en la edición fueron lógicas y deductivas, diseñadas a partir del cuestionario, evidencia de estadísticas vitales y otras fuentes de información, y prácticas consolidadas en operaciones oficiales como la Encuesta de Caracterización socioeconómica (Casen) y la Encuesta Nacional de Empleo (ENE). El objetivo de los procesos de validación y edición fue identificar y corregir valores implausibles y contradicciones internas de los datos, priorizando siempre la coherencia intrahogar.

Por otro lado, el proceso de imputación tuvo por objetivo identificar y corregir falta de respuesta al ítem en variables críticas como sexo, edad y parentesco. Se combinaron enfoques de donantes (*hot-deck*) con métodos de proximidad (K-Nearest Neighbors), definiendo la similitud a partir de información de la persona y del hogar (como parentesco, nivel educativo, tamaño y composición del hogar o características del/de la jefe(a)) y limitando la reutilización de donantes para evitar homogeneización excesiva.

La variable edad se consolidó primero en el proceso de validación-edición, confrontando la edad declarada con la fecha de nacimiento bajo reglas de plausibilidad y consistencia. Para los casos con falta de respuesta, se siguió una secuencia en dos momentos. Primero, se recuperó información cuando la edad había quedado consignada inadvertidamente en campos de texto (P. ej., en nombres u observaciones) mediante expresiones regulares, cuidando distinguir marcas de enumeración del hogar. Luego, para registros sin señales aprovechables, se imputó con KNN cuando existían covariables informativas; en hogares con escasa información auxiliar se aplicó *hot-deck* aleatorio, restringiendo su uso para resguardar la variabilidad.

En la variable sexo la no respuesta fue excepcional debido a los controles de la aplicación electrónica. Los pocos casos faltantes se abordaron, en primer término, recuperando marcas explícitas registradas por el censista en el texto libre (indicaciones como “hombre” o “mujer”) y, en su ausencia, utilizando un diccionario de nombres propios donde cada nombre se encontraba asociado a un sexo.

Para el parentesco, la imputación se basó en *hot-deck* con estratos de similitud definidos por sexo y edad de la persona, así como por la edad y sexo del/de la jefe(a) el tamaño del hogar. La asignación se verificó contra las reglas de consistencia intrahogar, a fin de asegurar la plausibilidad de los vínculos y roles. En los casos excepcionales en los que no se encontró ningún parentesco imputable consistente con otros integrantes del hogar, se asignó el parentesco “otro parentesco”.

---

[12] Por sus siglas en inglés Computer-Assisted Personal Interviewing.
[13] Por sus siglas en inglés Computer-Assisted Web Interviewing.

La evaluación de los resultados de estos procesos se organizó comparando las bases previas y posteriores al procesamiento en los niveles nacional, regional y comunal. Se analizaron las tasas de no respuesta, edición e imputación por variable, y se verificó que la edición no introdujera distorsiones en la estructura demográfica de sexo y edad. Se aplicaron indicadores estándar para detectar preferencias de dígitos en la declaración de edad (índices de Whipple y Myers), se revisó la distribución por edades simples y quinquenales, y se examinó la razón de masculinidad total y por quinquenios. Estos contrastes internos y externos, junto con series históricas, reforzaron la interpretación de las variaciones observadas.

A su vez, se verificó la replicabilidad de los procedimientos implementados, en particular de aquellos que incorporan componentes aleatorios, como la selección de donantes en los métodos *hot-deck* o la inicialización de algoritmos de imputación basados en proximidad. Para ello, se fijaron semillas de aleatoriedad controladas y se documentaron los parámetros utilizados en su ejecución.

El proceso dejó constancia de la trazabilidad de cada cambio y documentó reglas, métricas e insumos auxiliares. En síntesis, el CPV-2024 aplicó un flujo de validación, edición e imputación consistente con estándares internacionales: completitud con resguardo de la estructura observada, coherencia dentro del hogar y documentación integral que permite auditar decisiones y evaluar su impacto en la calidad de los resultados.

7.3 Censo Colombia 2018

La calidad de los datos del Censo Nacional de Población y Vivienda (CNPV) 2018 en Colombia estuvo condicionada por diversas situaciones de campo, como el desconocimiento de la información solicitada, el rechazo de las personas a proporcionar datos, la no participación de hogares en la investigación, entre otros factores. Estas situaciones generaron datos faltantes o inconsistentes, lo que, de no ser tratado adecuadamente, podría introducir sesgos en los resultados, especialmente en niveles de desagregación geográfica, donde el censo es la principal fuente de estadísticas confiables. Para abordar estos desafíos, se aplicaron técnicas de validación, consistencia e imputación, utilizando como insumo tanto Registros Administrativos como la información recolectada en la misma operación censal.

Enfoque Tradicional y Avances Metodológicos

Históricamente, el tratamiento de datos inconsistentes o faltantes en Colombia se ha basado en métodos como el *hot-deck*, que sustituye valores faltantes utilizando datos de donantes con características similares. Sin embargo, siguiendo los avances metodológicos propuestos por Rubin (1976, 1987), se ha adoptado un marco conceptual más robusto, que incluye técnicas como el algoritmo *Expectation Maximization* (EM) y la Imputación Múltiple (IM).

