ISSN 2545-7179

Encuesta Nacional de Gastos de los Hogares 2017-2018

# NOTA TÉCNICA

## Metodología de imputación

Mayo de 2020

NOTAS TÉCNICAS INDEC N° 5

Ministerio de Economía
Argentina

Instituto Nacional de Estadística y Censos
República Argentina

**Encuesta Nacional de Gastos de los Hogares 2017-2018**

**Metodología de imputación**

**Nota técnica n° 5 – Mayo de 2020**

Instituto Nacional de Estadística y Censos (INDEC)

**Dirección:** Marco Lavagna
**Dirección Técnica:** Pedro Ignacio Lines
**Dirección Nacional de Difusión y Comunicación:** María Silvina Viazzi

Esta publicación fue realizada por la **Dirección Nacional de Metodología Estadística**, a cargo de Gerardo Antonio Mitas, y el equipo de trabajo integrado por Isabel Alegre, Mariana Mendiburu y Gregorio García, con la colaboración de Ileana Penna y Carla Barreca, de la **Dirección de Estudios de Ingresos y Gastos de los Hogares**.

**Coordinación de Producción Gráfica y Editorial:** Marcelo Costanzo
**Diseño y diagramación:** Juan Garavaglia e Ignacio Pello
**Revisión y corrección:** Mariana Alonso, Horacio Barisani, Soledad Daffra y María Victoria Piñera

ISSN 2545-7179
ISBN 978-950-896-579-0

Instituto Nacional de Estadística y Censos - I.N.D.E.C.
Encuesta Nacional de Gastos de los Hogares 2017-2018 : metodología de imputación : nota técnica n°5 / 1a ed . - Ciudad Autónoma de Buenos Aires : Instituto Nacional de Estadística y Censos - INDEC, 2020.
Libro digital, PDF - (Notas técnicas ; 5)

Archivo Digital: descarga y online
ISBN 978-950-896-579-0

1. Encuestas. 2. Estadísticas. 3. Gastos. I. Título.
CDD 318



Buenos Aires, mayo de 2020

# Índice

1. Introducción ..... 3
2. Objetivos y supuestos asumidos para el tratamiento de datos faltantes ..... 4
3. Variables de la encuesta sujetas a imputación ..... 5
4. Metodología y procedimientos de imputación ..... 6
5. Determinación de las clases de imputación para el método Hot Deck ..... 7
6. Selección del donante en el método Hot Deck ..... 8
7. Imputación de gastos en los cuestionarios 3 y 4 ..... 9
7.1 Imputación de gasto en servicios públicos y comida fuera del hogar ..... 9
7.2 Imputación de gastos de transporte público ..... 11
7.3 Imputación de gastos personales con no respuesta total ..... 13
8. Calidad de las imputaciones en los gastos ..... 16
8.1 Estructura de los gastos personales ..... 17
8.2 Efectos de la imputación de los gastos personales ..... 18
8.3 Proporción de gastos nulos y repetición de donantes ..... 20
9. Imputación de ingreso ..... 22
9.1 Ingreso por Ocupación principal, Jubilaciones y pensiones, y Ocupación secundaria ..... 23
9.2 Ingreso de ocupaciones anteriores y otras fuentes ..... 26
10. Calidad de las imputaciones en los ingresos ..... 26
10.1 Efectos de la imputación del ingreso asalariado ..... 27
10.2 Repetición de donantes ..... 29
11. Síntesis y recomendaciones para el uso de los datos con imputaciones ..... 30
Referencias ..... 33
Glosario ..... 34
Anexo I. Variables que jerarquizan los árboles de regresión para gasto ..... 37
Anexo II. Variables que jerarquizan los árboles de regresión para asalariados ..... 41

# 1. Introducción

El interés principal de encuestas como la Encuesta Nacional de Gastos de los Hogares (ENGHo) es hacer inferencia sobre parámetros poblacionales como totales (por ejemplo, el total de gasto de consumo de los hogares en alimentos y bebidas no alcohólicas), promedios (por ejemplo, el gasto medio mensual per cápita), o razones o cocientes (por ejemplo, el porcentaje del gasto de consumo en salud). Al igual que todo operativo estadístico, una vez finalizado, lleva a enfrentar el problema de los datos faltantes por no respuesta.

Este es un fenómeno común en casi todas las encuestas, pero indeseable, porque influye en la calidad de la inferencia. Por un lado, involucrar solo una porción de los datos en las estimaciones resulta en una pérdida de información que puede llevar a un incremento del error de muestreo debido a la reducción de la muestra. Un problema más delicado es que los estimadores basados en datos que incluyen faltantes por no respuesta pueden sufrir sesgo, que se agudiza si los que no responden a la encuesta y quienes lo hacen tienen un comportamiento diferente en los temas relevados. Esto es generalmente así, ya que la no respuesta raramente es fruto del azar.

Es habitual distinguir entre la no respuesta total y la no respuesta parcial. En el primer caso, la unidad seleccionada no responde ninguna pregunta de la encuesta, lo que se traduce en una pérdida total de información. El rechazo a participar de un individuo seleccionado o la inhabilidad para establecer el contacto con el informante son las causas más comunes de este tipo de no respuesta. En cambio, la no respuesta parcial ocurre cuando la ausencia de información se limita a algunas variables del estudio o ítems de la encuesta, ya sea porque el individuo no contesta algunas preguntas o porque la información que brinda es rechazada por una inconsistencia detectada durante la etapa de revisión y edición y, por lo tanto, considerada inválida.

La reponderación y la imputación son dos técnicas de corrección de la no respuesta una vez concluido el operativo. Por lo general, la reponderación es empleada para compensar la no respuesta total, consiste en incrementar los factores de expansión de los que responden la encuesta para tener en cuenta a los que no lo hacen.[1] En cambio, la imputación es el método más habitual que emplean las oficinas de estadística para el tratamiento de la no respuesta parcial. Consiste en reemplazar los datos faltantes por otros determinados para completarlos.

La presente nota técnica se centra en la descripción de la metodología y los procedimientos de imputación adoptados para el tratamiento de la no respuesta parcial en la ENGHo. Específicamente se describe la imputación de gastos del hogar en servicios públicos; de gastos personales; y del ingreso por fuente, relevado por la encuesta a cada miembro del hogar.

La nota incluye una evaluación de la calidad de la imputación sobre las principales variables de la encuesta afectadas por la no respuesta parcial y una serie de recomendaciones y advertencias para los usuarios de los datos sobre cómo tratar las estimaciones cuando participa en los cálculos alguna de ellas.

[1] Ver INDEC (2020). *Encuesta Nacional de Gastos de los Hogares. Factores de expansión, estimación y cálculo de los errores de muestreo*. Nota técnica n° 4. https://www.indec.gob.ar/ftp/cuadros/menusuperior/engho/engho2017_18_nota_tecnica_4.pdf.

## 2. Objetivos y supuestos asumidos para el tratamiento de datos faltantes

El principal objetivo de cualquier técnica de imputación es disminuir en lo posible el sesgo por no respuesta de los estimadores, y al mismo tiempo proveer un conjunto de datos completos que permita obtener resultados consistentes para los distintos tipos de análisis que surgen a partir de la encuesta.

De esta forma, la imputación habilita a seguir empleando el factor de expansión calculado para las estimaciones de la encuesta, evitando ponderadores alternativos asociados a cada variable con no respuesta parcial; lo que constituye una propiedad atractiva desde el punto de vista de los usuarios de los datos.

Sin embargo, para alcanzar este objetivo se deben asumir algunos riesgos dado que, por lo general, las técnicas se sostienen al emplear una serie de supuestos de difícil verificación en la práctica. Inicialmente, y desde el punto de vista del formalismo matemático y estadístico, se asume la existencia de un mecanismo de respuesta que es el responsable de los datos faltantes o perdidos en las variables de interés.

En rigor, ese mecanismo es desconocido y tratado por el estadístico a través de un conjunto de postulados probabilísticos o modelo probabilístico teórico. Estos le permiten asignar una probabilidad de respuesta, *a priori* desconocida, a cada unidad de la población; y con la ayuda de las hipótesis las vincula con las componentes de los datos observados y no observados, para caracterizarla.

Los mecanismos de respuesta[2] son de tres tipos: “no respuesta al azar”, “no respuesta completamente al azar” y “no respuesta no al azar”, y se diferencian según las hipótesis que los definen (Rubin, 1976; Schafer y Graham, 2002; Seaman y otros, 2013).

El que se adopta para la metodología de imputación en la ENGHo es el primero, que en la bibliografía aparece como MAR, por su sigla en inglés (*missing at random*). En este se postula un principio de independencia, el cual asevera que la probabilidad de respuesta no se ve afectada por la componente no observada de los datos y puede ser predicha por la información observada por la encuesta o el estudio.

Los mecanismos de respuesta MAR son del tipo “ignorable” porque su efecto sobre el sesgo, la validación, la precisión y la potencia predictiva pueden ser mitigados sin la necesidad de explicitar el modelo probabilístico que lo define. Por el contrario, los mecanismos bajo “no respuesta no al azar” o MNAR (*missing not at random*), son de tipo “no ignorable”, porque obligan a incorporar un modelo explícito y correcto para explicar la no respuesta, lo que puede llevar a la necesidad de sumar información de la componente no observada a través de fuentes externas a la encuesta.

Aceptar el supuesto MAR tiene implicancias sobre todo el proceso de imputación, ya que, si este se cumple, la distribución de la variable con datos faltantes condicional a la información que proveen las variables auxiliares es igual entre los que responden a la encuesta y los que no.

Como consecuencia, el valor a imputar a un dato faltante puede ser generado de la propia distribución de la variable a partir de los datos observados en la encuesta para dicha variable. Si bien el supuesto MAR no es verificable en la práctica, el estadístico puede tener indicios con los cuales sostenerlo; y cuanto más tenga, el modelo de imputación tendrá más potencia predictiva.

Como se puede advertir, gran parte del éxito de un procedimiento de imputación depende de la selección de las variables auxiliares predictivas a emplear en el modelo y de la relación de estas con la probabilidad de respuesta. Pero no son los únicos factores que intervienen, el nivel o la tasa de no respuesta que afecta a la variable a imputar y la calidad de los datos obtenidos en

[2] Indistintamente, también en la bibliografía se lo menciona como "mecanismo de no respuesta", "mecanismo de pérdida de datos" y "mecanismo de datos faltantes". Todos hacen referencia al mismo concepto.

el operativo también entran en juego al intentar disminuir el sesgo por datos faltantes en los resultados de una encuesta.

## 3. Variables de la encuesta sujetas a imputación

A diferencia de otras encuestas, la ENGHo tiene características especiales y generalmente se ve afectada por las altas tasas de no respuesta por parte de los hogares. Esto se debe, por ejemplo, a que las principales variables de estudio son el gasto y el ingreso de los hogares, temáticas sensibles a la respuesta.

También se releva información sobre variables demográficas, de ocupación y educación de los miembros del hogar, como así también sobre las características y el equipamiento de la vivienda, y de las cantidades consumidas en el hogar de alimentos y bebidas, entre otras.

Para obtener toda esta información, la encuesta emplea cinco cuestionarios:

- Cuestionario 1 (C1). Características de los hogares: se releva información para caracterizar el hogar y a cada uno de sus miembros a partir de aspectos socioeconómicos, demográficos, educacionales y de las características de la vivienda.

- Cuestionario 2 (C2). Gastos diarios: cada hogar registra el gasto en alimentos y en otros bienes y servicios de consumo frecuente, correspondiente a la semana en la que estuvo bajo estudio.

- Cuestionario 3 (C3). Gastos varios: se apuntan los gastos correspondientes a bienes y servicios adquiridos en períodos de tiempo comprendidos entre el mes y el año anteriores a la semana de la encuesta.

- Cuestionario 4 (C4). Gastos personales: las personas de 10 años y más asientan, durante la semana de referencia, el gasto en transporte público, comidas fuera del hogar, cigarrillos y otros gastos personales.

- Cuestionario 5 (C5). Ingresos: se consignan los ingresos que percibió cada uno de los miembros del hogar en los seis meses anteriores a la semana de la encuesta.[3]

Articular cada uno de los cuestionarios dentro de un hogar posee un alto grado de complejidad, tanto en los procedimientos que se aplican como en la estrategia indagatoria para los distintos requerimientos de información que se solicita. Compromete al hogar a responder con detalle los gastos diarios a nivel de hogar y a nivel personal de sus miembros, e incluso algunos por recordación para distintos períodos de referencia.[4]

Esta complejidad potencia la presencia de la no respuesta total y la parcial en varios de sus cuestionarios e ítems. Esto obliga a una tarea importante de control, edición y consistencia sobre la información durante el operativo y después. Aun así, queda una proporción de casos no resueltos o con faltantes de datos.

Como se adelantó en la introducción, en esta nota se describe la imputación de gastos consignados en el C3, referido a los gastos en servicios públicos (electricidad, gas y agua) y a la imputación de los gastos correspondientes al C4. Este cuestionario presenta un desafío doble, ya que se imputa el gasto en comida fuera del hogar y en transporte público en los cuestionarios con respuesta parcial, y también se

[3] Para los cuentapropistas o patrones agropecuarios y las sociedades jurídicas se consignan los ingresos de los doce meses anteriores a la encuesta.
[4] Para un detalle de los procedimientos y las definiciones metodológicas empleadas en la ENGHO 2017-2018, ver https://www.indec.gob.ar/ftp/cuadros/sociedad/engho_2017_2018_informe_gastos.pdf

realiza la imputación de los C4 sin respuesta total. Por último, se da un detalle de los procedimientos de imputación aplicados a los ingresos, correspondiente a la información relevada en el C5.

En el proceso de imputación para la ENHGo es crucial la información relevada en el C1, por dos motivos. Por un lado, porque en él se consigna si en el hogar los miembros realizan determinados gastos u obtienen algún ingreso, lo cual define la información faltante. El otro motivo es que permite la construcción de variables auxiliares predictivas, disponibles tanto para los que tienen la variable a imputar como para los que no, para incorporarlas al modelo de imputación.[5]

Una metodología de imputación utiliza diversas estrategias para completar o imputar la información faltante, que van de las más simples e intuitivas a las más complejas. Recurren a supuestos específicos propios de los métodos que se aplican en cada situación y que se describirán en los siguientes apartados.

## 4. Metodología y procedimientos de imputación

Al inicio de la etapa de imputación de una encuesta, la tarea central reside en seleccionar uno o varios procedimientos de imputación. Para que se consideren apropiados, estos procedimientos en lo posible deben poseer un grado de automatización que favorezca el flujo de procesamiento de la encuesta para alcanzar las estimaciones y se deben poder reproducir bajo cualquier circunstancia. Asimismo, debe poder emplearse en forma eficiente la información válida y disponible de las unidades que responden y de las que no lo hacen.

