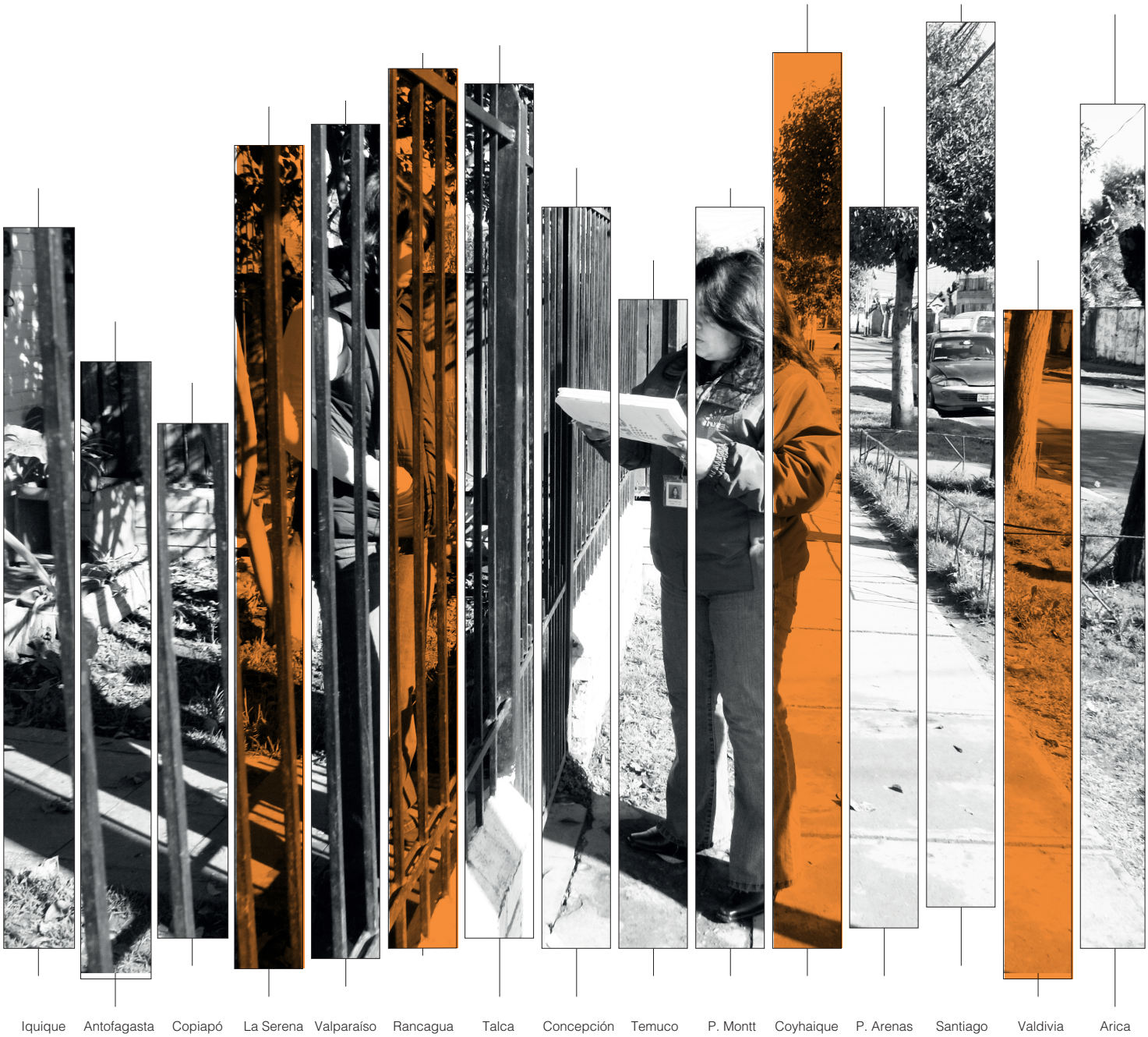
VII EPF

ENCUESTA DE PRESUPUESTOS FAMILIARES

nstituto Nacional de Estadísticas · Chile

DOCUMENTO DE TRABAJO

Métodos de imputación VII EPF:
Gastos diarios e ingresos de la actividad laboral principal y jubilaciones
VII Encuesta de Presupuestos Familiares