Para el CNPV 2018, se decidió imputar únicamente los datos faltantes parciales o inconsistentes, es decir, aquellos casos en los que se contaba con parte de la información recolectada. Para los registros con información completamente faltante, se realizó la estimación de la omisión. El método *hot-deck* fue seleccionado por su capacidad para aprovechar la cercanía geográfica y las características comunes entre donantes y receptores.

Proceso de Imputación

El proceso de imputación en el CNPV 2018 se basó en normas condicionadas a los datos observados, plasmadas en un documento de normas de validación e imputación. Este documento fue elaborado

con aportes de mesas internas y recomendaciones de expertos evaluadores del censo. Las normas de imputación consisten en un conjunto de reglas condicionales que permiten identificar y transformar datos faltantes o inconsistentes en valores válidos, preservando la consistencia en las respuestas secuenciales de los individuos. El proceso de imputación se dividió en dos grandes fases:

- La creación de tablas paralelas de calidad: Estas tablas contenían valores de 0, 1 o -1, indicando si un dato era inconsistente, consistente o faltante, respectivamente.
- Aplicación de normas de imputación: Se aplicaron las normas establecidas para corregir datos inconsistentes o faltantes en las tablas de viviendas, hogares, personas y personas fallecidas.

Imputación de Variables Clave

Identificación de personas: Se utilizaron registros administrativos, como el Archivo Nacional de Identificación (ANI) y el Registro de Menores de la Registraduría Nacional del Estado Civil (RNEC), para imputar datos faltantes en variables de identificación, como la fecha de nacimiento. Este proceso se realizó de manera determinística, utilizando cruces por tipo y número de documento, así como por nombres y fechas de nacimiento, con un proceso de transformación fonético para evitar sesgos en el emparejamiento. En cuanto a la ubicación y viviendas, se garantizó la imputación del uso y ocupación de las viviendas, así como la geocodificación espacial de cada encuesta. Con respecto a hogares y personas, la imputación se realizó de manera lineal, siguiendo el orden de las preguntas en el cuestionario, para asegurar la consistencia en las respuestas.

Resultados y Calidad de los Datos

Como resultado del proceso de imputación, se obtuvieron tablas de ubicación, viviendas, hogares, personas y personas fallecidas depuradas y consistentes. Este proceso se llevó a cabo en la segunda fase del procesamiento general de la base de datos, asegurando los universos de los flujos y filtros del cuestionario (DANE, 2021).

La experiencia del CNPV 2018 demuestra que la combinación de métodos tradicionales, como el *hot-deck*, con el uso de registros administrativos y normas de validación robustas permite mejorar significativamente la calidad de los datos censales. Además, la aplicación de técnicas de imputación basadas en criterios geográficos y demográficos asegura la consistencia y representatividad de los resultados, minimizando los sesgos asociados con la falta de respuesta o la inconsistencia en los datos.

8. PROCESOS DE IMPUTACIÓN EN ENCUESTAS DE HOGARES

Guatemala

El Instituto Nacional de Estadística de Guatemala implementó un riguroso proceso de imputación en la Encuesta Nacional de Condiciones de Vida (ENCOVI) 2023, con el objetivo de garantizar la calidad y representatividad de los datos utilizados para la medición de la pobreza. La imputación se centró en la sección de alimentos consumidos en el hogar, dentro del capítulo "Gastos y Autoconsumo", donde se identificaron retos específicos, como valores extremos y el doble conteo en productos procesados.

Para abordar estos desafíos, se establecieron criterios para la identificación de valores denominados “super extremos”, definiéndolos como aquellos que excedían en más de 12 desviaciones estándar el promedio del producto. En total, se detectaron 331 casos de este tipo, lo que representó el 0.06% del total de valores reportados. Los valores extremos fueron sustituidos mediante imputación basada en la mediana de hogares con características similares, considerando el tamaño del hogar, la ubicación geográfica y el tipo de producto.

El doble conteo de alimentos procesados fue tratado mediante la exclusión de aquellos reportados como "producción propia", evitando así la inflación del agregado de consumo. Esta decisión se basó en un análisis detallado del origen de los productos, diferenciando entre los adquiridos por compra, donación o producción interna del hogar. Además, se realizaron ajustes en los valores de alimentos comprados en supermercados para asegurar la consistencia utilizando la frecuencia reportada de compra. Esto permitió corregir posibles inconsistencias entre los montos totales de gasto y los valores individuales de cada producto registrado.