Los procedimientos se clasifican en dos grandes grupos: los determinísticos y los aleatorios. Los primeros son aquellos que, al repetir el mecanismo de imputación basado en el conjunto de unidades que responden, llevan siempre al mismo conjunto de datos completo. Ejemplos de este tipo son la imputación por el promedio, por el cociente, o por el vecino más cercano. También se incluyen en este grupo los métodos deductivos, o sea, aquellos que por información brindada por la unidad en otros ítems o variables del cuestionario permiten deducir el valor faltante.

En contraste, los procedimientos de imputación aleatorios (o estocásticos) son aquellos que llevan a un conjunto de datos completo distinto cada vez que el proceso de imputación es repetido. En general, estos métodos pueden verse como una imputación determinística más la suma de una componente aleatoria. Una de sus características es que tienden a preservar la distribución de la variable que requiere imputación, pero incorporan una componente adicional de error en la estimación debido al empleo de un mecanismo de imputación aleatorio (De Waal, Pannekoek y Scholtus, 2011; Chen y Haziza, 2019).

La complejidad señalada en el apartado anterior, los agregados de información y la relación existente entre la declaración del hogar informante y sus miembros en los distintos cuestionarios, sumados a la calidad de los datos capturados por la encuesta, induce a que la metodología general de imputación adoptada para la ENGHo involucre procedimientos tanto determinísticos como aleatorios.

Los métodos determinísticos empleados en la encuesta son: el promedio, el deductivo y, excepcionalmente, una fuente registral externa. En su mayoría fueron aplicados en clases de imputación definidas por variables geográficas: región, subregión, o provincia o jurisdicción, según la fuente a imputar.

El método aleatorio aplicado en la encuesta, cuando no se imputa por los precedentes, se basa en la imputación por Hot Deck. Su versión más general consiste en reemplazar el valor faltante de una o más variables de una unidad que no respondió (receptor) por valores observados de otra que tiene la información (donante) y que es similar al receptor con respecto a características observadas en ambos casos (Andridge y Little, 2010).

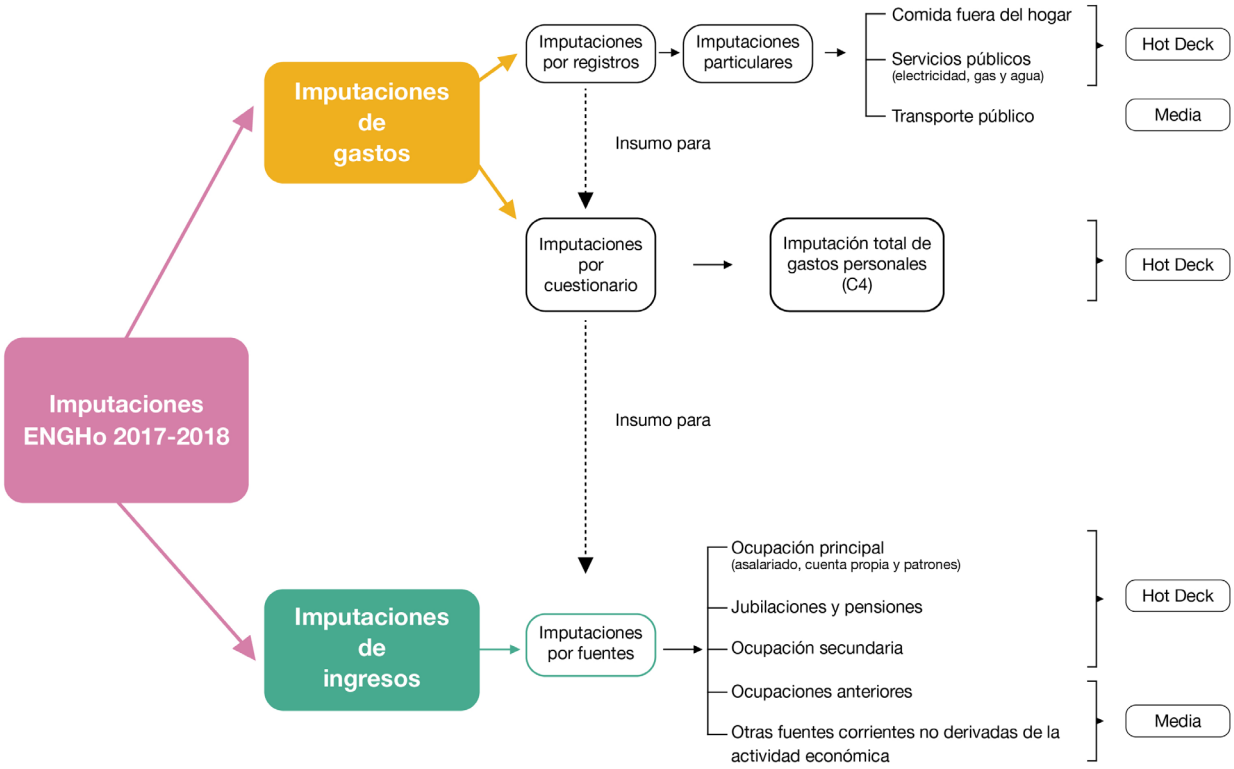
[5] En menor medida, los gastos registrados en los C 2 y 3 también contribuyen a la conformación de la información auxiliar que se emplea en los procedimientos de imputación.

El método por Hot Deck posee una serie de ventajas: es simple, existe un sin número de herramientas informáticas para ponerlo en práctica, y no obliga a asumir un modelo paramétrico para obtener el valor que se va a imputar.

Un detalle no menor en la ENGHo es que algunas de las variables a imputar pueden exhibir un gran número de observaciones con valor cero válido; bajo esta circunstancia, el método puede tratar la mixtura entre valores nulos y no nulos como donantes sin tener que recurrir a modelar especialmente este fenómeno dentro de la estrategia de imputación.

En el siguiente diagrama se presenta una síntesis de la operatoria empleada sobre los tipos de gastos y fuentes de ingresos que se imputaron y los métodos aplicados para completar los faltantes en la encuesta:

**Diagrama 1. Imputaciones de la ENGHo 2017-2018**




**Fuente**: INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

## 5. Determinación de las clases de imputación para el método Hot Deck

Para poner en práctica las distintas variantes del Hot Deck aleatorio se debe clasificar a los donantes y receptores en clases o celdas de imputación constituidas a partir de variables auxiliares propias de la encuesta. Una manera habitual para definirlas es a través de la clasificación cruzada de variables discretas, lo que obliga a categorizar las variables continuas y controlar el número de variables para no generar celdas con pocos donantes o ninguno.

Una posibilidad de remediar la falta de donantes en las celdas es agrupar categorías o eliminar variables hasta obtener donantes. Esto suele sumar una complejidad al método, conocido como "Hot Deck jerárquico", dado que es necesario determinar la manera de guiar y estructurar la jerarquía entre las variables, y cómo proceder ante la necesidad de realizar los agrupamientos entre celdas cuando no hay donantes.

Para la ENGHo, se buscó una metodología para la creación de las clases de imputación que permitiera utilizar variables auxiliares continuas sin llevar a cabo una categorización previa y garantizar donantes en las celdas sin una jerarquización *a priori* de las variables ni la confección de un algoritmo *ad hoc* para los reagrupamientos de las clases.

Existen distintas alternativas que cumplen con estos requisitos, entre ellas, las que emplean algoritmos de detección automática de interacciones (Kass, 1980), *k*-vecinos más cercanos (Altman, 1992) y los que aplican árboles de clasificación y regresión (Creel y Krotki, 2006; Saar-Tsechansky y Provost, 2007).

La metodología adoptada para la ENGHo se basa en árboles de regresión, ya que, de acuerdo con lo observado en distintas evaluaciones, brinda la mejor *performance* en los estimadores estudiados y satisface los requisitos señalados en los párrafos anteriores.

El árbol de regresión es un método no paramétrico de aprendizaje supervisado, que permite predecir una variable no observada mediante la partición recursiva del conjunto de datos al utilizar variables auxiliares. Entre sus bondades se encuentra la de permitir incluir efectos e interacciones entre las variables sin tener que incorporarlas de manera explícita al modelo predictivo.

El algoritmo parte de un conjunto de datos y, en cada paso, lo divide en dos grupos llamados "nodos", al elegir entre todas las variables auxiliares y sus posibles valores el corte óptimo que minimiza la dispersión de la variable respuesta. Este proceso continúa hasta que no es posible crear más divisiones. Luego, con un procedimiento conocido como "poda", se encuentra el subárbol óptimo que alcanza un compromiso entre la mayor homogeneidad y el menor número de nodos finales.

En la construcción de los árboles es habitual determinar algunas opciones para que el algoritmo tome decisiones, entre ellos: cantidad mínima de observaciones que puede tener un nodo para ser dividido, cantidad mínima de observaciones que puede tener un nodo final, profundidad máxima del árbol, criterio de poda, asignación a un nodo de las observaciones con valores faltantes en la variable auxiliar que lo define.

Para la imputación, la metodología no emplea los árboles como predictores, sino que se considera el conjunto de nodos finales del subárbol óptimo como clases de imputación homogéneas. Los receptores, o sea, los que tienen el valor faltante, son ubicados en la clase que les corresponde de acuerdo a las reglas de construcción del árbol según los valores que presentan sus variables auxiliares.

En todos los árboles de regresión construidos para la ENGHo se predice el logaritmo de la variable gasto o ingreso según corresponda. El conjunto de variables auxiliares, sumadas al algoritmo y los distintos parámetros empleados para la determinación de los nodos definitivos o clases de imputación, varía según el tipo de gasto o ingreso a imputar. La imputación se efectúa a nivel jurisdicción o región según el análisis de la información disponible y su calidad, la tasa de no respuesta al ítem o la variable, los criterios de bondad de ajuste, la homogeneidad interna en las clases y el número de donantes.

## 6. Selección del donante en el método Hot Deck

Una vez determinadas las celdas de imputación, el método Hot Deck lleva a seleccionar un donante para imputar el valor faltante. Existen distintas alternativas para la versión aleatoria del método; la adoptada para la ENGHo se basa en el "*k*-vecinos más cercano".

Esta modalidad emplea variables auxiliares, no necesariamente las utilizadas en los árboles de regresión, para definir una distancia entre el receptor y los posibles donantes. La selección se realiza al elegir con igual probabilidad entre las *k* unidades más cercanas al receptor en términos de distancia.

Con esto se busca acentuar la semejanza entre donante y receptor, pero manteniendo la aleatoriedad de un método estocástico, que preserva mejor las estructuras entre las variables bajo estudio.

La definición analítica de la distancia depende del tipo de variables que en conjunto la definen. Existen distintas alternativas según sean continuas, discretas o ambas. Es habitual estandarizar las variables para que la distancia no se vea afectada por las unidades de medida.

Toda la metodología de imputación de la ENGHo implícitamente asume que los que no responden y los que responden tienen la misma distribución en la variable a imputar dentro de cada celda. Es decir que, si las unidades se parecen en las variables auxiliares, esta información alcanza para inferir el valor de la variable a imputar. Esto implica suponer, como ya se adelantó, que el mecanismo de respuesta de la variable a imputar solo se asocia con las variables auxiliares observadas propias de la encuesta.

## 7. Imputación de gastos en los cuestionarios 3 y 4

Luego de la etapa de edición, con la información consistida, se analizó la falta de respuesta parcial en cada una de las variables de los C 3 y 4. En función de su relevancia, los niveles de completitud y calidad, y la información auxiliar disponible se decidió imputar:

- los gastos en servicios para la vivienda (electricidad, gas de red, gas envasado y agua) del C3
- los gastos en comida fuera del hogar del C4
- los gastos en transporte público (colectivo, subte, tren) de corta distancia del C4
- los gastos personales (C4) en su totalidad, si del proceso de análisis, edición y consistencia resultaba como un dato perdido o con no respuesta

La imputación se realizó a través de Hot Deck, de acuerdo a la metodología detallada en los apartados 4 a 6, excepto para el gasto en transporte público donde se imputó por el promedio.

### 7.1 Imputación de gasto en servicios públicos y comida fuera del hogar

Para los servicios públicos, los receptores se definieron como todos aquellos hogares que declararon en el C1 contar con el servicio correspondiente, respondieron el C3, pero no registraron en él el gasto y tampoco justificaron su ausencia.

En el caso de imputación de la variable “comida fuera del hogar”, los receptores fueron todas aquellas personas que declararon comer habitualmente fuera del hogar en el C1, respondieron el C4, pero no contaban con registro de gasto en dicho rubro ni con una aclaración que justificara su ausencia.

En ambos casos, los donantes se definieron como todos aquellos que cumplían con las dos primeras condiciones mencionadas para los receptores, pero que, a diferencia de estos, tenían registrado gasto positivo en el rubro en cuestión.

Para cada variable a imputar se elaboró un árbol de regresión por región geográfica, con excepción del GBA, donde se construyó uno para la CABA y otro para los partidos del GBA. Las variables explicativas fueron seleccionadas en función de su correlación con el gasto a imputar y, por ello, difieren para cada uno de los árboles de regresión empleados.

En el cuadro 1 se muestran las tasas de imputación para cada variable, que quedaron definidas en función de los criterios mencionados por región, en correspondencia con el nivel de aplicación del modelo de imputación,

**Cuadro 1. Tasa de imputación para servicios públicos y comida fuera del hogar, por región**

| Región | Tasa de imputación | | | | |
|---|---|---|---|---|---|
| | Electricidad | Gas de red | Gas envasado | Agua | Comida fuera del hogar |
| | | | % | | |
| CABA | 2,3 | 1,6 | 2,6 | 6,4 | 6,8 |
| Partidos del GBA | 2,0 | 3,2 | 4,8 | 37,9 | 31,1 |
| Cuyo | 1,3 | 2,7 | 8,2 | 31,1 | 25,9 |
| NEA | 4,8 | ---- | 10,3 | 10,6 | 28,9 |
| NOA | 1,1 | 2,8 | 3,7 | 29,3 | 22,1 |
| Pampeana ([1]) | 2,2 | 3,2 | 5,8 | 57,1 | 31,9 |
| Patagonia | 7,0 | 7,4 | 7,1 | 80,1 | 36,1 |

([1]) Incluye el resto de los partidos de la provincia de Buenos Aires que no conforman el GBA.
**Fuente**: INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

En cuanto al cálculo de la distancia en el método del *k*-vecinos más cercanos, en todos los casos, se emplearon variables continuas, asumiendo que poseen una alta correlación con las variables a imputar: gasto total del hogar del C2, semana de relevamiento y edad.

Este tipo de variables permite recurrir a la distancia euclídea. En cuanto al donante, se lo seleccionó de manera aleatoria entre los 3 vecinos más cercanos al receptor.