Subdirección Técnica
Encuesta de Presupuestos Familiares (EPF)

Métodos de imputación VII EPF:
Gastos diarios e ingresos de la actividad laboral principal y jubilaciones
VII Encuesta de Presupuestos Familiares
Diciembre / 2014

Jefe EPF: Leonardo González Allendes
Encargada equipo técnico: Beatriz Salinas Quiroga
Analista(s) investigador(es): Leonardo González Allendes
Beatriz Salinas Quiroga

Paseo Presidente Bulnes 418, Santiago de Chile
Código Postal 8330532
Fono: (56) 22892 4000
Sitio web: www.ine.cl
Correo electrónico: ine@ine.cl
facebook:chileINE
twitter:@INE_Chile
Santiago de Chile

ISBN: 978956323158-8

CONTENIDO

Índice de tablas

Índice de gráficos

PRESENTACIÓN

El Instituto Nacional de Estadísticas (INE) es el organismo responsable de producir y difundir las estadísticas oficiales de Chile, proporcionando información confiable y accesible a los usuarios para la toma de decisiones. Su producción estadística, busca permanentemente cumplir con lineamientos que permitan aumentar la confianza de la ciudadanía en la información producida y difundida, además de fomentar la aplicación de los mejores métodos y prácticas internacionales en materia de estadísticas públicas y oficiales.

En este contexto, el INE pone a disposición de los usuarios el presente texto, el cual es parte de los documentos metodológicos elaborados por el equipo de la Encuesta de Presupuestos Familiares (EPF). Estos documentos buscan dar a conocer las decisiones técnicas y metodologías de trabajo que acompañaron el proceso de producción estadística de la VII EPF.

La EPF es una investigación relevante para el INE y el país. Es la principal fuente de información para la actualización de la canasta de bienes y servicios que componen el Índice de Precios al Consumidor (IPC), así como también el principal insumo para la actualización de la línea de pobreza e indigencia para la medición de pobreza por ingresos. La EPF es, además, una fuente de abundante información para una diversidad de usuarios, tanto de instituciones públicas como privadas, ya que es la única encuesta oficial que mide ingresos y gastos de los hogares con una cobertura temporal de un año.

La EPF es una de las encuestas de hogares más compleja que elabora la institución, particularmente por el tiempo de presencia en los hogares y el volumen de información solicitada a los mismos. Esta encuesta cuenta con una larga tradición al interior de la institución, realizándose su primera versión a fines de los años 50.

A lo largo del tiempo, las EPF han introducido modificaciones tanto en sus objetivos como en la metodología de recopilación de los datos y su posterior tratamiento, a fin de mejorar la calidad de la información que se pone a disposición del país. En ese marco de perfeccionamiento continuo, la VII EPF introdujo importantes mejoras, entre las cuales se cuenta el estudio e implementación de métodos de imputación frente a la falta de respuesta parcial de gastos e ingresos.

El presente documento explica y analiza detalladamente el tratamiento dado a la falta de respuesta parcial en la libreta de gastos individuales y en la libreta de ingresos aplicada en la encuesta. Específicamente, describe el proceso de imputación de la no respuesta de gastos individuales e ingresos para la VII EPF, realizando análisis de la falta de respuesta en cuanto a su aleatoriedad, porcentaje y caracterización de la misma, revisando además los distintos métodos de imputación evaluados para los datos a imputar, y la decisión del método a utilizar para cada una de las dos temáticas tratadas en el informe.

En el marco de mejoras continuas de los productos estadísticos de la institución y al amparo de las buenas prácticas recomendadas por organismos internacionales, este documento, busca ser un aporte tanto para la discusión de futuras EPF como para la discusión sobre métodos de imputación para otras encuestas de hogares internas o externas a la institución. Al respecto, quisiera agradecer de manera especial las contribuciones realizadas por el jefe de la VII EPF, Francisco Bilbao Quiroga, y la encargada técnica de la encuesta durante el periodo 2011-2014, Rocío Miranda Rocco, bajo cuya dirección se generó el presente documento.