El proceso de imputación y limpieza de datos fue sometido a una revisión técnica rigurosa por parte del equipo del INE, que validó la coherencia de las imputaciones y minimizó la subjetividad en la toma de decisiones. Como resultado, la aplicación de estos procedimientos evitó la sobreestimación del consumo total de los hogares y garantizó la representatividad de los datos empleados en la evaluación de la pobreza extrema. La limpieza y correcciones aplicadas evitaron la sobreestimación del agregado de consumo y garantizaron la representatividad de los datos en el análisis de la línea de pobreza extrema. Además, la exclusión de valores de alimentos procesados y la imputación de valores extremos mejoraron la coherencia entre categorías de datos, reduciendo posibles sesgos. Durante este proceso, se identificaron diversos retos, destacando la complejidad en la identificación y exclusión de valores super extremos, aunque este esfuerzo permitió una depuración más efectiva de la información. Asimismo, el tratamiento del doble conteo en alimentos procesados requirió la formulación de supuestos, aunque su impacto fue mínimo.

Entre las buenas prácticas implementadas, se resalta el uso de criterios estadísticos definidos, como la evaluación de desviaciones estándar y el análisis de origen de los datos, lo que facilitó el tratamiento de valores atípicos. Además, el trabajo colaborativo del equipo técnico aseguró la coherencia en la toma de decisiones. Como recomendaciones para futuras implementaciones, se sugiere optimizar el diseño de las preguntas en campo para captar datos más precisos y minimizar errores, así como fortalecer la automatización de procesos de validación en tiempo real para reducir la ocurrencia de valores extremos e inconsistentes.

En términos de planes futuros, se contempla el fortalecimiento de los sistemas electrónicos de validación en campo, el perfeccionamiento de los criterios para la identificación de valores problemáticos mediante un análisis más detallado de subgrupos específicos y la aplicación de validaciones y patrones de salto directamente en la recolección de datos. Finalmente, se prevé la evaluación de modelos de imputación para su aplicación en futuras ediciones de ENCOVI y la Encuesta Nacional de Ingresos y Gastos de los Hogares, con el fin de mejorar la calidad de los datos recolectados en relación con el consumo de alimentos.

Ecuador

En el caso de Ecuador, la ENEMDU es una encuesta de hogares de tipo probabilístico bietápico de elementos, con estratificación geográfica por dominios de estudio y con representatividad nacional, urbano y rural recogida mensualmente que se dirige a individuos de 5 años y más de edad, y se realiza en aproximadamente 9.000 viviendas, lo que equivale a alrededor de 30.000 individuos en todo el territorio nacional.

La encuesta contempla un esquema de levantamiento de información a través de un formulario dividido en varias secciones, y que a su vez sigue un esquema de flujos controlado[14], lo que permite el direccionamiento de los informantes de acuerdo con sus características y situación laboral. En ese sentido, al levantarse cómputos de las respuestas obtenidas en las distintas preguntas que conforman la encuesta, se observan valores perdidos o *missing* en preguntas abiertas, o dentro de la categoría "no sabe". Esto se asocia con la negativa o desconocimiento de los informantes para proveer información. Con esto en consideración, se seleccionaron las preguntas tenencia de Registro Único de Contribuyentes (RUC)[15] del lugar de trabajo (p49) e ingreso laboral (ila).

En el caso de la variable ila, se construyó una nueva variable, la cual se denominó new_ila. La variable mantuvo los criterios metodológicos oficiales[16], pero en el caso de los ingresos negativos no los asignó como valores faltantes, sino que preservó su valor. Además, se categorizó a la variable en tres grupos: 1.Ingreso mayor o igual al salario básico unificado - SBU[17], 2. Ingreso menor o igual al SBU, y 3. Ingresos negativos.

Se probó y evaluó el desempeño de distintas técnicas de machine learning, como aprendizaje por ensambles (Random Forest - RF y XGBoost), redes neuronales (Multilayer Perceptron - MLP) y métodos supervisados simples (Support Vector Machine – SVM) sobre las variables p49 y new_ila en la ENEMDU 2021. Se utilizó la métrica de precisión (accuracy) de cada modelo para seleccionar la técnica más efectiva en la imputación definitiva de la información perdida. Los resultados mostraron que el modelo XGBoost tuvo un rendimiento superior, con una precisión del 88,7% para la variable p49 y del 83,8% para la variable new_ila, en comparación con las técnicas RF (precisión del 88,1% para p49 y 83,3% para new_ila), MLP (precisión del 88,4% para p49 y 83,3% para new_ila) y SVM (precisión del 88,4% para p49 y 83,2% para new_ila). Por lo tanto, se seleccionó XGBoost para realizar la imputación definitiva.

Una vez determinada la técnica con mejor rendimiento predictivo (XGBoost), se imputó la información perdida en las preguntas p49 y new_ila. Luego, con la información completa, se generaron las variables de sectorización del empleo (que incluye la variable p49 en su construcción) y clasificación de la condición de actividad de los empleados (que incluye la variable new_ila en su construcción).

Como resultado de la imputación, se observaron cambios en la distribución de las variables. En el caso de la variable p49, se encontró que de las 424.904 personas empleadas que respondieron "no sabe" en la pregunta, 307.525 personas fueron categorizadas como "sí" y 117.378 personas como "no". Por otro lado, al imputar la variable new_ila, se logró predecir los ingresos para un total de 66.443 empleados que inicialmente no proporcionaron información sobre sus ingresos laborales. Estos empleados se distribuyeron de la siguiente manera: 52.828 personas fueron clasificadas como 1. Ingreso mayor o igual al SBU, 13.605 personas como 2.Ingreso menor al SBU y 10 personas como 3. Ingreso negativo.