En el cuadro 2 se detallan las variables propuestas para la construcción de los árboles y las utilizadas para cálculo de la distancia. Se especifica el nivel de gasto, a nivel de hogar o personal.

**Cuadro 2. Variables auxiliares propuestas para los árboles y las utilizadas en la distancia de vecino más cercano (VMC), según el nivel de gasto**

| Nivel de gasto | Variables propuestas para los árboles ([1]) | Utilizadas en | Variables VMC |
|---|---|---|---|
| Hogar | Jurisdicción (25)<br>Trimestre (4)<br>Tipo de hogar (4)<br>Tipo de vivienda (3)<br>Cantidad de miembros del hogar<br>Cantidad de habitaciones de uso exclusivo<br>Situación ocupacional del jefe o la jefa del hogar (3) | Electricidad<br>Gas de red<br>Gas envasado<br>Agua | Logaritmo del gasto total en el C2<br>Semana de relevamiento en campo (52) |
| | Tipo de combustible utilizado para cocinar (4 o 2)<br>Tipo de calefacción (4 o 2) | Electricidad<br>Gas de red<br>Gas envasado | |
| | Nivel educativo del jefe o la jefa del hogar (8) | Electricidad<br>Agua | |
| | Tenencia de aire acondicionado (3) | Electricidad | |
| | Tenencia de baño (2)<br>Tipo de descarga del baño (3)<br>Cantidad de baños<br>Tenencia de jardín (3)<br>Tenencia de piscina (3)<br>Tenencia de huerta (3)<br>Condición de actividad del jefe o la jefa del hogar (3) | Agua | |
| Personal | Sexo (2)<br>Grupo de edad (9)<br>Nivel educativo (5)<br>Situación conyugal (5)<br>Tipo de hogar (4)<br>Tipo de vivienda (3)<br>Condición de actividad (3)<br>Calificación ocupacional (8)<br>Jerarquía ocupacional (8)<br>Horas trabajadas<br>Jurisdicción (25)<br>Trimestre (4) | Comida fuera del hogar | Edad<br>Semana de relevamiento en campo (52) |

([1]) Entre paréntesis se indica la cantidad de categorías, cuando corresponda.
**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

## 7.2 Imputación de gastos de transporte público

Para la no respuesta parcial en la variable transporte público se emplearon métodos determinísticos; en la mayor parte de los casos, imputación por el promedio, diferenciando el medio de transporte utilizado y la jurisdicción. Se consideraron receptores de la imputación del gasto en transporte público (colectivo, tren y subte) a todas aquellas personas que declararon habitualidad en el uso del medio de transporte correspondiente en el C1, respondieron el C4, pero no declararon gastos en ese rubro, no justificaron su falta ni tenían registros de transporte en la página de bienes y servicios recibidos gratuitamente.

En los cuadros 3 y 4, que se presentan a continuación, se indican las tasas de imputación de los medios de transporte al nivel en que se realizó la imputación.

**Cuadro 3. Tasa de imputación del medio de transporte colectivo, por jurisdicción**

| Jurisdicción | Tasa de imputación |
|---|---|
| | % |
| CABA | 2,8 |
| Partidos del GBA | 11,8 |
| Resto de Buenos Aires | 11,8 |
| Catamarca | 9,6 |
| Córdoba | 15,6 |
| Corrientes | 15,2 |
| Chaco | 23,7 |
| Chubut | 10,7 |
| Entre Ríos | 11,7 |
| Formosa | 9,4 |
| Jujuy | 11,3 |
| La Pampa | 4,4 |
| La Rioja | 8,1 |
| Mendoza | 5,9 |
| Misiones | 8,6 |
| Neuquén | 14,7 |
| Río Negro | 8,2 |
| Salta | 8,8 |
| San Juan | 32,7 |
| San Luis | 16,2 |
| Santa Cruz | 7,4 |
| Santa Fe | 11,9 |
| Santiago del Estero | 15,3 |
| Tucumán | 9,4 |
| Tierra del Fuego | 12,7 |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

**Cuadro 4. Tasas de imputación para los medios de transporte tren en las jurisdicciones que corresponden**

| Jurisdicción | Tasa de imputación | |
|---|---|---|
| | Subte | Tren |
| | % | |
| CABA | 2,8 | 1,2 |
| Partidos del GBA | 2,6 | 8,0 |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Para determinar el valor a imputar en un faltante de gasto en colectivo se decidió emplear:

- *Promedios mensuales* del gasto en el boleto de colectivo por jurisdicción multiplicado por la cantidad declarada como habitual en el C1 cuando existían datos disponibles para todos los meses y la variación mensual, en caso de ser negativa, no superaba una disminución del 10%.
- *Promedios trimestrales* del gasto en el boleto de colectivo por jurisdicción multiplicado por la cantidad declarada como habitual en el C1 en el caso de que no fuera posible

usar los promedios mensuales (en función del criterio anterior), la información trimestral disponible no tenía datos faltantes y no se presentaban disminuciones mayores al 10% de un trimestre al otro.

- *Información externa* del precio del boleto de colectivo, según índice de precios al consumidor (IPC), multiplicado por la cantidad declarada como habitual en el C1 cuando faltaba algún dato trimestral o cuando había caídas de más del 10% entre los promedios de dos trimestres consecutivos.

Como resultado se imputa el gasto con:

- *Promedios mensuales* a las siguientes jurisdicciones: Chubut, CABA, Jujuy, La Rioja, Mendoza, partidos del GBA, Salta, Santiago del Estero y Tucumán.
- *Promedios trimestrales* en las siguientes jurisdicciones: provincia de Buenos Aires exceptuando los partidos del GBA, Catamarca, Córdoba, Corrientes, Entre Ríos, Misiones, Neuquén, San Luis y Santa Fe.
- *Promedio mensual* del precio de colectivo de acuerdo a la información brindada por la Dirección de Índices de Precios de Consumo del INDEC a: Chaco, Formosa, La Pampa, Río Negro, San Juan, Santa Cruz y Tierra del Fuego.

Para el cálculo de los promedios del gasto que salen de la propia encuesta, se eliminaron los valores extremos que superaban el percentil 95 de la cola superior de la distribución.

Para imputar los gastos en subte, se utilizó el promedio mensual del gasto en boleto de subte según jurisdicciones, para la CABA y los partidos del GBA, multiplicado por la cantidad declarada como habitual en el C1.

El cálculo se hizo por separado para la CABA y los partidos del GBA, porque, aunque el subte solo se encuentre en la CABA, al existir la tarifa de transporte integrada, se estima que aquellos que se trasladan desde los partidos del GBA ya tomaron previamente algún transporte, por lo que el precio que pagan por el subte es menor. El gasto faltante de aquellas observaciones de otras jurisdicciones que declararon habitualidad en el uso de subte, pero no declararon monto, es imputado con el valor promedio calculado para la CABA.

Los gastos en tren fueron imputados con el promedio mensual del gasto en boleto de tren según jurisdicción, para la CABA y los partidos del GBA, multiplicado por la cantidad declarada como habitual en el C1.

## 7.3 Imputación de gastos personales con no respuesta total

Luego de haberse realizado la imputación de los gastos parciales, se procedió a imputar los C4 con falta de información total o los resultantes del proceso de edición que llevó a anularlos por información errónea o inconsistente. En este caso se imputa a cada persona que no contesta un cuestionario completo informado por otro individuo.

La selección de este cuestionario se realiza dentro de una celda de imputación construida con árboles de regresión a través del método *k*-vecinos más cercanos. Al imputar un cuestionario completo muchas variables de la encuesta se imputan simultáneamente usando un único donante, con lo que se busca preservar en lo posible las relaciones entre ellas. En el cuadro 5 se hace referencia a la tasa de imputación del C4 por jurisdicción.

**Cuadro 5. Tasa de imputación del C4, por jurisdicción**

| Jurisdicción | Tasa de imputación |
|---|---|
| | % |
| CABA | 7,4 |
| Partidos del GBA | 16,9 |
| Resto de Buenos Aires | 9,6 |
| Catamarca | 4,1 |
| Córdoba | 16,0 |
| Corrientes | 16,6 |
| Chaco | 39,6 |
| Chubut | 6,2 |
| Entre Ríos | 15,9 |
| Formosa | 5,6 |
| Jujuy | 9,0 |
| La Pampa | 10,7 |
| La Rioja | 10,0 |
| Mendoza | 12,1 |
| Misiones | 15,8 |
| Neuquén | 21,1 |
| Río Negro | 7,8 |
| Salta | 16,8 |
| San Juan | 18,2 |
| San Luis | 11,8 |
| Santa Cruz | 31,5 |
| Santa Fe | 12,1 |
| Santiago del Estero | 17,5 |
| Tucumán | 14,8 |
| Tierra del Fuego | 17,7 |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

La variable a imputar es el gasto personal, que puede exhibir un gran número de observaciones con valor cero válido o nulo. En efecto, cuando los miembros del hogar informan que en la semana de referencia no realizaron ningún gasto en los bienes y servicios relevados en el C4, y esta información es fidedigna, los cuestionarios se consideran válidos, pero con gasto cero.

La validación de esta información fue realizada en la etapa de edición de la encuesta. Estos valores deben ser contemplados como posibles en la distribución de la variable; por lo tanto, también son tratados como donantes.

En el cuadro 6 se presenta el porcentaje del C4 con gasto cero, por jurisdicción.

**Cuadro 6. Porcentaje de donantes con gasto total cero en el C4, por jurisdicción**

| Jurisdicción | Gasto cero |
|---|---|
| | % |
| CABA | 13,9 |
| Partidos del GBA | 25,9 |
| Resto de Buenos Aires | 23,8 |
| Catamarca | 21,6 |
| Córdoba | 26,3 |
| Corrientes | 47,3 |
| Chaco | 53,0 |
| Chubut | 24,4 |
| Entre Ríos | 43,1 |
| Formosa | 30,0 |
| Jujuy | 22,1 |
| La Pampa | 41,6 |
| La Rioja | 35,2 |
| Mendoza | 23,2 |
| Misiones | 39,4 |
| Neuquén | 23,7 |
| Río Negro | 31,9 |
| Salta | 27,2 |
| San Juan | 35,5 |
| San Luis | 30,4 |
| Santa Cruz | 51,5 |
| Santa Fe | 31,7 |
| Santiago del Estero | 23,8 |
| Tucumán | 28,7 |
| Tierra del Fuego | 30,5 |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Para la elección de las variables auxiliares que participan en los árboles de regresión, se decidió aplicar un "principio de parsimonia", en el sentido de dejar pocas que resultaran significativas para explicar el gasto total del C4.

Con este objetivo se analizó la correlación de las variables utilizadas en las ENGHo anteriores y otras propias de la encuesta 2018, con la variable respuesta logaritmo del gasto total del C4. Resultaron seleccionadas aquellas con correlaciones más altas, al asumir que es beneficioso para la construcción de los árboles. Una vez seleccionadas las variables explicativas, los árboles se construyeron por jurisdicción.

En el cuadro 7 se enumeran las variables seleccionadas propuestas para la construcción de los árboles y las variables utilizadas para calcular la distancia en el método *k*-vecinos más cercanos.

**Cuadro 7. Variables auxiliares propuestas para los árboles las utilizadas en la distancia de vecino más cercano (VMC)**

| Variables propuestas para los árboles ([1]) | Variables VMC |
|---|---|
| Aglomerado principal (2)<br>Asistencia a establecimiento educativo (2)<br>Cantidad de miembros del hogar<br>Cantidad de viajes en transporte público<br>Come afuera del hogar (2)<br>Condición de actividad (3)<br>Edad (9)<br>Estrato de área (5)<br>Gasto de consumo en el hogar per cápita<br>Jubilado o pensionado (2)<br>Nivel de educación (4)<br>Número de autos en el hogar (3)<br>Perceptor de ingreso (2)<br>Sexo (2)<br>Viaja en transporte público (2) | Edad<br>Semana de relevamiento en campo (52) |

([1]) Entre paréntesis se indica la cantidad de categorías, cuando corresponda.
**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

En la elección de las variables para el cálculo de distancia en el método *k*-vecinos más cercanos se buscó el resguardo de las relaciones entre donante y receptor por edad, utilizando aquí la variable sin categorizar (en años cumplidos) y procurando, a la vez, no alejarse de la semana de referencia del receptor. Por ejemplo, se buscó evitar que un individuo con gasto en transporte privado o en tabaco y derivados fuese donante de un niño de 10 años.

Al ser variables continuas, se emplea la distancia euclídea. Para que ambas variables tengan la misma importancia, a cada una se la estandariza a través del promedio y el desvío dentro de la celda.

En Anexo I. "Variables que jerarquizan los árboles de regresión para gasto", se presentan las variables con las que se conformaron las celdas de imputación en cada jurisdicción. Están señaladas las variables que definen el primer nivel de corte de cada árbol, las correspondientes al segundo nivel de corte y a los niveles posteriores en conjunto.

Para la construcción de los árboles se utilizó el procedimiento HPSPLIT incluido en la versión de SAS v9.4. Se determinó en 20 el número mínimo de observaciones que debe contener cada nodo resultante de una división para que esta se considere, para garantizar al menos esa cantidad de donantes en cada celda.

El criterio de poda elegido fue el de costo complejidad y el árbol óptimo se encontró usando validación cruzada, que es la selección preestablecida por el procedimiento.

## 8. Calidad de las imputaciones en los gastos

Para evaluar que los valores de gasto imputados son razonables y respetan las hipótesis en las que se basan, se realizan algunas comprobaciones, comparando distintas medidas calculadas antes y después de la imputación.

En primera instancia se estudia la estructura de la distribución de los gastos parciales, luego se evalúa la media o promedio de la variable imputada en forma univariada, tomando solo los gastos no nulos. A continuación, la relación bivariada entre el gasto y distintas variables por medio de la correlación. También se incluye el cálculo del coeficiente de determinación $\boldsymbol{R^2}$ que surge de ajustar un modelo de regresión lineal

entre el gasto y otras variables relevantes para evaluar el resultado de la imputación en una dimensión multivariada. Por último, como indicadores adicionales de calidad, se presentan la proporción de gastos nulos y la cantidad de veces que se emplea una observación como donante.

## 8.1 Estructura de los gastos personales

Los gastos registrados en el C4 (alimentos y bebidas consumidas fuera del hogar, transporte diario en medios públicos, combustible, peaje y estacionamiento diario, artículos de librería, entradas a espectáculos y juegos de azar, cigarrillos, etc.) se agruparon en tres rubros: comida, transporte y otros gastos.

La imputación del C4 se lleva a cabo asignando a un individuo que no lo respondió, el cuestionario completo de otra persona que es similar en términos de las variables auxiliares. De esta forma se garantiza la coherencia interna de los cuestionarios. Pero, además, es importante comprobar que la imputación no altera la proporción de gasto en cada rubro, es decir, la estructura de gastos personales.