Por último, esperamos que esta información contribuya al conocimiento de estas materias, en cumplimiento con los principios de compromiso con la calidad, de precisión y fiabilidad, de coherencia y comparabilidad, así como de accesibilidad a la información y claridad de la misma.

Ximena Clark Nuñez
Directora Nacional
Instituto Nacional de Estadísticas

I. INTRODUCCIÓN

En la aplicación de una encuesta, es recurrente que no se pueda obtener del informante todos los datos que se solicitan. Existen diversas razones por las que el informante no entrega toda la información solicitada, entre las que se pueden mencionar el desconocimiento sobre la utilidad de la información recolectada con la encuesta, aburrimiento, vergüenza, omisión involuntaria o simplemente el no querer responder temas que pueden ser sensibles para él. La existencia de las situaciones mencionadas, nos llevan a que encontremos encuestas con valores faltantes en determinadas preguntas[1].

La ausencia de datos para algunas variables es frecuente en las encuestas y puede afectar a una o más variables, o en uno o más hogares para una misma variable, así como también puede existir presencia de datos anómalos o poco probables. Estas dos situaciones llevan a plantearse las siguientes interrogantes: ¿Podemos aceptar perder la información de un hogar por no haber recogido todas las preguntas de la encuesta mientras que al mismo tiempo el hogar cooperó proporcionando otra información relevante para el estudio? y ¿podemos descartar aquellas encuestas (en el caso de la EPF se refiere a libretas) incompletas a pesar de que el resto de la información solicitada al hogar se recolectó satisfactoriamente?

El primer y principal resguardo ante la falta de información debe venir de la planificación del trabajo de campo. En el caso de la VII EPF, la no respuesta podía estar asociada a la comprensión de las preguntas y alternativas de respuesta o a la disposición de los informantes a revelar información sensible (sobre todo en preguntas asociadas al ingreso). Para aminorar la no respuesta derivada de la dificultad de comprensión de las preguntas y sus alternativas por parte de los informantes, antes del levantamiento oficial de la encuesta se realizaron pruebas de los cuestionarios que se aplicarían posteriormente. Además, se realizó una profunda capacitación inicial de investigadores de hogares junto a una Prueba Piloto y Marcha Blanca del proyecto para posteriormente, realizar un trabajo continuo de reforzamiento conceptual durante el período oficial del trabajo de campo de la encuesta.

El trabajo de campo realizado en la Prueba Piloto y Marcha Blanca del proyecto, junto a la capacitación continua de investigadores, además de entregar conocimientos que permitieran a los encuestadores tener un argumento sólido y claro, buscaron construir mecanismos y estrategias para la generación de confianza con los informantes, medidas que buscaron disminuir la no respuesta ligada a la disposición de los informantes a revelar información sensible.

A pesar de los resguardos tomados en el trabajo de campo, la experiencia institucional, nacional e internacional muestra que la existencia de datos faltantes persiste en las encuestas de hogares. La presencia de información faltante tiene implicancias importantes para la interpretación de los resultados de una encuesta, como la disminución del tamaño muestral, la estimación de estadísticos sesgados y la falta de significancia estadística de los mismos (IBM-SPSS, 2012).

Ante la presencia de datos perdidos en las encuestas, la Oficina Europea de Estadística (Statistical Office of the European Communities [EUROSTAT], 2003) recomienda imputar el dato faltante, antes que trabajar con el dato perdido, recomienda también que si un hogar presenta muchos datos faltantes, este hogar puede considerarse como no-cooperante y la mejor opción sería eliminarlo de la muestra, ajustando los factores de expansión respectivos. En el caso de la Asociación Americana para la Investigación de la Opinión Pública (American Association for Public Opinion Research [AAPOR], 2011) considera que deben existir consideraciones explícitas a priori sobre qué requisitos distinguen una encuesta completa de una parcial y qué requisitos distinguen una encuesta parcial de una encuesta que pese a entregar información, puede considerarse no cooperante (break-off).