---

[14] El cumplimiento de los flujos del formulario es controlado a partir de mallas de validación, implementados en los sistemas de captación (el INEC de Ecuador denomina al sistema de captación de información como Sistema Integrado de Producción Estadística - SIPE).

[15] El Registro Único de Contribuyentes (RUC) es el documento que identifica e individualiza a los contribuyentes, personas físicas o jurídicas, para fines tributarios (SRI, 2022).

[16] La metodología oficial del INEC para la construcción del ila se encuentra en el siguiente link: https://www.ecuadorencifras.gob.ec/documentos/web-inec/Bibliotecas/Revista_Estadistica/Aspectos_metodologicos_sobre_la_medicion_de_la_pobreza_por_ingresos_en_el_Ecuador.pdf

[17] En el Ecuador el salario básico unificado (SBU) para el año 2021 fue de $400 dólares en el territorio continental y de $720 dólares en la región insular de Galápagos.

Una vez observados los cambios en la distribución de las variables imputadas, se procedió a comparar los resultados de los indicadores generados a partir de estas variables, específicamente el sector del empleo y la condición de actividad, antes y después de la imputación. El análisis de los indicadores estimados basados en la variable de sectorización del empleo, tanto antes como después de la imputación, reveló un aumento significativo en varias de sus categorías; sin embargo, no se encontraron cambios en la dispersión en comparación con las cifras oficiales. En cuanto a los indicadores derivados de la variable condición de actividad de los empleados, los resultados obtenidos no revelaron cambios en la dispersión.

Entre las principales recomendaciones derivadas de este estudio se encuentran: la necesidad de implementar técnicas de búsqueda de hiperparámetros y análisis exploratorios de variables más exhaustivos. Lo cual permitirá mejorar la exactitud y el rendimiento de los modelos de imputación utilizados. Además, es importante explorar otras técnicas de machine learning que sean más eficientes computacionalmente en comparación con las utilizadas en el estudio. Algunas opciones que considerar son los algoritmos genéticos, la eliminación recursiva de variables y la optimización bayesiana. Estas técnicas pueden ofrecer resultados aún más precisos y eficientes en términos de tiempo de procesamiento.

Asimismo, se sugiere que se realice un contraste entre la imputación mediante técnicas de machine learning y las técnicas de imputación tradicionales, como hot deck, cold deck o vecino más cercano. Este contraste permitirá evaluar y comparar el desempeño de diferentes enfoques de imputación y determinar cuál es el más adecuado para cada situación.

Costa Rica

El diseño muestral de la Encuesta Nacional de Hogares (ENAHO) es probabilístico, estratificado y bietápico. Para los valores de ingreso que no se reportan, se lleva a cabo dos tipos de imputación: una para los ingresos por trabajo de personas mayores de 15 años y otra para ingresos por rentas y transferencias, aplicable a personas de 12 años y más. Desde el año 2010, se utiliza la metodología de medias condicionales, decisión que se debe a que la tasa de no respuesta en las variables de ingreso no es significativa. Este enfoque es adecuado cuando la no respuesta es baja, ya que permite imputar valores basados en patrones observados en los datos disponibles, manteniendo la coherencia interna de la información.

El proceso de imputación se lleva a cabo en cuatro fases:

- Fase 1: Se utilizan cuatro variables: zona, sexo, escolaridad y ocupación. Se realiza un cruce de datos; si se obtiene un grupo que coincida, se calcula un promedio que se asigna. Si no hay coincidencias, se avanza a la siguiente fase.
- Fase 2: Se utilizan tres variables: zona, sexo y ocupación. Se repite el proceso de cruce y cálculo de promedios. Si no hay coincidencias, se pasa a la siguiente fase.
- Fase 3: Se utilizan dos variables: zona y sexo. Se realiza el cruce y se asigna un promedio si hay coincidencias. De lo contrario, se avanza a la siguiente fase.
- Fase 4: Se utilizan dos variables: sexo y escolaridad. Se sigue el mismo procedimiento de cruce y asignación de promedios.

Esta metodología ha mostrado porcentajes de imputación que varían entre un 2% y un 11%, lo que indica que la mayoría de los datos de ingresos son reportados directamente por los encuestados, y solo una pequeña proporción requiere imputación.

Brasil

La Encuesta Nacional por Muestra de Domicilios Continua (PNAD Continua) fue implementada por el Instituto Brasileño de Geografía y Estadística (IBGE) en todo el territorio nacional en enero de 2012. Se realiza mediante muestreo probabilístico y tiene como principal objetivo recopilar información sobre la fuerza de trabajo y sus variaciones a corto, mediano y largo plazo.

Adicionalmente, la encuesta recopila información sobre otros temas, como ingresos laborales y de otras fuentes, educación, vivienda, otras formas de trabajo (trabajo para el propio consumo o uso, trabajo voluntario, tareas domésticas y cuidado de personas), trabajo infantil, acceso a televisión e Internet, y posesión de teléfono móvil para uso personal, entre otros temas relevantes para el estudio del desarrollo socioeconómico del país.