El método propuesto por cuestionario completo fue evaluado a partir de distintas simulaciones con los datos de la ENGHo 2004-2005, estudiando el desempeño de los estimadores para las proporciones de los gastos personales agrupados, en distintas subpoblaciones. Los resultados obtenidos permitieron concluir que el método no alteraba la estructura de los gastos.

Para ver el efecto que tiene la imputación sobre la estructura de gastos en la ENGHo 2017-2018, se calculó la proporción ponderada de gasto en cada rubro por jurisdicción, antes y después de la imputación.

Dichas proporciones y sus diferencias se muestran en el cuadro 8. En él se puede observar que las diferencias entre las proporciones resultaron pequeñas en los tres rubros, las mayores corresponden al gasto en comida, en Córdoba y Santa Cruz, y a otros gastos, también en Santa Cruz. Es decir, en líneas generales la imputación no altera la estructura de gastos en las jurisdicciones.

**Cuadro 8. Proporción de gastos antes y después de la imputación, por jurisdicción**

| Jurisdicción | Proporción de gasto | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Comida | | | Transporte | | | Otros gastos | | |
| | Antes | Después | Diferencia | Antes | Después | Diferencia | Antes | Después | Diferencia |
| | % | | | | | | | | |
| CABA | 45,9 | 45,9 | 0,0 | 37,0 | 37,1 | -0,1 | 17,1 | 17,0 | 0,1 |
| Partidos del GBA | 30,5 | 30,4 | 0,1 | 52,6 | 52,4 | 0,2 | 16,9 | 17,1 | -0,2 |
| Resto de Buenos Aires | 26,1 | 26,4 | -0,3 | 54,5 | 54,2 | 0,3 | 19,3 | 19,4 | -0,1 |
| Catamarca | 24,5 | 24,7 | -0,2 | 53,3 | 53,3 | 0,0 | 22,2 | 22,0 | 0,2 |
| Córdoba | 29,3 | 31,3 | -2,0 | 46,2 | 45,4 | 0,8 | 24,5 | 23,2 | 1,3 |
| Corrientes | 22,3 | 24,2 | -1,9 | 55,1 | 53,5 | 1,6 | 22,7 | 22,4 | 0,3 |
| Chaco | 18,6 | 20,5 | -1,9 | 55,8 | 54,1 | 1,7 | 25,6 | 25,4 | 0,2 |
| Chubut | 15,7 | 16,3 | -0,6 | 61,0 | 60,2 | 0,8 | 23,3 | 23,6 | -0,3 |
| Entre Ríos | 20,2 | 21,9 | -1,7 | 59,6 | 58,2 | 1,4 | 20,3 | 19,9 | 0,4 |
| Formosa | 22,5 | 23,9 | -1,4 | 47,2 | 46,6 | 0,6 | 30,3 | 29,5 | 0,8 |
| Jujuy | 23,6 | 24,1 | -0,5 | 49,6 | 49,6 | 0,0 | 26,8 | 26,3 | 0,5 |
| La Pampa | 23,0 | 23,3 | -0,3 | 57,1 | 57,6 | -0,5 | 20,0 | 19,1 | 0,9 |
| La Rioja | 25,6 | 26,2 | -0,6 | 58,3 | 58,0 | 0,3 | 16,1 | 15,8 | 0,3 |
| Mendoza | 23,6 | 23,6 | 0,0 | 57,4 | 57,0 | 0,4 | 19,0 | 19,4 | -0,4 |
| Misiones | 15,1 | 14,3 | 0,8 | 64,3 | 64,8 | -0,5 | 20,6 | 20,8 | -0,2 |
| Neuquén | 24,8 | 25,1 | -0,3 | 49,1 | 48,6 | 0,5 | 26,1 | 26,3 | -0,2 |
| Río Negro | 21,7 | 21,7 | 0,0 | 60,9 | 61,1 | -0,2 | 17,4 | 17,3 | 0,1 |
| Salta | 29,9 | 29,2 | 0,7 | 47,9 | 48,6 | -0,7 | 22,2 | 22,3 | -0,1 |
| San Juan | 22,4 | 22,0 | 0,4 | 60,0 | 59,3 | 0,7 | 17,6 | 18,7 | -1,1 |
| San Luis | 24,0 | 24,5 | -0,5 | 54,2 | 53,4 | 0,8 | 21,8 | 22,1 | -0,3 |
| Santa Cruz | 19,8 | 23,1 | -3,3 | 54,8 | 54,9 | -0,1 | 25,4 | 22,0 | 3,4 |
| Santa Fe | 29,2 | 30,6 | -1,4 | 51,6 | 50,1 | 1,5 | 19,1 | 19,3 | -0,2 |
| Santiago del Estero | 23,0 | 22,4 | 0,6 | 47,8 | 48,1 | -0,3 | 29,2 | 29,6 | -0,4 |
| Tucumán | 27,2 | 28,6 | -1,4 | 54,2 | 53,1 | 1,1 | 18,5 | 18,4 | 0,1 |
| Tierra del Fuego | 29,8 | 29,9 | -0,1 | 51,4 | 51,0 | 0,4 | 18,8 | 19,1 | -0,3 |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

## 8.2 Efectos de la imputación de los gastos personales

El impacto de la imputación del gasto se evalúa a través de la estimación del gasto medio no nulo por jurisdicción. En el cuadro 9 se muestran el gasto promedio estimado antes y después de la imputación, la diferencia entre ellos y, por último, la diferencia relativa al gasto promedio observado expresada en porcentaje. En todos los casos, las medidas fueron calculadas con los valores ponderados, surgidos de los estimadores por expansión empleados para la encuesta.

**Cuadro 9. Estimación del gasto medio no nulo antes y después de la imputación, por jurisdicción**

| Jurisdicción | Gasto medio no nulo | | Diferencia | |
|---|---|---|---|---|
| | Antes | Después | absoluta | relativa |
| | Pesos | | | % |
| CABA | 3.097 | 3.089 | 8 | 0,3 |
| Partidos del GBA | 2.267 | 2.165 | 102 | 4,5 |
| Resto de Buenos Aires | 2.274 | 2.301 | -27 | -1,2 |
| Catamarca | 1.639 | 1.653 | -14 | -0,9 |
| Córdoba | 2.275 | 2.251 | 24 | 1,1 |
| Corrientes | 1.664 | 1.684 | -20 | -1,2 |
| Chaco | 1.606 | 1.608 | -2 | -0,1 |
| Chubut | 2.056 | 2.032 | 24 | 1,2 |
| Entre Ríos | 2.240 | 2.228 | 12 | 0,5 |
| Formosa | 1.148 | 1.171 | -23 | -2,0 |
| Jujuy | 1.552 | 1.550 | 2 | 0,1 |
| La Pampa | 2.391 | 2.507 | -116 | -4,9 |
| La Rioja | 1.696 | 1.697 | -1 | -0,1 |
| Mendoza | 2.056 | 2.006 | 50 | 2,4 |
| Misiones | 1.556 | 1.530 | 26 | 1,7 |
| Neuquén | 2.720 | 2.647 | 73 | 2,7 |
| Río Negro | 1.945 | 1.978 | -33 | -1,7 |
| Salta | 1.476 | 1.446 | 30 | 2,0 |
| San Juan | 1.884 | 1.790 | 94 | 5,0 |
| San Luis | 1.907 | 1.913 | -6 | -0,3 |
| Santa Cruz | 2.517 | 2.503 | 14 | 0,6 |
| Santa Fe | 2.785 | 2.701 | 84 | 3,0 |
| Santiago del Estero | 1.303 | 1271 | 32 | 2,5 |
| Tucumán | 1.758 | 1.725 | 33 | 1,9 |
| Tierra del Fuego | 3.223 | 3.121 | 102 | 3,2 |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Las diferencias relativas no superan el 5% y no hay un patrón, dado que en algunas jurisdicciones la imputación provoca un aumento en la media y en otras, una disminución.

A continuación, para evaluar una dimensión bivariada del efecto de la imputación de los gastos personales, el cuadro 10 presenta la correlación del total de estos gastos antes y después de la imputación con dos variables continuas, elegidas a modo ilustrativo: la edad y el gasto de consumo en el hogar. El interés de este análisis radica en ver si las correlaciones cambian como resultado de la imputación.

**Cuadro 10. Correlación de los gastos no nulos con la edad y con el gasto de consumo del hogar antes y después de la imputación, por jurisdicción**

| Jurisdicción | Correlación del gasto total del C4 | | | |
|---|---|---|---|---|
| | Edad | | Gasto de consumo del hogar | |
| | Antes | Después | Antes | Después |
| CABA | 0,06 | 0,07 | 0,36 | 0,34 |
| Partidos del GBA | 0,12 | 0,12 | 0,42 | 0,41 |
| Resto de Buenos Aires | 0,15 | 0,15 | 0,34 | 0,33 |
| Catamarca | 0,14 | 0,13 | 0,33 | 0,33 |
| Córdoba | 0,10 | 0,06 | 0,25 | 0,25 |
| Corrientes | 0,17 | 0,13 | 0,22 | 0,21 |
| Chaco | 0,21 | 0,16 | 0,21 | 0,11 |
| Chubut | 0,19 | 0,18 | 0,29 | 0,29 |
| Entre Ríos | 0,13 | 0,11 | 0,33 | 0,29 |
| Formosa | 0,14 | 0,12 | 0,25 | 0,24 |
| Jujuy | 0,14 | 0,13 | 0,29 | 0,28 |
| La Pampa | 0,18 | 0,16 | 0,31 | 0,28 |
| La Rioja | 0,24 | 0,24 | 0,31 | 0,29 |
| Mendoza | 0,12 | 0,12 | 0,43 | 0,42 |
| Misiones | 0,13 | 0,10 | 0,19 | 0,19 |
| Neuquén | 0,14 | 0,16 | 0,32 | 0,32 |
| Río Negro | 0,25 | 0,25 | 0,36 | 0,39 |
| Salta | 0,16 | 0,13 | 0,33 | 0,27 |
| San Juan | 0,10 | 0,06 | 0,25 | 0,24 |
| San Luis | 0,11 | 0,09 | 0,27 | 0,22 |
| Santa Cruz | 0,17 | 0,14 | 0,27 | 0,27 |
| Santa Fe | 0,12 | 0,11 | 0,32 | 0,32 |
| Santiago del Estero | 0,16 | 0,16 | 0,45 | 0,42 |
| Tucumán | 0,12 | 0,10 | 0,26 | 0,23 |
| Tierra del Fuego | 0,18 | 0,18 | 0,28 | 0,23 |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Se observa que en general la imputación no provoca cambios en las correlaciones. El caso más extremo es el de Chaco, donde la correlación con el gasto de consumo del hogar disminuye a la mitad, como probable consecuencia de la alta tasa de faltantes en la provincia que afecta a la calidad de la imputación.

Por último, para considerar una dimensión multivariada del gasto se calculó el coeficiente de determinación de la regresión del gasto con las variables explicativas efectivamente utilizadas en la construcción de los árboles. Esto se muestra en el cuadro 11.

**Cuadro 11. R cuadrado de la regresión de los gastos no nulos con las variables del árbol antes y después de la imputación, por jurisdicción**

| Jurisdicción | $R^2$ | |
|---|---|---|
| | Antes | Después |
| CABA | 0,36 | 0,35 |
| Partidos del GBA | 0,35 | 0,33 |
| Resto de Buenos Aires | 0,34 | 0,32 |
| Catamarca | 0,32 | 0,31 |
| Córdoba | 0,24 | 0,22 |
| Corrientes | 0,29 | 0,25 |
| Chaco | 0,27 | 0,21 |
| Chubut | 0,33 | 0,33 |
| Entre Ríos | 0,25 | 0,22 |
| Formosa | 0,23 | 0,22 |
| Jujuy | 0,31 | 0,29 |
| La Pampa | 0,30 | 0,26 |
| La Rioja | 0,33 | 0,31 |
| Mendoza | 0,34 | 0,32 |
| Misiones | 0,22 | 0,21 |
| Neuquén | 0,24 | 0,26 |
| Río Negro | 0,36 | 0,34 |
| Salta | 0,30 | 0,26 |
| San Juan | 0,23 | 0,20 |
| San Luis | 0,22 | 0,20 |
| Santa Cruz | 0,31 | 0,29 |
| Santa Fe | 0,27 | 0,25 |
| Santiago del Estero | 0,38 | 0,34 |
| Tucumán | 0,30 | 0,26 |
| Tierra del Fuego | 0,23 | 0,22 |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Se aprecia que el valor de $R^2$ es similar en todas las jurisdicciones antes y después de la imputación, es decir, se mantiene la relación entre el gasto y el conjunto de variables consideradas. La diferencia más notoria ocurre en Chaco donde disminuye el coeficiente, por la misma razón manifestada en los párrafos anteriores.

## 8.3 Proporción de gastos nulos y repetición de donantes

Debido a la magnitud que alcanzan los cuestionarios de personas de 10 años o más que manifiestan justificadamente no haber realizado gastos en la semana de referencia y que son considerados donantes, se evalúa el impacto de la imputación en la proporción ponderada de gasto nulo.

El cálculo de esta proporción se realiza por jurisdicción, antes y después de la imputación; dichas proporciones y sus diferencias se muestran en el cuadro 12.

Las diferencias entre las proporciones ponderadas son bajas, menores al 6%; las más altas se registran en Santa Cruz y San Juan. También se observa que, salvo en Neuquén, en ninguna jurisdicción la imputación provoca un aumento en la proporción de gastos nulos.

**Cuadro 12. Porcentaje ponderado de gastos nulos antes y después de la imputación, por jurisdicción**

| Jurisdicción | Gastos nulos | | |
|---|---|---|---|
| | Antes | Después | Diferencia |
| | | % | |
| CABA | 13,5 | 13,4 | 0,1 |
| Partidos del GBA | 24,7 | 24,4 | 0,3 |
| Resto de Buenos Aires | 22,0 | 21,1 | 0,9 |
| Catamarca | 19,2 | 18,7 | 0,5 |
| Córdoba | 30,3 | 27,8 | 2,5 |
| Corrientes | 45,3 | 42,7 | 2,6 |
| Chaco | 54,6 | 52,1 | 2,5 |
| Chubut | 25,5 | 24,9 | 0,6 |
| Entre Ríos | 42,9 | 40,5 | 2,4 |
| Formosa | 30,9 | 30,1 | 0,8 |
| Jujuy | 21,3 | 20,8 | 0,5 |
| La Pampa | 40,9 | 39,3 | 1,6 |
| La Rioja | 35,6 | 34,7 | 0,9 |
| Mendoza | 24,4 | 24,0 | 0,4 |
| Misiones | 38,3 | 36,2 | 2,1 |
| Neuquén | 23,0 | 24,7 | -1,7 |
| Río Negro | 33,0 | 32,1 | 0,9 |
| Salta | 26,5 | 26,2 | 0,3 |
| San Juan | 35,7 | 30,8 | 4,9 |
| San Luis | 27,1 | 25,6 | 1,5 |
| Santa Cruz | 52,1 | 46,4 | 5,7 |
| Santa Fe | 30,7 | 29,2 | 1,5 |
| Santiago del Estero | 23,2 | 21,5 | 1,7 |
| Tucumán | 27,4 | 27,2 | 0,2 |
| Tierra del Fuego | 30,1 | 28,1 | 2,0 |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

El método de selección de donantes permite que estos se utilicen reiteradamente. No hay una regla determinada que indique el número máximo de repeticiones, pero es importante comprobar que se usan un número razonable de veces. Luego de realizar la imputación se calculó la cantidad de veces que se repitió cada donante, lo que se muestra en el cuadro 13.