1 El valor faltante puede ser una respuesta válida al usar códigos de no sabe o no responde.

Con el objetivo de asegurar un nivel mínimo de calidad para la incorporación de los datos recogidos a través de los instrumentos de recolección de la información, la VII Encuesta de Presupuestos Familiares incorporó un proceso automatizado de revisión sistemática a través de una grilla técnica[2] de mínimos de calidad, la cual busca asegurar un mínimo de cantidad y calidad de datos exigidos a cada libreta y a la encuesta en su conjunto para entrar a la base de datos de los hogares sobre los que se extraerán conclusiones del fenómeno estudiado.

Dado que el objetivo de una encuesta es hacer inferencia estadística de una muestra a la población manteniendo su capacidad de reproducir la realidad, la imputación de datos es un método viable para la corrección de la no respuesta parcial, teniendo en cuenta siempre que los datos luego de la imputación deben entregar resultados consistentes con los fenómenos que busca estudiar la encuesta. No obstante, antes de estudiar las distintas formas de abordar la no respuesta parcial en la VII EPF, cabe mencionar que la mejor forma de enfrentar la falta de respuesta es evitándola. Por lo tanto, un dato imputado no se puede considerar bajo ninguna forma mejor que un dato observado.

El presente documento analiza y explica la no respuesta parcial de las libretas de gastos individuales[3] e ingresos en el contexto de la VII EPF, es decir, en concreto cómo se trataron los días faltantes en la libreta de gastos individuales y los ingresos de la ocupación principal y de jubilaciones en la libreta de ingresos de algunos integrantes de aquellos hogares considerados en la base final (quienes superan la grilla técnica)[4].

Una distinción importante que se debe realizar antes de estudiar los diferentes métodos de imputación de ingresos y de gastos individuales, es la diferencia entre la imputación de la no respuesta abordada en este documento y el ajuste por no respuesta realizado sobre el factor de expansión de los hogares para corregir la no respuesta. Este documento aborda el problema de la no respuesta parcial y no el de la no respuesta total de encuestas.

El ajuste por no respuesta presente en el cálculo de los factores de expansión responde al hecho de que existe no respuesta total de hogares en la muestra. Este ajuste se realiza a nivel de manzana y que busca repartir el peso teórico del factor de expansión de las viviendas no levantadas entre las viviendas que sí fueron levantadas en una manzana determinada para un mes determinado[5].

Previo a realizar cualquier tipo de imputación, se estableció una depuración y tratamiento de las bases de datos para la revisión de datos anómalos, con el fin de contrastar la información digitada con la registrada en las libretas o para intentar recuperar con el hogar información importante para el estudio. Se buscó minimizar de la manera más exhaustiva los errores no muestrales asociados al levantamiento y digitación de la información.

Finalmente, podemos decir que este documento busca describir el proceso adoptado de imputación de la no respuesta de gastos individuales e ingresos para la VII EPF desde el análisis de la no respuesta en cuanto a su aleatoriedad, porcentaje de no respuesta y caracterización de la misma, hasta los distintos métodos de imputación evaluados para los datos a imputar y la decisión final del método a utilizar para cada una de las dos temáticas tratadas en el informe.

"*Cada situación es diferente, y la tasa de no respuesta y su distribución espacial cambia entre encuestas, por lo que no es conveniente adoptar —a priori— el mismo procedimiento de imputación para todas las variables, en todas las encuestas*" (MEDINA, 2007)

2 Para mayor detalle sobre la grilla técnica de la encuesta, revisar la "Metodología VII Encuesta de Presupuestos Familiares", capítulo 8, pág.106, disponible en: http://www.ine.cl/epf/

3 Respecto al resto de las libretas de gastos, al igual que en las versiones anteriores de la encuesta, en la VII EPF no se realizó imputación de libretas faltantes.