Los indicadores relacionados con la fuerza de trabajo se publican mensualmente para Brasil a través de trimestres móviles; trimestralmente se desglosan para Brasil, Grandes Regiones, Unidades de la Federación, Regiones Metropolitanas que incluyen el municipio de la capital y las capitales. Anualmente, se publican indicadores complementarios con el mismo desglose geográfico que los publicados trimestralmente. Los demás temas de la encuesta, en general, se publican anualmente. Sin embargo, el intervalo de tiempo entre dos ediciones consecutivas para el mismo tema puede reducirse o ampliarse según la variación de sus indicadores a lo largo del tiempo.

Los datos de la PNAD Continua pasan por etapas de crítica, es decir, un proceso para detectar errores en los datos estadísticos (Chambers, 2006), y de imputación, un procedimiento en el que se inserta un valor para un ítem específico en el que la respuesta está ausente o es inutilizable (ONU, 2000). Estos procesos permiten aumentar la calidad del producto y generar una base de datos completa. Desde enero de 2012 hasta marzo de 2020, las entrevistas de la PNAD Continua se realizaron exclusivamente mediante el modo de recolección CAPI (computer-assisted personal interviewing), es decir, con contacto cara a cara entre el entrevistador y el informante, utilizando un cuestionario electrónico implementado en el Dispositivo Móvil de Recolección (DMC), un teléfono inteligente con algunas funciones bloqueadas.

El objetivo de este texto es describir cómo se realiza la imputación de los ingresos laborales. Se eligieron estas variables debido a que requieren un procesamiento más complejo, dada su naturaleza cuantitativa y su asociación con variables como el nivel educativo, el tipo de ocupación, la ubicación geográfica, entre otras. Además, los ingresos son una de las informaciones más sensibles y difíciles de recolectar, debido a la inseguridad de los encuestados al proporcionar dichos datos. Este grupo de variables presenta el mayor porcentaje de rechazo en las respuestas, lo que implica mayores porcentajes de imputación.

La imputación en la PNAD Continua se realiza de forma probabilística o determinística. En este contexto, se utilizan informaciones auxiliares para establecer, con base en reglas predefinidas, un único valor posible que se asignará al registro con falta de respuesta o error de medición. Esta etapa se ejecuta utilizando el paquete estadístico SAS Enterprise Guide.

En ausencia de información que permita la imputación determinística, la PNAD Continua adopta la imputación probabilística. En este caso, la imputación es parte de un proceso aleatorio y, por lo tanto, si se repite, puede generar valores diferentes. El método elegido para la encuesta es el vecino más cercano (NIM, del inglés Nearest-neighbour Imputation Methodology), implementado en el Canadian Census Edit and Imputation System (CANCEIS), un paquete de crítica e imputación desarrollado por Statistics Canada. La elección del CANCEIS se debe a que este método ha sido utilizado en otros operativos del IBGE, como el Censo Agropecuario de 2006, la Encuesta Nacional

por Muestra de Domicilios (PNAD) de 2007 a 2015, el Censo Demográfico 2010 y la Encuesta de Presupuestos Familiares (POF) 2008-2009 y 2017-2018. Posteriormente, se utilizó en la Encuesta Nacional de Salud (PNS) 2013 y 2019, y también está siendo utilizado en la Encuesta Nacional de Salud y Demografía (PNDS).

La crítica de los ingresos laborales comienza durante la recolección de datos, con reglas de advertencia y error en el Dispositivo Móvil de Recolección (DMC) y en el Sistema e Indicadores Gerenciales de Recolección (SIGC), que genera informes de registros por encima del percentil 95%, basados en datos de períodos anteriores para cada Unidad de la Federación (UF). Estos registros son supervisados por los respectivos Coordinadores Estatales o entrevistadores y pueden ser confirmados con el informante.

Después del cierre de la base mensual, los registros pasan por una primera etapa de imputación determinística en SAS, donde se realizan consistencias con códigos de ocupación y posición en la ocupación. Luego, se lleva a cabo una inspección visual de ingresos extremos, basada en la distribución empírica de la variable, utilizando un criterio que separa grupos de registros homogéneos con características similares. La descripción de este método está disponible en BAPTISTA ET AL, 2022. Después de la confirmación o ajuste de los valores, se realiza una segunda etapa de imputación determinística, con la posibilidad de recuperar registros de entrevistas anteriores. En caso de que la imputación determinística no sea posible, se procede a la imputación probabilística utilizando el método NIM (Nearest-neighbour Imputation Methodology).