**Cuadro 13. Cantidad de observaciones según el número de veces que son utilizadas como donantes, por jurisdicción**

| Jurisdicción | Número de veces que se repite el donante | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| CABA | 260 | 20 | 1 | | |
| Partidos del GBA | 597 | 106 | 16 | 5 | 1 |
| Resto de Buenos Aires | 290 | 31 | 4 | | |
| Catamarca | 84 | 8 | 1 | | |
| Córdoba | 309 | 57 | 8 | 3 | 1 |
| Corrientes | 193 | 43 | 6 | 1 | 1 |
| Chaco | 281 | 112 | 37 | 9 | 4 |
| Chubut | 89 | 12 | | | |
| Entre Ríos | 191 | 45 | 4 | 2 | 1 |
| Formosa | 89 | 19 | 4 | 1 | |
| Jujuy | 135 | 19 | 2 | | |
| La Pampa | 96 | 15 | 3 | 1 | |
| La Rioja | 162 | 25 | 4 | 1 | |
| Mendoza | 211 | 18 | 2 | 1 | |
| Misiones | 145 | 32 | 7 | 1 | |
| Neuquén | 134 | 23 | 7 | 1 | |
| Río Negro | 129 | 15 | 2 | | |
| Salta | 346 | 57 | 9 | 3 | |
| San Juan | 207 | 66 | 16 | 4 | 1 |
| San Luis | 107 | 16 | 2 | | |
| Santa Cruz | 138 | 42 | 12 | 7 | 1 |
| Santa Fe | 236 | 46 | 3 | 1 | |
| Santiago del Estero | 226 | 53 | 9 | 3 | 1 |
| Tucumán | 284 | 35 | 5 | 3 | |
| Tierra del Fuego | 96 | 21 | 2 | 1 | 1 |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Por ejemplo, según se ve en el cuadro 6, en la CABA hay 260 observaciones que se utilizaron como donantes una sola vez; 20 que se utilizaron 2 veces y una observación que fue donante de tres receptores. Es decir, con 281 donantes se imputaron 301 receptores. Es importante destacar que ningún donante se repitió más de cinco veces en todo el proceso de imputación.

## 9. Imputación de ingreso

Uno de los objetivos de la ENGHo es estimar el ingreso medio neto de un perceptor, quien puede obtener sus ingresos de una o varias fuentes. Los ingresos se registran según la fuente que los genera:

- Ocupación principal (que abarca las subfuentes Cuenta propia, Patrón y Asalariado)
- Ocupación secundaria (con las subfuentes Cuenta propia o Patrón y Asalariado)
- Jubilación
- Ocupación anterior
- Otras fuentes corrientes (Rentas, Transferencias, Autoconsumo)

Es posible que la información falte o esté incompleta, lo que lleva a considerar la imputación de alguna de las fuentes declaradas por los miembros del hogar. A tal efecto, se definió como no respuesta la situación en que un perceptor declaró en el C1 que tenía ingresos de una fuente, pero no los detalló en el C5. La imputación se realizó por perceptor y se consideró cada fuente por separado, con independencia de la completitud en las restantes fuentes.

La metodología aplicada para la imputación de ingresos sigue los mismos lineamientos generales empleados para los gastos (ver diagrama del apartado 4); o sea, Hot Deck con celdas de imputación

construidas a través de árboles de regresión y selección del donante a través del método *k*-vecinos más cercanos o por el promedio, como se muestra en el cuadro 14.

**Cuadro 14. Método de imputación por cada fuente de ingreso**

| Fuente | | Imputación |
|---|---|---|
| Ocupación principal | Asalariado | Hot Deck con árbol de regresión por jurisdicción y VMC |
| | Cuenta propia | Hot Deck con árbol de regresión por jurisdicción y VMC ([1]) |
| | Patrón | Hot Deck con árbol de regresión por región y VMC |
| Ocupación secundaria | | Hot Deck con árbol de regresión por región y VMC |
| Jubilación | | Hot Deck con árbol de regresión por jurisdicción y VMC |
| Ocupación anterior | | Ingreso medio por región |
| Otras fuentes corrientes no derivadas de la actividad económica | | Ingreso medio por región/subregión ([2]) |

([1]) En las provincias de Chaco, Mendoza, Neuquén, Santa Cruz y Tierra del Fuego los árboles de regresión fueron construidos por región, en virtud de la falta de donantes en cada una de las mencionadas.
([2]) El detalle de la modalidad de imputación para las fuentes comprendidas se muestra en el cuadro 18.
**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

## 9.1 Ingreso por Ocupación principal, Jubilaciones y pensiones, y Ocupación secundaria

En los siguientes cuadros se presenta la tasa de no respuesta para las principales fuentes. La desagregación geográfica que presentan los cuadros se relaciona con la magnitud de la no respuesta. El ingreso correspondiente a Asalariado y Jubilación (cuadro 15) se presenta por jurisdicción, mientras que el de Cuenta propia, Patrón y Ocupación secundaria (cuadro 16), por región.

**Cuadro 15. Tasa de imputación de las fuentes de ingreso correspondientes a Asalariados y Jubilación, por jurisdicción**

| Jurisdicción | Tasa de imputación | |
|---|---|---|
| | Asalariado | Jubilación |
| | % | |
| CABA | 17,5 | 10,4 |
| Partidos del GBA | 25,9 | 13,4 |
| Resto de Buenos Aires | 9,7 | 7,2 |
| Catamarca | 8,8 | 8,3 |
| Córdoba | 6,0 | 1,9 |
| Corrientes | 13,7 | 11,8 |
| Chaco | 34,2 | 32,1 |
| Chubut | 5,4 | 3,8 |
| Entre Ríos | 10,0 | 6,9 |
| Formosa | 4,8 | 4,2 |
| Jujuy | 5,4 | 5,3 |
| La Pampa | 9,6 | 10,4 |
| La Rioja | 13,2 | 11,9 |
| Mendoza | 32,5 | 19,5 |
| Misiones | 24,7 | 12,0 |
| Neuquén | 22,9 | 13,8 |
| Río Negro | 10,5 | 6,9 |
| Salta | 6,3 | 9,1 |
| San Juan | 11,8 | 8,1 |
| San Luis | 5,0 | 5,7 |
| Santa Cruz | 38,1 | 29,3 |
| Santa Fe | 9,6 | 6,5 |
| Santiago del Estero | 14,4 | 9,4 |
| Tucumán | 5,3 | 7,7 |
| Tierra del Fuego | 7,3 | 9,2 |

**Fuente**: INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

**Cuadro 16. Tasa de imputación de las fuentes de ingreso correspondientes a Cuenta propia, Patrón y Ocupación secundaria, por jurisdicción**

| Región | Tasa de imputación | | |
|---|---|---|---|
| | Cuenta propia | Patrón | Ocupación secundaria |
| | % | | |
| CABA | 23,0 | 40,0 | 15,0 |
| Partidos del GBA | 30,5 | 45,1 | 28,1 |
| Pampeana ([1]) | 21,1 | 22,6 | 13,3 |
| NOA | 15,4 | 20,9 | 10,5 |
| NEA | 32,7 | 34,5 | 11,5 |
| Cuyo | 27,5 | 40,6 | 24,0 |
| Patagonia | 24,1 | 30,3 | 17,5 |

([1]) Incluye el resto de partidos de la provincia de Buenos Aires que no conforman el GBA.
**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Los árboles se construyeron modelando el logaritmo del ingreso por subfuente (Asalariados, Cuenta propia y Patrón) para Ocupación principal; para Jubilaciones y pensiones en conjunto y para Ocupación secundaria. Se realizaron a nivel de jurisdicción cuando la cantidad de observaciones y la tasa de no respuesta lo permitió; en los casos donde no se pudo, los árboles se construyeron por región.

Al igual que en la imputación de gastos personales con no respuesta total, para la construcción de los árboles se utilizó el procedimiento HPSPLIT, con un mismo número mínimo de observaciones por nodo.

Fue necesario poner un límite al número de nodos finales de cada árbol (entre 5 y 20, aproximadamente) para asegurar donantes suficientes en las celdas, teniendo en cuenta las tasas de imputación. Para esto se utilizó la regla de "Breiman 1-SE" (SAS, 2015) que amplía el rango de subárboles a ser elegidos.

La imputación de ingreso se realizó con los datos consistidos, los valores de ingreso que se consideraron erróneos fueron consideraros a imputar. Aun así, se realizó un análisis de los ingresos que se corresponden a valores reales, es decir, que pasaron las reglas de consistencia.

Para esto se analizó cada uno de los valores que superaban el percentil 98. Para evaluar si alguno de estos valores tenía un comportamiento atípico se realizó un estudio multivariado, se construyeron clases de detección de valores extremos con un árbol de regresión creado con las mismas variables que el de imputación. Luego, en cada clase se consideró un ingreso como extremo severo si este se encontraba a más de 3 veces el rango intercuantil sobre el tercer cuartil de la clase. A los valores así identificados no se los utilizó como donantes.

Las variables auxiliares propuestas para la construcción de los árboles de regresión fueron aquellas con correlaciones más altas con el logaritmo de la variable ingreso, aplicando el "principio de parsimonia" ya señalado. En el cuadro 17 se presentan las variables correspondientes a Asalariados (Ocupación principal).[6]

**Cuadro 17. Variables auxiliares propuestas para los árboles y utilizadas en la distancia de vecino más cercano, para la fuente de ingreso Ocupación principal-asalariado**

| Variables propuestas para los árboles ([1]) | Variables VMC |
|---|---|
| Aglomerado principal (2) | Cantidad de horas trabajadas |
| Asalariado registrado (3) | Semana de relevamiento en campo (55) |
| Beneficiario de programa social (2) | |
| Calificación ocupacional (4) | |
| Cantidad de fuentes de ingreso del perceptor | |
| Cantidad de horas trabajadas | |
| Cantidad de miembros del hogar | |
| Cantidad de perceptores de ingreso del hogar | |
| Cobertura médica (3) | |
| Come afuera del hogar (2) | |
| Edad (10) | |
| Jerarquía ocupacional (3) | |
| Logaritmo del gasto del hogar por perceptor | |
| Nivel educativo (4) | |
| Número de autos en el hogar (3) | |
| Rama de actividad ([2]) | |
| Sexo (2) | |
| Tamaño del establecimiento donde trabaja (6) | |
| Tipo de empleo (3) | |
| Trimestre de la encuesta (4) | |

([1]) Entre paréntesis se indica la cantidad de categorías, cuando corresponda.
([2]) Se crearon 21 variables *dummy* que corresponden a las letras de la Clasificación de Actividades Económicas para Encuestas Sociodemográficas (CAES).
**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Es importante destacar que para la imputación del ingreso se recurrió a las variables de gasto ya imputadas, por los procedimientos apuntados en el apartado 6.

Como las variables seleccionadas para calcular la distancia del VMC son continuas, se emplea la distancia euclídea. Para que ambas variables tengan la misma importancia, a cada una se la estandariza a través del promedio y el desvío dentro de la celda.

[6] En la publicación se presentan los cuadros correspondientes a esta fuente, ya que representa la mayor proporción de los ingresos en los miembros de los hogares.

En Anexo II. Variables que jerarquizan los árboles de regresión para asalariados se presentan las variables con las que se conformaron las celdas de imputación en cada jurisdicción. Están señaladas las variables que definen el primer nivel de corte de cada árbol, las correspondientes al segundo nivel de corte y a los niveles posteriores en su conjunto.

### 9.2 Ingreso de ocupaciones anteriores y otras fuentes

Estas fuentes incluyen los ingresos corrientes derivados de una ocupación anterior a la semana de referencia y los ingresos corrientes de aquellas fuentes que no derivan directamente del desarrollo de una actividad económica. Es decir, ingresos por jubilación o pensión no contributivas, rentas de la propiedad y del capital, otros ingresos corrientes como contribución por separación o cuota alimentaria, ayuda familiar permanente en dinero de otros hogares residentes en el país o en el exterior, Asignación Universal por Hijo (AUH), Asignación Universal por Embarazo (AUE), seguros de desempleo, becas de estudio de todos los niveles educativos, transferencias para la adquisición de bienes o servicios para el hogar como atención de enfermos o alimentos y otras formas de asistencia social.

En todos estos casos, como se señala en cuadro 18, se utilizó el ingreso medio de cada fuente como procedimiento de imputación, con la excepción de la AUH donde se utilizaron registros administrativos para asignar montos en los casos a imputar. Los promedios fueron calculados sobre los donantes en clases de imputación definidas por las variables región, subregión o provincia, según la fuente. De este modo, cada valor faltante en una fuente se imputó con el promedio de la clase a la cual pertenece el valor a ser imputado.

**Cuadro 18. Resumen del método de imputación en otras fuentes**

| Fuente | Clases de imputación | Imputación |
|---|---|---|
| AUH | Patagonia/resto | Registro |
| AUE | Patagonia/resto | Ingreso medio |
| Plan Progresar | Subregión | Ingreso medio |
| Ayuda familiar | Subregión | Ingreso medio |
| Rentas (propiedad/capital) | Región | Ingreso medio |
| Transferencias de privado | Región | Ingreso medio |
| Productos de autoconsumo | Región | Ingreso medio |
| Ocupación anterior | Región | Ingreso medio |
| Menores de 10 años | Región | Ingreso medio |
| Seguro de desempleo | Subregión | Ingreso medio |
| Cuota alimentos | Subregión | Ingreso medio |
| Subsidios de la seguridad | Región | Ingreso medio |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

## 10. Calidad de las imputaciones en los ingresos

Al igual que con la imputación de los gastos personales, para verificar que los valores imputados correspondientes al ingreso asalariado son razonables se realiza la comparación de distintas medidas calculadas antes y después de la imputación.