4 Para una discusión sobre la no respuesta total en el contexto de la VII EPF revisar el documento "Introducción a la medición de no respuesta considerando la Hoja de Ruta (HR) en la VII Encuesta de Presupuestos Familiares" disponible en: http://www.ine.cl/epf/

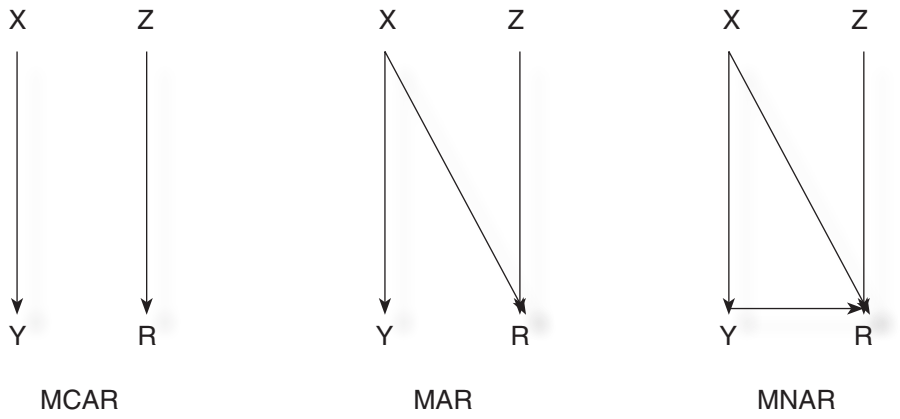
5 Este tipo de ajuste es una práctica extendida en las distintas encuestas de hogares que busca ajustar la representatividad de las viviendas que no cooperaron con la encuesta entre las que sí cooperaron con ella. Para una explicación metodológica más detallada ver "Metodología VII Encuesta de Presupuestos Familiares", capítulo 4, disponible en: http://www.ine.cl/epf/

II TEORÍA DE LA NO RESPUESTA Y VALORES PERDIDOS

Como menciona Wang (2007), las técnicas para lidiar con la no respuesta dependen del tipo de mecanismo que la genera. Si bien la formalización del concepto **mecanismo generador de la no respuesta** fue introducida en 1976 por Rubin, recién en 2002 y en conjunto con Little diseñan una prueba de hipótesis (test de Little) para diferenciar entre dos de los tres mecanismos de generación de no respuesta: MCAR (No Respuesta Completamente Aleatoria), MAR (No Respuesta Aleatoria) y NMAR (No Respuesta No Aleatoria). El test diferencia un mecanismo MAR y uno MCAR (Rubin, 1976 y Little y Rubin, 2002).

Las diferentes formas del mecanismo de no respuesta se ejemplifican en el diagrama siguiente.

DIAGRAMA 1: Diferentes formas de mecanismos de no respuesta



Fuente: Huisman, 2010.

Se pueden plantear los mecanismos en torno a cuatro vectores de información, donde X son los casos observados, Y son los faltantes, Z es un vector de componentes independientes de X e Y que explica por qué existe la no respuesta y finalmente R es un indicador de respuesta.

Como se mencionó, la prueba de Little permite distinguir entre un proceso de datos faltantes completamente aleatorio y otro de tipo aleatorio (por sus siglas en ingles MCAR y MAR). Cuando la no respuesta es no aleatoria el sesgo es eminente y por lo tanto, no se puede ignorar a la hora de analizar la información disponible y esta información se debe reportar y tener en consideración a la hora de realizar imputaciones. Sólo cuando se identifica el mecanismo de no respuesta se puede valorar si los métodos de imputación utilizados son los adecuados.

La hipótesis nula (H0) de la prueba corresponde al supuesto de que los datos con no respuesta tienen un patrón completamente aleatorio- MCAR, donde el estadístico d^2 tiene una distribución χ^2 con f grados de libertad.

Cuando no se rechaza la hipótesis nula el mecanismo es MCAR (Little, R 1988 y Medina y Galván, 2007). Los valores p superiores a 0,05 indican que

con una exigencia del 95% no se rechaza la H0 de aleatoriedad en la no respuesta. Cuando el mecanismo generador de la no respuesta sigue un patrón MCAR, la teoría plantea que es posible aplicar cualquier método de imputación, mientras que cuando se rechaza H0 se recomienda un proceso de máxima verosimilitud, debido a que en la primera etapa de estos modelos no se asume a priori el parámetro de distribución de la no respuesta y el proceso es iterativo hasta encontrar convergencia.

Cuando se explicita que la no respuesta es MCAR, entonces las unidades observadas tienen un comportamiento estadístico que se comparte con las unidades no observadas, por lo que la no respuesta se puede identificar solo gracias a las variables Z (conocidas y observadas).