La similitud entre receptores y donantes se mide mediante las siguientes variables auxiliares: condición en el hogar, sexo, edad, posición en la ocupación, años de estudio, horas trabajadas y contribución a la seguridad social en cualquier trabajo durante la semana de referencia. Otras informaciones importantes para comprender y reproducir el proceso de imputación probabilística de la PNAD Continua se enumeran a continuación:

- La imputación se realiza de forma conjunta para toda la población de la encuesta y no por subgrupos poblacionales (clases de imputación). Por lo tanto, un registro puede ser imputado por un donante de otra Unidad de la Federación (UF).
- La similitud entre registros no significa igualdad en todas las variables auxiliares. Por ejemplo, el sexo del donante puede diferir del sexo del receptor.
- Un donante puede ser utilizado para más de un receptor.
- Los pesos muestrales no se consideran en la función de distancia.

En caso de falla en la imputación, una nueva etapa en SAS realiza la imputación utilizando la mediana de la variable definida por subgrupos, como UF y posición en la ocupación.

La última etapa del tratamiento de los ingresos laborales consiste en la identificación e imputación de ingresos extremos que fueron confirmados por la red de recolección y considerados válidos en la inspección visual de valores extremos. Sin embargo, debido a que poseen valores muy elevados, estos pueden generar impactos artificiales en los indicadores de ingresos medios o totales, así como en la desigualdad (outliers no representativos).

En esta etapa, se consideran outliers los ingresos superiores al valor de la media más seis desviaciones estándar dentro de los estratos definidos por las Grandes Regiones. Estos valores se imputan de manera determinística por el valor más alto entre aquellos que no superan el límite, es

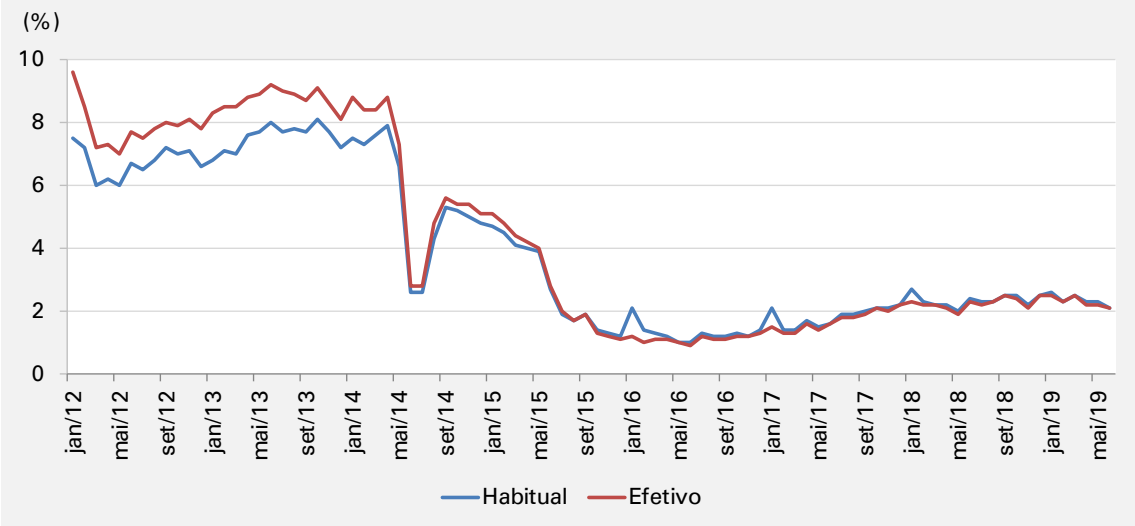
decir, que no son considerados outliers. Para obtener más detalles sobre este método, se puede consultar el Anexo 8 de las notas técnicas de la encuesta (IBGE, 2020).

Para ejemplificar el resultado final del proceso, la Figura 1 presenta las tasas de imputación de los ingresos habituales y efectivos del trabajo principal desde el inicio de la encuesta en 2012 hasta junio de 2019. Se puede observar que, incluso durante el período de implementación de la encuesta, la tasa de imputación no superó el 10%. A partir de 2015, esta tasa disminuyó considerablemente y se mantuvo por debajo del 3% hasta junio de 2019.

Hypólito (2020) realizó un estudio sobre el porcentaje de la varianza total de la estimación de ingresos que se debe a la falta de respuesta y al proceso de imputación, y concluyó que, para la gran mayoría de los indicadores analizados, estos valores son bajos y no afectan la calidad de las estimaciones.

Para llegar a la configuración actual de la imputación de ingresos en la PNAD Continua, se realizaron varios cambios a lo largo del proceso desde la implementación de la encuesta en 2012. El principal punto de cambio fue la publicación del ingreso medio en una Unidad de la Federación (UF) que se destacó significativamente debido a un valor extremadamente alto, lo que alteró la distribución de los ingresos en esa localidad. La etapa final de tratamiento determinístico se creó después de que ocurriera esta situación.