Primero se evalúa la distribución de la variable imputada en forma univariada a través de la media de los ingresos. A continuación, la relación bivariada entre el ingreso y distintas variables por medio de la correlación. Luego se calcula el coeficiente de determinación $R^2$ que surge de ajustar un modelo lineal

entre el ingreso y otras variables relevantes, para evaluar una dimensión multivariada. Por último, se muestra la cantidad de veces que se usa una observación como donante.

## 10.1 Efectos de la imputación del ingreso asalariado

El impacto de la imputación en la distribución del ingreso asalariado se evalúa a través de la estimación del ingreso medio por jurisdicción. Todas las medidas se consideran ponderadas.

En el cuadro 19 se presentan el ingreso promedio estimado antes y después de la imputación, la diferencia entre ellos y la diferencia relativa al ingreso promedio observado, expresada en porcentaje.

**Cuadro 19. Estimación del ingreso asalariado medio antes y después de la imputación, por jurisdicción**

| Jurisdicción | Ingreso asalariado | | Diferencia | |
|---|---|---|---|---|
| | Antes | Después | absoluta | relativa |
| | Pesos | | | % |
| CABA | 25.454 | 25.434 | 20 | 0,1 |
| Partidos del GBA | 18.083 | 17.859 | 224 | 1,2 |
| Resto de Buenos Aires | 16.636 | 16.785 | -149 | -0,9 |
| Catamarca | 12.846 | 12.760 | 87 | 0,7 |
| Córdoba | 14.059 | 14.217 | -158 | -1,1 |
| Corrientes | 11.725 | 12.074 | -349 | -3,0 |
| Chaco | 10.337 | 11.105 | -768 | -7,4 |
| Chubut | 26.722 | 26.754 | -32 | -0,1 |
| Entre Ríos | 13.613 | 13.818 | -205 | -1,5 |
| Formosa | 10.883 | 10.984 | -101 | -0,9 |
| Jujuy | 13.261 | 13.230 | 31 | 0,2 |
| La Pampa | 17.343 | 17.246 | 97 | 0,6 |
| La Rioja | 11.376 | 11.621 | -246 | -2,2 |
| Mendoza | 15.780 | 15.717 | 62 | 0,4 |
| Misiones | 12.171 | 12.560 | -389 | -3,2 |
| Neuquén | 21.545 | 20.785 | 760 | 3,5 |
| Río Negro | 20.152 | 19.809 | 343 | 1,7 |
| Salta | 12.879 | 12.885 | -6 | 0,0 |
| San Juan | 12.104 | 11.811 | 293 | 2,4 |
| San Luis | 13.926 | 14.202 | -276 | -2,0 |
| Santa Cruz | 22.541 | 22.408 | 133 | 0,6 |
| Santa Fe | 17.206 | 16.803 | 404 | 2,3 |
| Santiago del Estero | 10.382 | 10.470 | -88 | -0,8 |
| Tucumán | 13.577 | 13.463 | 114 | 0,8 |
| Tierra del Fuego | 31.815 | 31.922 | -107 | -0,3 |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Como se puede advertir, no existen grandes diferencias entre ambos promedios con la excepción de Chaco, tal como se observó en párrafos anteriores.

A continuación, para evaluar una dimensión bivariada del efecto de la imputación del ingreso asalariado, en el cuadro 20 se presentan los cálculos de la correlación del ingreso antes y después de la imputación, con dos variables continuas elegidas a modo ilustrativo: la edad y el gasto del hogar. El interés de este análisis radica en ver si las correlaciones cambian como resultado de la imputación, no siendo relevantes las medidas en sí mismas.

**Cuadro 20. Correlación del ingreso asalariado con la edad y con el gasto del hogar antes y después de la imputación, por jurisdicción**

| Jurisdicción | Correlación del ingreso asalariado | | | |
|---|---|---|---|---|
| | Edad | | Gasto del hogar | |
| | Antes | Después | Antes | Después |
| CABA | 0,19 | 0,19 | 0,49 | 0,47 |
| Partidos del GBA | 0,22 | 0,18 | 0,40 | 0,40 |
| Resto de Buenos Aires | 0,16 | 0,19 | 0,29 | 0,32 |
| Catamarca | 0,35 | 0,34 | 0,40 | 0,39 |
| Córdoba | 0,19 | 0,20 | 0,45 | 0,44 |
| Corrientes | 0,12 | 0,14 | 0,27 | 0,29 |
| Chaco | 0,32 | 0,22 | 0,23 | 0,18 |
| Chubut | 0,22 | 0,24 | 0,35 | 0,35 |
| Entre Ríos | 0,27 | 0,28 | 0,31 | 0,31 |
| Formosa | 0,24 | 0,25 | 0,40 | 0,41 |
| Jujuy | 0,37 | 0,37 | 0,43 | 0,43 |
| La Pampa | 0,30 | 0,29 | 0,43 | 0,43 |
| La Rioja | 0,28 | 0,28 | 0,35 | 0,34 |
| Mendoza | 0,14 | 0,17 | 0,49 | 0,46 |
| Misiones | 0,20 | 0,16 | 0,43 | 0,38 |
| Neuquén | 0,19 | 0,18 | 0,39 | 0,42 |
| Río Negro | 0,24 | 0,24 | 0,62 | 0,61 |
| Salta | 0,25 | 0,23 | 0,44 | 0,44 |
| San Juan | 0,23 | 0,23 | 0,42 | 0,40 |
| San Luis | 0,26 | 0,27 | 0,31 | 0,31 |
| Santa Cruz | 0,12 | 0,05 | 0,20 | 0,22 |
| Santa Fe | 0,18 | 0,18 | 0,47 | 0,47 |
| Santiago del Estero | 0,36 | 0,34 | 0,35 | 0,33 |
| Tucumán | 0,30 | 0,30 | 0,42 | 0,42 |
| Tierra del Fuego | 0,18 | 0,18 | 0,21 | 0,21 |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Se observa que, en general, la imputación no provoca cambios en las correlaciones. Los cambios más importantes se produjeron con la variable edad en Chaco y Santa Cruz, provincias afectadas por la alta tasa de valores faltantes o con no respuesta.

Por último, para considerar una dimensión multivariada del ingreso, se calculó el coeficiente de determinación de la regresión del ingreso con las variables explicativas efectivamente utilizadas en la construcción de los árboles, tal como se presenta en el cuadro 21.

**Cuadro 21. R cuadrado de la regresión del ingreso asalariado con las variables del árbol antes y después de la imputación, por jurisdicción**

| Jurisdicción | $R^2$ | |
|---|---|---|
| | Antes | Después |
| CABA | 0,62 | 0,59 |
| Partidos del GBA | 0,58 | 0,55 |
| Resto de Buenos Aires | 0,53 | 0,53 |
| Catamarca | 0,59 | 0,56 |
| Córdoba | 0,60 | 0,58 |
| Corrientes | 0,60 | 0,55 |
| Chaco | 0,61 | 0,54 |
| Chubut | 0,58 | 0,56 |
| Entre Ríos | 0,57 | 0,57 |
| Formosa | 0,61 | 0,61 |
| Jujuy | 0,55 | 0,55 |
| La Pampa | 0,66 | 0,62 |
| La Rioja | 0,65 | 0,62 |
| Mendoza | 0,60 | 0,55 |
| Misiones | 0,56 | 0,56 |
| Neuquén | 0,65 | 0,60 |
| Río Negro | 0,66 | 0,64 |
| Salta | 0,58 | 0,57 |
| San Juan | 0,59 | 0,57 |
| San Luis | 0,61 | 0,59 |
| Santa Cruz | 0,52 | 0,36 |
| Santa Fe | 0,59 | 0,57 |
| Santiago del Estero | 0,60 | 0,56 |
| Tucumán | 0,62 | 0,62 |
| Tierra del Fuego | 0,57 | 0,54 |

**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Se aprecia que el valor de $R^2$ es similar en todas las jurisdicciones antes y después de la imputación, es decir, se mantiene la relación entre el ingreso y el conjunto de variables consideradas. La diferencia más extrema se produce en Santa Cruz, donde disminuye notoriamente el coeficiente como consecuencia a la alta tasa de valores faltantes que afecta a esta provincia.

## 10.2 Repetición de donantes

Al igual que en la imputación de gasto, se calculó el número de veces que se repite cada donante, lo que se muestra en el cuadro 22.

**Cuadro 22. Cantidad de observaciones según el número de veces que son utilizadas como donantes, por jurisdicción**

| Jurisdicción | Número de veces que se repite el donante | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | más de 4 ([1]) |
| CABA | 221 | 42 | 3 | 1 |
| Partidos del GBA | 341 | 71 | 9 | 2 |
| Resto de Buenos Aires | 124 | 8 | | |
| Catamarca | 66 | 8 | | |
| Córdoba | 64 | 1 | | |
| Corrientes | 64 | 9 | | |
| Chaco | 104 | 27 | 7 | 1 |
| Chubut | 34 | 2 | | |
| Entre Ríos | 53 | 5 | | |
| Formosa | 38 | 2 | | |
| Jujuy | 39 | | | |
| La Pampa | 39 | 2 | | |
| La Rioja | 90 | 16 | | |
| Mendoza | 168 | 41 | 8 | 2 |
| Misiones | 82 | 10 | 3 | 1 |
| Neuquén | 70 | 11 | 1 | |
| Río Negro | 79 | 4 | | 1 |
| Salta | 62 | 1 | | |
| San Juan | 76 | 5 | | 1 |
| San Luis | 21 | 1 | | |
| Santa Cruz | 91 | 23 | 5 | 2 |
| Santa Fe | 71 | 10 | 2 | |
| Santiago del Estero | 75 | 11 | 1 | |
| Tucumán | 42 | 3 | | |
| Tierra del Fuego | 29 | | | |

([1]) Las observaciones indicadas en esta columna se utilizaron como donantes 4 veces, salvo una observación en Santa Cruz que se utilizó 6 veces
**Fuente:** INDEC, Encuesta Nacional de Gastos de los Hogares 2017-2018.

Por ejemplo, según se ve en el cuadro 22, en la CABA hay 221 observaciones que se utilizaron como donantes una sola vez; 42 que se utilizaron 2 veces; 3 que se repitieron 3 veces y, por último, una que fue donante de cuatro receptores. Es decir, con 267 donantes se imputaron 318 receptores.

## 11. Síntesis y recomendaciones para el uso de los datos con imputaciones

La mayoría de las operaciones estadísticas que emplean la recolección de datos por muestreo, censos o registros se enfrentan al problema de datos faltantes en todas o en algunas de las variables de interés.

En la actualidad, para la etapa de difusión de los resultados, los institutos nacionales de estadística ponen a disposición de los usuarios una base de microdatos. En ella se incorporan las principales variables de la encuesta y toda información adicional (factores de expansión, variables indicadoras de dominios de estimación, etc.), sin violar las restricciones del secreto estadístico, para que los usuarios puedan efectuar sus propias estimaciones.

Esto obliga, por lo general, a aplicar algún tratamiento de los datos faltantes para completarla, sumar indicadores de dato faltante o perdido para las principales variables de la encuesta en la base, y transparentar la metodología de imputación aplicada en la encuesta.

Retomando lo señalado en el apartado 2, la presencia de datos faltantes es indeseable porque lleva a: a) que los estimadores que se emplean para las estimaciones sean vulnerables al sesgo por no respuesta y, b) sufrir un aumento del error muestral si se emplea solo la información de los que responden, como consecuencia de la reducción de la muestra.

En el caso la ENGHo, esta nota técnica sigue en línea lo expresado en los párrafos anteriores. Para paliar las dificultades señaladas en a) y b) y presentar una base completa de una encuesta, se empleó un método de imputación bajo algunos supuestos. Como se señala en el apartado, de acuerdo con la práctica habitual para la mayor parte de los métodos existentes, el principal supuesto que habilita a emplearlo es el de no respuesta al azar o MAR.

La metodología se basa en imputar un dato faltante a través de seleccionar un donante al azar por Hot Deck, entre los vecinos más cercanos. Las celdas para la selección son construidas con la ayuda de árboles de regresión, para ganar eficiencia en el procedimiento de imputación y buscar sostener lo mejor posible el supuesto MAR. En el proceso se controlaron los valores aberrantes en las variables y la repetición de donantes que pueden afectar al comportamiento del modelo de imputación y el de los estimadores de la encuesta.

Como se adelantó, uno de los principales objetivos de un proceso de imputación es tratar de disminuir en lo posible el sesgo por no respuesta parcial o en ítems en los estimadores que emplea la encuesta, y poner a disposición de los usuarios una base sin datos faltantes o completa para facilitar el cálculo de las estimaciones empleando el único factor de expansión de la encuesta.

Es de esperar que la metodología aplicada en la mayor parte de las estimaciones para gastos e ingresos haya logrado el primer objetivo, pero el usuario debe saber que no se puede evaluar o cuantificar la reducción del sesgo en forma directa a partir de los datos de la encuesta. La naturaleza propia del fenómeno, en el que solo se tiene información de las variables de interés en el conjunto de unidades observadas, y no en las unidades inobservables o sin respuesta, lo impide.

Por lo tanto, es importante que, a la hora de emplear los datos de la ENGHo, los usuarios tengan presentes las siguientes consideraciones y recomendaciones:

- La imputación es un recurso estadístico que intenta disminuir el sesgo por no respuesta sobre un estimador, empleado para una estimación o agregado de una variable de interés; pero que no lo elimina.

- En ningún caso los valores imputados se pueden tomar como verdaderos, ya que surgen de un modelo de imputación que trata de predecir la componente no observada por falta de respuesta o dato perdido; solo deben ser tomados como valores de referencia para ser tratados en forma agregada y adoptando los recaudos señalados en el presente apartado.

- No todos los estimadores se ven afectados por el sesgo de no respuesta de la misma manera; algunos sufren un impacto mayor que otros. Por ejemplo, es probable que los empleados en un estudio para tasas o promedios tengan menos sesgo que los utilizados para parámetros de orden (quintil, decil, etc.), índice de Gini u otra medida de desigualdad.

- La tasa de no respuesta de la encuesta afecta a las estimaciones; no solo puede deteriorar la precisión de los estimadores, a tasas más altas los modelos de imputación pueden perder su capacidad predictiva. Se deberá poner atención en dominios de estimación o subpoblaciones en donde los niveles fueron altos (por ejemplo, más del 30% o 35%); o cuando se advierte que el número de unidades muestrales y con respuesta son pocas (por ejemplo,

menos de 100 observaciones); o cuando el total de datos imputados supera del total de observaciones en el dominio (por ejemplo, en un 40% o más).

- El indicador de la tasa de no respuesta no es la única guía para dimensionar el problema del sesgo; este no tiene en cuenta lo que aportan en el tratamiento del sesgo las variables auxiliares incorporadas al modelo de imputación. A tal efecto, una interpretación en conjunto con los indicadores de la calidad de las imputaciones, presentados en esta nota técnica, pueden orientar y ayudar al usuario a validar sus estimaciones.