Si bien lo que se busca es que la no respuesta sea MCAR, trabajar y realizar consultas sobre la información parcial (ignorando las celdas vacías), puede llevar a un investigador a tomar decisiones y realizar informes con información errónea o con un error efectivo mayor. Una de las formas de combatir y buscar soluciones para la no respuesta es a través de la imputación de datos. Existen variados métodos de imputación, que se ajustan de acuerdo a las particularidades de la información contenida en una base de datos. Respecto a la forma para determinar cuáles son las mejores aproximaciones a los datos, existen numerosos documentos de trabajo y metodologías planteadas para las encuestas[6].

Por otra parte, si bien el desarrollo de las herramientas de cálculo para la imputación pone al alcance del investigador el uso de métodos más sofisticados, en una encuesta con diseño complejo en la que el porcentaje de datos faltantes es bajo y está concentrado en una porción con características especiales. Los autores Medina y Galván (2007) señalan que potencialmente un método de imputación simple reproduzca mejor las características de la subpoblación de interés que un algoritmo que considera a toda la muestra[7].

En este documento se recopiló información sobre cuáles son los pilares teóricos de los métodos recomendados para lidiar con la no respuesta de los conceptos objetivos para la VII EPF, a saber, el gasto y el ingreso. Los métodos elegidos para realizar los procesos de imputación de datos en la VII EPF fueron seleccionados teniendo en cuenta las recomendaciones internacionales y las mejores prácticas detectadas en la revisión bibliográfica.

La práctica internacional muestra que existen diversas adaptaciones de técnicas estadísticas para realizar imputaciones de gastos y, tal como Dayal et al. (2001) observan, para la imputación de gastos no existen pruebas formales apropiadas para distinguir entre conjuntos alternativos de datos imputados, razón que ayuda a explicar que existan diversas metodologías para realizar este tipo de imputaciones.

Cabe señalar que existen distintos patrones de consumo dadas las diferentes realidades internacionales, como también diferentes metodologías de captura y procesamiento de los datos. Debido a esto, la experiencia internacional se debe/puede tomar como referencia y evidencia, mas no como una rutina que se debe igualar completamente.

En general el período de referencia para ingresos es el mes anterior al momento de realización de la encuesta, sin embargo para el gasto, el período de referencia varía dependiendo del tipo de gasto que se está registrando. La medición del gasto, que combina distintos períodos de referencia, es ajustada en el caso de Chile al mes de referencia.

Por otro lado, los gastos de los hogares son compilados a través de más de un instrumento (cuestionario o libreta), que consideran diferentes períodos de referencia. Los gastos del día a día son recopilados a través de una libreta de autoregistro, quincenal para el caso de Chile, en la cual cada integrante del hogar de quince años o más debe completar. El período de referencia para la aplicación de estos instrumentos varía de país en país. Mientras que Argentina, Brasil, Colombia, Costa Rica, Ecuador, Italia, México, Perú y Uruguay utilizan una semana para el registro de estos gastos, Australia, Canadá, Chile, EEUU, Francia, Irlanda, Portugal, Sudáfrica y UK utilizan libretas de registro bisemanales. España utiliza una doble metodología donde el administrador de gastos del hogar debe registrar durante dos semanas los gastos del día a día suyos y los adquiridos para satisfacer las necesidades de todos los miembros del hogar, mientras que el resto de los integrantes del hogar deben registrar durante una semana sus gastos del día a día.

Otro tipo de gastos son recopilados a nivel hogar y utilizando distintos períodos de referencia, por ejem-

6 Por ejemplo, una búsqueda sobre el artículo de Little (2002) arroja un total de 32 mil artículos en varias áreas de investigación como la biométrica, ingeniería eléctrica, estudios laborales, etc.

7 Se recomienda a Medina y Galván (2007) para una revisión detallada de los métodos de imputación usados en Latinoamérica, como también una comparación de las ventajas y limitaciones de cada método.

plo el mes anterior para las cuentas mensuales y los 3, 6 o 12 últimos meses para los gastos de mayor cuantía y menor frecuencia de compra[8].