Figura 1
Tasa de imputación de las variables de ingresos del trabajo principal, por tipo de ingreso, Brasil, enero de 2012 a junio de 2019



Fonte: IBGE. PNAD Contínua. Janeiro de 2012 a junho de 2019

9. CONCLUSIONES

9.1 Resumen de Hallazgos y reflexionaes finales

El documento constata que la imputación es un componente imprescindible del procesamiento estadístico en censos y encuestas de hogares de la región, siempre subordinado a una primera fase robusta de prevención, control y validación de la no respuesta. La evidencia comparada muestra que los métodos tradicionales (p. ej., eliminación por lista/pares y medias condicionadas) solo resultan aceptables en escenarios muy acotados (baja no respuesta, mecanismos cercanos a MCAR, variables poco influyentes), pues tienden a perder información y a introducir sesgos o subestimar la varianza. En contraste, los enfoques que modelan explícitamente la incertidumbre y preservan la estructura de los datos son los que, en promedio, mejor sostienen la calidad de las estimaciones: imputación por donantes bien diseñada (*hot-deck* secuencial o por vecino más cercano con clases homogéneas y límites de uso del donante), imputación por regresión estocástica y, sobre todo, imputación múltiple (MI) con combinación adecuada de estimadores y varianzas.

Un hallazgo transversal es que el "mejor método" depende del contexto: tipo de variable, patrón/mecanismo de no respuesta (MAR/MNAR), tasas y distribución de faltantes, disponibilidad de auxiliares (incluidos registros administrativos), y diseño muestral complejo. En encuestas con diseño estratificado y por conglomerados, las prácticas más sólidas incorporan ponderadores, estratos y conglomerados tanto en la fase de imputación como en el análisis, o bien utilizan estrategias que respetan la estructura del diseño (p. ej., clases de donación dentro de estratos y dominios, o MI con modelos compatibles con el diseño). Cuando existen auxiliares de alta calidad, la combinación de registros administrativos con donación vecinal local mejora notablemente la coherencia interna, mientras que en dominios pequeños y variables sensibles (ingresos, empleo, cuidado) la MI y el *predictive mean matching* reducen sesgos y mantienen la variabilidad.

Las experiencias de las ONE confirman estos puntos. En contextos censales con dificultades operativas no previstas (p. ej., pandemia, catástrofes naturales, etc.), la imputación de no respuesta total a nivel vivienda basada en donantes cercanos geográfica y contextualmente, complementada con supuestos conservadores para variables no observadas, permitió estabilizar distribuciones sin alterar estructuras demográficas clave. En encuestas continuas de empleo e ingresos, las medias condicionadas han sido suficientes cuando la no respuesta es baja y el cuestionario está bien encadenado; sin embargo, al aumentar la complejidad y la proporción de faltantes, las oficinas que migraron a MI o a donación con clases finas observaron mejoras en precisión y en la trazabilidad de la incertidumbre. Asimismo, pilotos con aprendizaje automático (p. ej., bosques aleatorios o XGBoost) han mostrado buen desempeño predictivo para variables categóricas y discretas, con dos condiciones críticas: i) fuerte gobernanza para evitar "cajas negras" y *leakage*; ii) evaluación rigurosa de sesgo y estabilidad por subpoblaciones, incorporando paradata y enfoque de género para no reproducir brechas.

Mirando hacia adelante, el mensaje central es doble. Primero, la imputación adecuada es una política de aseguramiento de calidad, no un remedio post-hoc: debe planificarse desde el diseño (GSBPM 5.2), con reglas claras de elegibilidad, jerarquías de métodos por variable, uso explícito de auxiliares, métricas de monitoreo (tasa de imputación, FMI, RVI), y protocolos de validación y *sensitivity analysis* (comparar estimaciones con/sin imputación, por dominios y subgrupos). Segundo, la región necesita lineamientos comunes y reutilizables: catálogos de clases de donación y variables auxiliares típicas por tema; plantillas de MI compatibles con diseños complejos; guías para integrar RRAA y paradata; y estándares de documentación y transparencia (*imputation flags*, tasas y reglas publicadas). Consolidar estas prácticas en un marco regional (con anexos reproducibles de código y ejemplos) permitirá a las ONEs mejorar la comparabilidad, reducir sesgos sistemáticos y comunicar con mayor claridad la incertidumbre asociada, fortaleciendo la credibilidad de las estadísticas oficiales.

BIBLIOGRAFÍA

BAPTISTA, F. K. R. C; HYPÓLITO, E. B.; CONDE F. Q. O tratamento das informações da PNAD Contínua. Textos para discussão. n. 61 . Rio de Janeiro, 2022. 25 p. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101953.pdf

Barnard, J., & Rubin, D. B. (1999). Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4), 948–955. https://doi.org/10.1093/biomet/86.4.948

Binder, D. A. (1996). Comment to articles by Rao, Fay and Rubin. *Journal of the American Statistical Association*, 91, 510–512.

Binder, D. A., & Sun, W. (1996). Frequency valid multiple imputation for surveys with complex designs, Business Survey Methods Division. Statistics, Canada.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, 15(4), 651–675. https://doi.org/10.1080/1070551 0802339072

CHAMBERS, R. L. (2006). Evaluation criteria for editing and imputation in EUREDIT. In *Statistical Data Editing, Volume No. 3, Impact on Data Quality*, U. E. S. Division (ed), 11. New York and Geneva: United Nations Statistical Commission, and United Nations Economic Commission for Europe.

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation* (Vol. 563). John Wiley & Sons.