- La metodología de imputación aplicada a la encuesta habilita a los usuarios a seguir empleando los estimadores y factores de expansión de la encuesta provistos por el Instituto; de esta manera, por ejemplo, dos usuarios que llevan un mismo análisis sobre el mismo conjunto de datos deberían alcanzar las mismas conclusiones.

- Como en la metodología para el cálculo del error de muestreo no se tuvieron en cuenta los métodos de imputación empleados, estos errores pueden verse ligeramente subestimados si la estimación se obtiene en dominios con altas tasas de no respuesta. En ese caso, se recomienda calcular estimaciones en los dominios geográficos o agregados definidos como dominios de estimación por el diseño muestral.

Finalmente, se sugiere la lectura de la nota técnica n° 4, *Encuesta Nacional de Gastos de los Hogares 2017-2018. Factores de expansión, estimación y cálculo de los errores de muestreo.*[7] Esta permite componer un diagnóstico integral de los procesos de expansión y los errores de muestreo e imputación aplicados en la encuesta, y sumar las recomendaciones que se señalan en ella.

[7] Ver *Encuesta Nacional de Gastos de los Hogares 2017-2018. Factores de expansión, estimación y cálculo de los errores de muestreo.* Nota técnica n°4. Disponible en: https://www.indec.gob.ar/ftp/cuadros/menusuperior/engho/engho2017_18_nota_tecnica_4.pdf.

## Referencias


Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, *The American Statistician*, 46(3), 175-185, DOI: 10.1080/00031305.1992.10475879.

Andridge, R. y Little, R. J. A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*, 78(1), 40-64, DOI: 10.1111/j.1751-5823.2010.00103.x.

Creel, D. y Krotki, K. (2006). Creating imputation classes using classification tree methodolgy. *Proceedings of the Survey Research Methods Section (ASA)*, 2884-2887. Recuperado de; https://www.researchgate.net/publication/251906313_Creating_Imputation_Classes_Using_Classification_Tree_Methodology.

Chen, S. y Haziza, D. (2019). Recent Developments in Dealing with Item Non-response in Survey: A Critical Review. *International Statistical Review*, 87(1), 192-218. DOI:10.1111/insr.12305.

De Waal, T., Pannekoek, J. y Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.

Grande, E. y Luzi, O. (2004). *Regression trees in the context of imputation of item non-response: an experimental application on business data*. Istat, Servizio Metodologie, Tecnologie e Software per la Produzione Statistica. Recuperado de: https://www.istat.it/it/files/2018/07/2004_11.pdf.

Kass, J. V (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29(2), 119-127. DOI:10.2307/2986296.

Loh, W. (2008). Classification and Regression Tree Methods. En F. Ruggeri, R. S. Kenett y F. W. Faltin (Eds.), *Encyclopedia of Statistics in Quality and Reliability*. DOI:10.1002/9780470061572.eqr492.

Loh, W. Y., Eltinge, J., Cho, M. y Li, Y. (2016). Classification and regression tree methods for incomplete data from sample surveys. Recuperado de: https://arxiv.org/abs/1603.01631.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592. DOI:10.2307/2335739.

Saar-Tsechansky, M. y Provost, F. (2007). Handling Missing Values when Applying Classification Models. *Journal of Machine Learning Research*, 8, 1625-1657. Recuperado de: http://jmlr.csail.mit.edu/papers/volume8/saar-tsechansky07a/saar-tsechansky07a.pdf.

SAS Institute Inc. (2015). SAS/STAT® 14.1 User's Guide. Cary, NC: SAS Institute Inc.

Schafer, J. L. y Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7, 147-177. DOI:10.1037//1082-989X.7.2.147.

Seaman, S., Galati J., Jackson, D. y Carlin, J. (2013). What is meant by missing at random? *Stat. Sci.*, 28(2), 257-268. DOI:10.1214/13-STS415.

# Glosario

**Aleatorio**. Concepto que permite calificar un evento vinculado a un resultado posible entre otros y desconocido antes de ser ejecutado. Dentro del muestreo probabilístico es el propio proceso de selección el que asegura que la muestra resultante no pueda ser predicha de antemano. En ese contexto, las respuestas a las variables indagadas por la encuesta son tratadas como valores fijos, y la componente aleatoria es solo atribuida al proceso de selección que origina la muestra.

**Árboles de regresión.** Técnica estadística de aprendizaje supervisado no paramétrica que predice valores de una variable respuesta, mediante reglas de decisión binarias derivadas de variables auxiliares predictoras. Las reglas recursivamente dividen los datos en conjuntos disjuntos, donde el modelo predictivo es ajustado; las variables auxiliares pueden ser continuas o categóricas. Los árboles de regresión son una variante de los árboles de decisión y están diseñados para aproximar una función a valores reales, a diferencia de los árboles de clasificación, que predicen variables cualitativas.

**Clases de imputación.** Subgrupos de observaciones determinados por variables auxiliares medidas tanto en los que responden como en los que no. Se los emplea para que algunos métodos de imputación (por ejemplo, Hot Deck, promedios por clase, etc.) tengan mejor potencia predictiva y los supuestos que los sostienen sean admisibles. Es deseable que internamente sean homogéneas, en términos de la probabilidad de respuesta o en algún aspecto de la distribución de la variable a ser imputada.

**Dominios de estimación**. Subconjuntos de la población objetivo cuyos elementos pueden ser identificados en el marco muestral sin ambigüedad y a los que, en la etapa de diseño de la encuesta, se les determina un tamaño de muestra y un nivel de precisión predefinido para obtener estimaciones de interés en ellos. En una encuesta a hogares, suelen ser agregados geográficos, o agrupamientos geopolíticos o administrativos del territorio (región, provincia, aglomerado o localidad principal, etcétera).

**Donante.** Valor obtenido del conjunto de los datos observados para completar un dato faltante.

**Error de muestreo, error muestral o error por muestra**. Error asociado con la no inclusión de todos los miembros de la población en la muestra. Se refiere a la diferencia entre la estimación derivada de la muestra y el valor "verdadero" que resultaría si se realizara un censo de toda la población bajo las mismas condiciones en las que se llevó adelante la muestra.

**Estimación**. Proceso por el cual se obtiene un valor numérico o un rango de valores para un parámetro desconocido de la población a partir de los datos de una muestra. También empleado para denominar el resultado del proceso.

**Estimador**. Expresión analítica de una función que, utilizada con los datos de una muestra, permite estimar un parámetro de interés desconocido.

**Factor de expansión.** Valor asociado a cada unidad elegible y que responde a la muestra, que se construye a partir de la inversa de la probabilidad de inclusión de cada unidad o peso muestral inicial. Puede incluir distintos tipos de ajustes (por cobertura, por no respuesta, por calibración) que llevan en general a los estimadores a ganar eficiencia y precisión.

**Hot Deck**. Familia de métodos de imputación no determinísticos, que emplea donantes seleccionados de los datos observados para completar datos faltantes. Por lo general, la selección es aleatoria y se lleva a cabo dentro de clases de imputación.

**Imputación**. Proceso empleado para completar un dato faltante, inválido o inconsistente que no fue aceptado en la etapa de edición o consistencia de los datos de una encuesta.

**Imputación aleatoria**. Proceso por el cual se completa un dato faltante con un valor surgido de un proceso de imputación que cada vez que se lo replica origina un valor distinto. La mayoría de los métodos Hot Deck son de imputación aleatoria.

**Imputación determinística**. Proceso por el cual se completa un valor faltante a través de una estimación o un valor que se caracteriza por no cambiar cada vez que se repite el proceso de imputación. Por ejemplo, imputar por la media o la mediana es una imputación determinística; si al valor surgido de un método determinístico se le suma una componente aleatoria, se transforma en una imputación aleatoria.

**Indicador de (no) respuesta**. Variable binaria que representa la falta de información o dato en una variable. Toma valor 1 cuando la unidad responde; 0, en caso contrario. Desde el punto de vista estadístico, es una variable aleatoria que acompaña a cada variable de interés, cuya distribución es desconocida. Está involucrada en el proceso estocástico que guía el mecanismo de no respuesta que origina los datos faltantes en las variables de la encuesta.

***k*-vecinos más cercanos.** Técnica empleada para determinar las *k* unidades más próximas a una unidad dada, según una métrica o distancia, y definida a partir de un conjunto de variables auxiliares medidas para todas las observaciones de la encuesta o estudio.

**Mecanismo de respuesta**. Mecanismo subyacente responsable o causal de los datos faltantes o perdidos. Este mecanismo es desconocido; pero, desde el punto de vista estadístico, se lo explicita mediante un conjunto de postulados probabilísticos. Estos describen la interrelación entre los indicadores de respuesta de las variables estudiadas en la encuesta, con las componentes observadas y no observadas de dichas variables. Dependiendo de su naturaleza, habilita a sostener el tipo de inferencia que se adopta para el encuadre metodológico del método de imputación.

**Mecanismo ignorable**. Mecanismos de no respuesta en donde los efectos sobre el sesgo y la precisión en un estimador, y la validación y potencia predictiva de la imputación, pueden ser mitigados sin tener que modelar explícitamente el mecanismo. Tanto los MAR como los MCAR son ignorables. (Ver **No respuesta al azar o MAR** y **No respuesta completamente al azar o MCAR**).

**Método de imputación**. Conjunto de supuestos, técnicas, reglas y tratamientos de los valores observados o no, que permiten completar un conjunto de datos con información faltante.

**No respuesta**. Imposibilidad de obtener datos sobre las unidades elegibles de la población objetivo, en un censo o una encuesta. Puede ser total, o sea, cuando para la unidad no se logra la información requerida por el cuestionario; o parcial, cuando solo para algunos de los ítems incluidos en el cuestionario se falla en obtener información.

**No respuesta al azar o MAR**. Supuesto sobre el mecanismo de no respuesta que postula que la probabilidad de respuesta es afectada por la información de la componente observada y no por aquella no observada en el estudio o encuesta. Por ejemplo, un respondente anciano puede terner dificultades para recordar un evento y no brindar información a causa de problemas de memoria. La no respuesta está relacionada a la edad, pero no al evento en sí mismo, en cuyo caso es un dato faltante por no respuesta al azar o MAR. Este supuesto sobre el mecanismo es clave, porque induce a que efectivamente los datos observados son suficientes para romper la dependencia entre la probabilidad de respuesta a una variable y el valor particular que esta variable toma. También es conocido como *missing at random* (MAR) por su terminología en inglés.

**No respuesta completamente al azar o MCAR**. Caso especial de MAR. Se caracteriza por postular la independencia de la probabilidad de respuesta con respecto a las componentes observadas o no de todas las variables del estudio o encuesta. Por ejemplo, la falta de respuesta en un ítem de alguna variable a causa de un salto a una pregunta del cuestionario no advertido por el encuestador es un dato faltante que se encuadra en este supuesto. Bajo este supuesto, al conjunto de los que no responden se los puede tomar como si provinieran de una muestra puramente al azar del conjunto completo de los datos. Suele ser mencionado en la bibliografía como *missing completely at random* (MCAR).

**No respuesta no al azar o MNAR**. Supuesto sobre el mecanismo de no respuesta en el que la probabilidad de respuesta se ve afectada por la componente no observada del conjunto de los datos. Por ejemplo, un respondente que decide no responder a una pregunta sensible es un dato que sufrió la no respuesta no al azar. Este supuesto aparece en la bibliografía como *missing not at random* (MNAR).

**Parámetros**. Medidas cuantitativas de interés desconocidas de la población objetivo o de cualquier dominio de estimación específico que son factibles de ser estimadas a partir de una muestra. Algunos, usualmente considerados en las encuestas por muestreo, son del tipo descriptivo (como totales, medias, proporciones, varianzas, etcétera).

**Precisión**. Consistencia con la que se obtienen los resultados o mediciones a partir de la muestra aplicando el mismo diseño muestral con respecto al valor verdadero o parámetro poblacional de interés. (Ver **Error de muestreo**).

**Probabilidad**. Cuantificación de la posibilidad de ocurrencia de un evento aleatorio. Toma valores entre 0 y 1, y es el pilar fundamental en el que sostiene el proceso de inferencia estadística.

**Probabilidad de respuesta**. Probabilidad que manifiesta la tendencia a responder a un ítem o variable una unidad de observación.

**Sesgo**. Diferencia entre el valor esperado de un estimador y el valor del parámetro poblacional.

**Sesgo por no respuesta**. Sesgo que ocurre cuando el valor observado del estimador se desvía del parámetro poblacional debido a diferencias entre quienes responden la encuesta y los que no lo hacen. En el contexto de imputación, por lo general se lo define a nivel de respuesta parcial a una variable o para un ítem de una variable del cuestionario de una encuesta.

**Tasa de imputación**. Proporción de unidades de la muestra elegibles con no respuesta parcial a la encuesta o estudio, o cuya respuesta a un ítem fue anulada por información inconsistente o errónea en el proceso de edición; estas unidades son sometidas a los procedimientos de imputación para completar el ítem.

**Tasa de respuesta**. Proporción de unidades de la muestra elegibles que respondieron a la encuesta o estudio. Se puede calcular la tasa de respuesta total y parcial de acuerdo a la ocurrencia de respuesta total (todo el cuestionario) o parcial (ítems con no respuesta), respectivamente.