Estas diferencias en los períodos de registro entre gastos e ingresos, además de la subdeclaración en algunos de los ítems (aunque en otros podría existir sobreestimación) pueden hacer que el ingreso y el gasto presenten grandes diferencias. Al respecto también se encuentra la influencia que pueda generar el gasto intertemporal o gasto con crédito. Estos desbalances entre gastos e ingresos derivados del registro de distintos períodos de referencia son reconocidos por Dayal et al. (2000), quienes plantean que estas diferencias pueden ser específicamente acusadas en el caso de los hogares con trabajadores independientes. Reforzando la idea anterior sobre el desbalance entre ingresos y gastos, la Oficina Nacional de Estadísticas del Reino Unido también reconoce que dado los diferentes períodos de recolección de la información, no es posible esperar que el gasto y el ingreso estén balanceados para un hogar o para un grupo de hogares (Office for National Statistics [ONS], 2010). Esta situación cobra peso al momento de considerar las variables que más pueden explicar el gasto y el ingreso al realizar imputaciones. Es así reconocido que este tipo de encuestas puede tener sesgos inducidos por la mezcla de distintos períodos de referencia de los instrumentos de recolección de datos.

Para el caso de la imputación de gastos, tal como observan Dayal et al. (2001), el introducir otras variables además del ingreso disponible del hogar en la formación de clusters de imputación, mejora las imputaciones de gastos efectuadas y replica de mejor forma la estructura del gasto a través de los distintos deciles de ingreso.

En la revisión internacional sobre métodos de imputación de gastos, se pudo constatar que no todas las encuestas de presupuestos familiares explicitan la metodología de imputación o los ajustes realizados a los datos de gastos. En este mismo sentido, si bien la mayoría de encuestas de presupuestos familiares realiza imputaciones sobre los gastos individuales, también existen casos en los cuales no se realiza, como en la encuesta de presupuestos familiares de Estados Unidos (aunque sí realiza imputación de ingresos y de otras variables demográficas).

A continuación se resume una la revisión internacional sobre la imputación de gastos. La revisión se centra sobre la imputación de libretas de gastos individuales, ya que el trabajo sobre la imputación de otras variables es más amplio y no pretende ser abordado en el presente documento:

Argentina. Argentina imputa libretas de gastos individuales sin respuesta a través de un procedimiento hot-deck jerárquico, en el que un cuestionario de gastos individuales faltante se imputa con el cuestionario de un donante que comparte algunas variables seleccionadas. Además, se imputaron servicios para la vivienda, impuestos inmobiliarios, alimentos y bebidas de cuestionarios de registros incompletos. La imputación se realizó a nivel de artículos, seleccionando donantes para los datos faltantes entre registros de características similares (Instituto Nacional de Estadísticas y Censos [INDEC], 2007).

Australia. Las libretas de gastos individuales incompletas son rellenadas imputando el registro de gastos de la misma persona, es decir, si de las dos semanas la persona sólo contestó una, entonces la semana que entregó información es utilizada para imputar la semana sin información. Para las libretas de gastos individuales faltantes, se imputó el gasto utilizando libretas de donantes que contestaron el 100% de los días y que compartían características comunes (Australian Bureau of Statistics [ABS], 2012).

Canadá. La imputación de gastos individuales es realizada cuando los informantes entregan una detalle de gasto sin el monto del gasto o cuando se entrega un monto total sin realizar un desglose de este monto. En estos casos, la imputación del gasto o la desagregación de éste, es hecha considerando un donante que comparta características comunes (vecino cercano), respetando la estacionalidad y la zona geográfica del informante (Statistics Canada, 2013).

España. Si existe algún valor declarado en las libretas de registros diarios sin su monto (libreta individual de cuentas que es llenada durante siete días por cada integrante del hogar salvo el administrador de gastos que llena la libreta de cuentas del hogar con sus gastos y los gastos para atender las necesidades básicas del hogar durante 14 días), se hacen cluster de informantes según características comunes (como comunidad autónoma [región], estrato,

8 Para una explicación más detallada de los distintos instrumentos utilizados en la EPF véase la METODOLOGÍA, VII Encuesta de Presupuestos Familiares, disponible en www.ine.cl/epf

etc.), se calculan los gastos unitarios (dividiendo monto por cantidad), se ordenan aleatoriamente los gastos de cada grupo y se imputa el monto unitario de gasto de la observación donante que se encuentra inmediatamente arriba del gasto a imputar (imputación hot-deck). Dado que la selección del donante es aleatoria, cada vez que se repita el procedimiento puede variar el resultado (Instituto Nacional de Estadística [INE] España, 2010).