Departamento Administrativo Nacional de Estadística (DANE). (2021). *Documento Metodológico - Censo Nacional de Población y Vivienda (CNPV) 2018*. Recuperado de https://microdatos. dane.gov.co/index.php/catalog/643/download/21334.

Donza, E. (2013). Método de imputación de la no respuesta en las preguntas de ingresos en la Encuesta Permanente de Hogares. Gran Buenos Aires 1990 - 2010. *X Jornadas de Sociología*. Facultad de Ciencias Sociales, Universidad de Buenos Aires, (págs. 1 - 25). Buenos Aires.

Eltinge, J. L., & Luery, D. M. (2003). Imputation in Three Federal Statistical Agencies. Presentation at the.

FREITAS, M. P. S. de; ANTONACI, G. de A. Sistema Integrado de Pesquisas Domiciliares: Amostra mestra 2010 e amostra da PNAD Contínua. *Texto para discussão* n. 50. Rio de Janeiro, 2014.

HYPÓLITO, E. B. *Erros não amostrais em pesquisas domiciliares: impactos na qualidade*. 2020. Tese (Doutorado) – Escola Nacional de Ciências Estatísticas (Ence), Rio de Janeiro, RJ, 2020.

IBGE (2015). Política de revisão de dados divulgados das operações estatísticas do IBGE: Rio de Janeiro, Brazil, https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=298009

IBGE. *Pesquisa Nacional por Amostra de Domicílios Contínua. Notas metodológicas. Volume 1*. Diretoria de Pesquisas. Coordenação de Trabalho e Rendimento. Rio de Janeiro, 2014. Disponível em:https://www.ibge.gov.br/estatisticas/sociais/trabalho/9171-pesquisa-nacional -por-amostra-de-domicilios-continua-mensal.html?=&t=downloads

IBGE. *Pesquisa Nacional por Amostra de Domicílios Contínua. Notas técnicas.* Versão 1.7. Diretoria de Pesquisas Coordenação de Trabalho e Rendimento. Rio de Janeiro, 2020. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101708_notas_tecnicas.pdf

Instituto Nacional de Estadística y Censos (INEC). (2021). *Diseño muestral de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU).*

Instituto Nacional de Estadística y Geografía (INEGI). (2021). *Censo de Población y Vivienda 2020. Marco conceptual.* INEGI.

Instituto Nacional de Estadística y Geografía (INEGI). (2023). *Encuesta Nacional de Ocupación y Empleo. Cómo se hace la ENOE: métodos y procedimientos* (3.ª ed.).

Kish, L. (1965). *Survey sampling*. John Wiley & Sons.

Lin, & Tsai. (2019). Missing value imputation: a review and analysis of the literature (2006 - 2017). *Artificial Intelligence Review*, 1-23.

Little, R. J. A., & Rubin, D. B. (2002). Bayes and multiple imputation. *Statistical analysis with missing data* (pp. 200–220). Wiley Online Library.

Medina, F., & Galván, M. (2007). Imputación de datos: teoría y práctica. Cepal.

Medina, F., & Galván, M. (2007). Imputación de datos: Teoría y práctica. Cepal.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., & Verbeke, G. (2014). *Handbook of missing data methodology*. CRC Press.

ONU. *Glossary of terms on statistical data editing.* Conference of european statisticians methodological material. Genebra, 2000. Disponível em https://ec.europa.eu/ eurostat/ramon/statmanuals/files/un_editing_glossary_2000.pdf

Restrepo, M., & Marín, J. (2012). Imputación de ingresos en la Gran Encuesta Integrada de Hogares (geih) de 2010. *Revista Desarrollo y Sociedad*, 219-243.

Rosati, G. (2021). Métodos de Machine Learning como alternativa para la imputación de datos perdidos. Un ejercicio en base a la Encuesta Permanente de Hogares. *Estudios del trabajo*, 1-24

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, 4(3), 227–241. https://doi.org/10.1177/1536867X0400400304

Rubin, D. B. (1988). An overview of multiple imputation. *Proceedings of the survey research methods section of the American statistical association* (Vol. 79, pp. 84). Citeseer.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Servicio de Rentas Internas (SRI). (2022). Servicio de Rentas Internas (SRI). Obtenido de https://www.sri.gob.ec/ruc-personas-naturales#%C2%BFqu%C3%A9-es

Steele, R. J., Wang, N., & Raftery, A. E. (2010). Inference from multiple imputation for missing data using mixtures of normals. *Statistical Methodology*, 7(3), 351–365. https://doi.org/10.1016/j.stamet.2010.03.002

Sthamer, C. (2021). Multiple imputation through machine learning in a survey of sport clubs Organisation: Statistics Poland.

Tlamelo, E., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, 1-37.

UNECE. (2022). *Machine Learning Imputation for Social Surveys: Random Forest imputation of ONS' Household Financial Survey*. MODERNSTATS, 1-12.

United Nations Economic Commission for Europe. (2025). *Generic Statistical Business Process Model (GSBPM)*, version 5.2 (CES endorsed). UNECE.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC Press.