# Anexo I. Variables que jerarquizan los árboles de regresión para gasto

**Cuadro 23. Variables utilizadas en los árboles de regresión de gasto según nivel de corte, por jurisdicción**

| Jurisdicción | Primer nivel | Segundo nivel | Niveles posteriores |
|---|---|---|---|
| CABA | Condición de actividad (3) | Come afuera del hogar (2)<br>Cantidad de viajes en transporte público | Gasto de consumo en el hogar per cápita<br>Nivel de educación (4) |
| Partidos del GBA | Condición de actividad (3) | Come afuera del hogar (2)<br>Cantidad de viajes en transporte público | Gasto de consumo en el hogar per cápita<br>Edad (9)<br>Nivel de educación (4)<br>Número de autos en el hogar (3)<br>Sexo (2)<br>Estrato de área (5)<br>Cantidad de miembros del hogar<br>Perceptor de ingreso (2)<br>Jubilado o pensionado (2) |
| Resto de Buenos Aires | Come fuera del hogar (2) | Condición de actividad (3) | Cantidad de viajes en transporte público<br>Viaja en transporte público (2)<br>Gasto de consumo en el hogar per cápita<br>Edad (9)<br>Nivel de educación (4)<br>Sexo (2)<br>Estrato de área (5) |
| Catamarca | Condición de actividad (3) | Gasto de consumo en el hogar per cápita<br>Aglomerado principal (2) | Come afuera del hogar (2)<br>Cantidad de viajes en transporte público<br>Viaja en transporte público (2)<br>Edad (9)<br>Nivel de educación (4)<br>Estrato de área (5)<br>Cantidad de miembros del hogar<br>Asistencia a establecimiento educativo (2) |
| Córdoba | Come afuera del hogar (2) | Viaja en transporte público (2)<br>Edad (9) | Condición de actividad (3)<br>Cantidad de viajes en transporte público<br>Gasto de consumo en el hogar per cápita<br>Aglomerado principal (2)<br>Nivel de educación (4)<br>Número de autos en el hogar (3)<br>Sexo (2)<br>Estrato de área (5)<br>Cantidad de miembros del hogar<br>Asistencia a establecimiento educativo (2) |
| Corrientes | Come afuera del hogar (2) | Condición de actividad (3)<br>Edad (9) | Cantidad de viajes en transporte público<br>Viaja en transporte público (2)<br>Gasto de consumo en el hogar per cápita<br>Aglomerado principal (2)<br>Nivel de educación (4)<br>Sexo (2)<br>Estrato de área (5)<br>Cantidad de miembros del hogar<br>Perceptor de ingreso (2) |

(continúa)

**Cuadro 23. (continuación)**

| Jurisdicción | Primer nivel | Segundo nivel | Niveles posteriores |
|---|---|---|---|
| Chaco | Condición de actividad (3) | Cantidad de viajes en transporte público<br>Aglomerado principal (2) | Come afuera del hogar (2)<br>Gasto de consumo en el hogar per cápita<br>Edad (9)<br>Nivel de educación (4)<br>Número de autos en el hogar (3)<br>Sexo (2)<br>Estrato de área (5)<br>Cantidad de miembros del hogar |
| Chubut | Condición de actividad (3) | Viaja en transporte público (2)<br>Número de autos en el hogar (3) | Come afuera del hogar (2)<br>Cantidad de viajes en transporte público<br>Gasto de consumo en el hogar per cápita<br>Edad (9)<br>Aglomerado principal (2)<br>Nivel de educación (4)<br>Sexo (2)<br>Estrato de área (5)<br>Cantidad de miembros del hogar |
| Entre Ríos | Come afuera del hogar (2) | Condición de actividad (3)<br>Edad (9) | Cantidad de viajes en transporte público<br>Gasto de consumo en el hogar per cápita<br>Aglomerado principal (2)<br>Nivel de educación (4)<br>Número de autos en el hogar (3)<br>Sexo (2) |
| Formosa | Condición de actividad (3) | Come afuera del hogar (2)<br>Cantidad de viajes en transporte público | Gasto de consumo en el hogar per cápita<br>Edad (9)<br>Nivel de educación (4)<br>Número de autos en el hogar (3)<br>Sexo (2)<br>Cantidad de miembros del hogar<br>Perceptor de ingreso (2) |
| Jujuy | Condición de actividad (3) | Come afuera del hogar (2)<br>Cantidad de viajes en transporte público | Gasto de consumo en el hogar per cápita<br>Edad (9)<br>Aglomerado principal (2)<br>Nivel de educación (4)<br>Número de autos en el hogar (3)<br>Sexo (2)<br>Estrato de área (5)<br>Cantidad de miembros del hogar<br>Perceptor de ingreso (2) |
| La Pampa | Condición de actividad (3) | Come afuera del hogar (2)<br>Gasto de consumo en el hogar per cápita | Viaja en transporte público (2) |
| La Rioja | Condición de actividad (3) | Come afuera del hogar (2)<br>Viaja en transporte público (2) | Gasto de consumo en el hogar per cápita<br>Nivel de educación (4)<br>Número de autos en el hogar (3)<br>Sexo (2)<br>Cantidad de miembros del hogar |

(continúa)

**Cuadro 23. (continuación)**

| Jurisdicción | Primer nivel | Segundo nivel | Niveles posteriores |
|---|---|---|---|
| Mendoza | Condición de actividad (3) | Viaja en transporte público (2)<br>Gasto de consumo en el hogar per cápita | Come afuera del hogar (2)<br>Cantidad de viajes en transporte público<br>Nivel de educación (4)<br>Número de autos en el hogar (3)<br>Sexo (2) |
| Misiones | Condición de actividad (3) | Cantidad de viajes en transporte público<br>Gasto de consumo en el hogar per cápita | Viaja en transporte público (2)<br>Edad (9)<br>Nivel de educación (4)<br>Número de autos en el hogar (3)<br>Sexo (2)<br>Estrato de área (5) |
| Neuquén | Gasto de consumo en el hogar per cápita | Condición de actividad (3)<br>Cantidad de viajes en transporte público | Come afuera del hogar (2)<br>Edad (9)<br>Nivel de educación (4)<br>Número de autos en el hogar (3)<br>Sexo (2)<br>Estrato de área (5)<br>Cantidad de miembros del hogar |
| Río Negro | Gasto de consumo en el hogar per cápita | Condición de actividad (3)<br>Viaja en transporte público (2) | Come afuera del hogar (2)<br>Edad (9)<br>Nivel de educación (4)<br>Número de autos en el hogar (3)<br>Sexo (2)<br>Estrato de área (5) |
| Salta | Condición de actividad (3) | Come afuera del hogar (2)<br>Cantidad de viajes en transporte público | Gasto de consumo en el hogar per cápita<br>Edad (9)<br>Aglomerado principal (2)<br>Nivel de educación (4)<br>Sexo (2) |
| San Juan | Come afuera del hogar (2) | Cantidad de viajes en transporte público<br>Edad (9) | Condición de actividad (3)<br>Gasto de consumo en el hogar per cápita<br>Aglomerado principal (2)<br>Sexo (2)<br>Estrato de área (5)<br>Cantidad de miembros del hogar<br>Jubilado o pensionado (2) |
| San Luis | Condición de actividad (3) | Come afuera del hogar (2)<br>Cantidad de viajes en transporte público | Viaja en transporte público (2)<br>Número de autos en el hogar (3) |
| Santa Cruz | Condición de actividad (3) | Come afuera del hogar (2)<br>Viaja en transporte público (2) | Cantidad de viajes en transporte público<br>Gasto de consumo en el hogar per cápita<br>Edad (9)<br>Aglomerado principal (2)<br>Número de autos en el hogar (3)<br>Estrato de área (5)<br>Jubilado o pensionado (2) |

(continúa)

**Cuadro 23. (continuación)**

| Jurisdicción | Primer nivel | Segundo nivel | Niveles posteriores |
|---|---|---|---|
| Santa Fe | Condición de actividad (3) | Come afuera del hogar (2)<br>Viaja en transporte público (2) | Gasto de consumo en el hogar per cápita<br>Edad (9)<br>Aglomerado principal (2)<br>Nivel de educación (4)<br>Número de autos en el hogar (3)<br>Sexo (2)<br>Estrato de área (5)<br>Cantidad de miembros del hogar |
| Santiago del Estero | Condición de actividad (3) | Come afuera del hogar (2)<br>Viaja en transporte público (2) | Cantidad de viajes en transporte público<br>Gasto de consumo en el hogar per cápita<br>Edad (9)<br>Aglomerado principal (2)<br>Número de autos en el hogar (3)<br>Sexo (2)<br>Estrato de área (5)<br>Cantidad de miembros del hogar<br>Jubilado o pensionado (2) |
| Tucumán | Condición de actividad (3) | Come afuera del hogar (2)<br>Viaja en transporte público (2) | Cantidad de viajes en transporte público<br>Gasto de consumo en el hogar per cápita<br>Edad (9)<br>Nivel de educación (4)<br>Sexo (2)<br>Estrato de área (5) |
| Tierra del Fuego | Condición de actividad (3) | Come afuera del hogar (2)<br>Nivel de educación (4) | Cantidad de viajes en transporte público<br>Gasto de consumo en el hogar per cápita<br>Edad (9)<br>Sexo (2) |

## Anexo II. Variables que jerarquizan los árboles de regresión para asalariados

**Cuadro 24. Variables utilizadas en los árboles de regresión para asalariados según nivel de corte, por jurisdicción**

| Jurisdicción | Primer nivel | Segundo nivel | Niveles posteriores |
|---|---|---|---|
| CABA | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Logaritmo del gasto del hogar por perceptor | Calificación ocupacional (4)<br>Edad (10)<br>Tamaño del establecimiento donde trabaja (6)<br>Número de autos en el hogar (3)<br>Nivel educativo (4)<br>Jerarquía ocupacional (3) |
| Partidos del GBA | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Logaritmo del gasto del hogar por perceptor | Calificación ocupacional (4)<br>Edad (10)<br>Sexo (2)<br>Tamaño del establecimiento donde trabaja (6)<br>Cantidad de fuentes de ingreso del perceptor |
| Resto de Buenos Aires | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Rama de actividad T | Logaritmo del gasto del hogar por perceptor<br>Calificación ocupacional (4)<br>Edad (10)<br>Sexo (2)<br>Trimestre de la encuesta (4) |
| Catamarca | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Logaritmo del gasto del hogar por perceptor | Calificación ocupacional (4)<br>Trimestre de la encuesta (4)<br>Nivel educativo (4)<br>Aglomerado principal (2) |
| Córdoba | Cantidad de horas trabajadas | Asalariado registrado (3) | Logaritmo del gasto del hogar por perceptor<br>Calificación ocupacional (4)<br>Edad (10)<br>Sexo (2)<br>Número de autos en el hogar (3) |
| Corrientes | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Logaritmo del gasto del hogar por perceptor | Calificación ocupacional (4)<br>Trimestre de la encuesta (4)<br>Número de autos en el hogar (3)<br>Cantidad de fuentes de ingreso del perceptor |
| Chaco | Asalariado registrado (3) | Logaritmo del gasto del hogar por perceptor<br>Rama de actividad T | Cantidad de horas trabajadas |
| Chubut | Asalariado registrado (3) | Cantidad de horas trabajadas | Logaritmo del gasto del hogar por perceptor<br>Calificación ocupacional (4)<br>Edad (10)<br>Tamaño del establecimiento donde trabaja (6)<br>Rama de actividad B |
| Entre Ríos | Asalariado registrado (3) | Cantidad de horas trabajadas | Logaritmo del gasto del hogar por perceptor<br>Calificación ocupacional (4)<br>Edad (10)<br>Trimestre de la encuesta (4)<br>Rama de actividad P |
| Formosa | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Logaritmo del gasto del hogar por perceptor | Calificación ocupacional (4)<br>Trimestre de la encuesta (4)<br>Tamaño del establecimiento donde trabaja (6)<br>Cantidad de miembros del hogar<br>Aglomerado principal (2) |

(continúa)

**Cuadro 24. (continuación)**

| Jurisdicción | Primer nivel | Segundo nivel | Niveles posteriores |
|---|---|---|---|
| Jujuy | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Logaritmo del gasto del hogar por perceptor | Calificación ocupacional (4)<br>Edad (10)<br>Trimestre de la encuesta (4)<br>Nivel educativo (4)<br>Aglomerado principal (2) |
| La Pampa | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Logaritmo del gasto del hogar por perceptor | Tamaño del establecimiento donde trabaja (6)<br>Nivel educativo (4) |
| La Rioja | Asalariado registrado (3) | Cantidad de horas trabajadas | Logaritmo del gasto del hogar por perceptor<br>Calificación ocupacional (4)<br>Edad (10)<br>Trimestre de la encuesta (4)<br>Número de autos en el hogar (3)<br>Nivel educativo (4)<br>Cantidad de miembros del hogar<br>Rama de actividad G |
| Mendoza | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Logaritmo del gasto del hogar por perceptor | Calificación ocupacional (4)<br>Sexo (2)<br>Trimestre de la encuesta (4)<br>Número de autos en el hogar (3)<br>Nivel educativo (4)<br>Aglomerado principal (2)<br>Cantidad de perceptores de ingreso del hogar |
| Misiones | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Logaritmo del gasto del hogar por perceptor | Calificación ocupacional (4) |
| Neuquén | Asalariado registrado (3) | Cantidad de horas trabajadas | Logaritmo del gasto del hogar por perceptor<br>Edad (10)<br>Sexo (2)<br>Cantidad de miembros del hogar |
| Río Negro | Asalariado registrado (3) | Logaritmo del gasto del hogar por perceptor<br>Calificación ocupacional (4) | Cantidad de horas trabajadas<br>Sexo (2)<br>Trimestre de la encuesta (4)<br>Tamaño del establecimiento donde trabaja (6)<br>Cantidad de miembros del hogar<br>Aglomerado principal (2)<br>Tipo de empleo (3) |
| Salta | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Logaritmo del gasto del hogar por perceptor | Calificación ocupacional (4)<br>Trimestre de la encuesta (4)<br>Tamaño del establecimiento donde trabaja (6)<br>Número de autos en el hogar (3)<br>Aglomerado principal (2)<br>Tipo de empleo (3)<br>Cantidad de perceptores de ingreso del hogar |
| San Juan | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Logaritmo del gasto del hogar por perceptor | Edad (10)<br>Sexo (2)<br>Número de autos en el hogar (3)<br>Cantidad de miembros del hogar |

(continúa)

**Cuadro 24. (continuación)**

| Jurisdicción | Primer nivel | Segundo nivel | Niveles posteriores |
|---|---|---|---|
| San Luis | Asalariado registrado (3) | Logaritmo del gasto del hogar por perceptor<br>Edad (10) | Cantidad de horas trabajadas<br>Calificación ocupacional (4)<br>Trimestre de la encuesta (4)<br>Número de autos en el hogar (3)<br>Nivel educativo (4)<br>Tipo de empleo (3) |
| Santa Cruz | Cobertura médica (3) | Cantidad de horas trabajadas | Logaritmo del gasto del hogar por perceptor<br>Calificación ocupacional (4)<br>Número de autos en el hogar (3) |
| Santa Fe | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Logaritmo del gasto del hogar por perceptor | Edad (10)<br>Rama de actividad T<br>Trimestre de la encuesta (4)<br>Tamaño del establecimiento donde trabaja (6)<br>Rama de actividad O |
| Santiago del Estero | Asalariado registrado (3) | Logaritmo del gasto del hogar por perceptor<br>Sexo (2) | Cantidad de horas trabajadas<br>Calificación ocupacional (4)<br>Nivel educativo (4)<br>Cantidad de miembros del hogar<br>Tipo de empleo (3)<br>Cantidad de fuentes de ingreso del perceptor<br>Rama de actividad G |
| Tucumán | Asalariado registrado (3) | Cantidad de horas trabajadas<br>Logaritmo del gasto del hogar por perceptor | Calificación ocupacional (4)<br>Edad (10)<br>Sexo (2)<br>Trimestre de la encuesta (4)<br>Cantidad de miembros del hogar<br>Rama de actividad A |
| Tierra del Fuego | Asalariado registrado (3) | Cantidad de horas trabajadas | Logaritmo del gasto del hogar por perceptor<br>Edad (10)<br>Tamaño del establecimiento donde trabaja (6) |