En este mismo sentido, la EPF de España realiza un ajuste de gasto para la segunda semana de registro del administrador de gastos del hogar[9]. Este ajuste se hace sólo para los gastos en alimentos y consiste en ajustar la magnitud de los gastos de la segunda semana de registro utilizando la información sobre la magnitud de los gastos de la primera semana. Por lo tanto, una corrección para la segunda semana de registro de sus informantes considera el efecto del cansancio en el llenado de dicha semana. Este ajuste se realiza, ya que se considera que el efecto cansancio en el llenado de libretas es también una de las principales causas que hace que existan libretas incompletas de gastos individuales en general para los diferentes países.

Por otro lado, en España las libretas de gastos individuales no recogidas son imputadas a través de un proceso bietápico de imputación que es realizado de forma trimestral y para cada comunidad autónoma. El proceso consiste en imputar el monto de gasto del cuestionario en la primera etapa para luego en la segunda distribuir el monto entre los distintos códigos de gasto. En ambas etapas de la imputación se forman clusters considerando las características más relacionadas con el objetivo de la imputación.

Inglaterra. La encuesta de presupuestos familiares del Reino Unido imputa libretas de gastos individuales para las personas que no contestaron la libreta a través de donantes que sí lo hicieron (ONS, 2010). Para buscar el donante de gastos individuales, se aplica un sistema de puntaje que identifica la persona con características más cercanas. Este sistema considera la edad del informante (8 puntos), relación de parentesco (4 puntos) con la persona definida como representante del hogar (equivalente al jefe de hogar), condición de actividad económica (2 puntos) y mes de aplicación de la encuesta (1 punto).

La potencial persona donante que tiene el puntaje más alto es seleccionada y la libreta de gastos individuales del donante es copiada con todos sus registros para reemplazar la libreta vacía de la persona que no contestó. Para usar una persona como donante, ésta tiene que obtener como mínimo un puntaje de 8 puntos. Para el año 2010, 181 hogares tuvieron libretas de gastos individuales imputadas, los que representaban aproximadamente un 3,5% de los hogares que respondieron la encuesta.

Italia. El tratamiento de los datos faltantes difiere si la variable a imputar es cuantitativa o cualitativa. Dado que el gasto individual es una variable cuantitativa, primero se verifica que cada partida del gasto se encuentre en intervalos de aceptación, mientras que en una segunda etapa, llamada etapa de corrección, se utiliza información de un donante para corregir los casos anómalos. Esta metodología es aplicada a través de un sistema llamado reconstrucción de la información con donación automática (Istituto Nazionale Di Statistica [ISTAT] Italia, 2011).

Para la realización de la imputación de gastos debe considerarse el rol que juegan los gustos y preferencias en la estructura de gastos de las personas, lo que responde a componentes subjetivos del proceso que deben ser considerados para acercarse de la forma más adecuada posible a los distintos patrones de consumo de los sujetos a ser imputados. Para realizar imputaciones de este tipo de gastos, el método se basa en la búsqueda de una persona con características similares a aquella que presenta el dato faltante.

Dentro de las modificaciones realizadas a los cuestionarios, la VII EPF introdujo la distinción entre días sin gasto y días sin registro en el llenado de sus cuestionarios de gastos individuales, lo que busca facilitar la imputación de gastos individuales para aquellos informantes que contestan la libreta de autollenado de forma parcial. Esta distinción genera un tratamiento distinto de los dos tipos de información, ya que los días sin registro son imputados, mientras que los días con gasto cero son tratados como un registro válido[10].

La imputación de gastos de la VII EPF se realiza para los informantes que tienen un registro parcial de datos en sus libretas de gastos individuales. Para el proceso de la VII EPF no se realizó imputación so-

9 España utiliza una metodología de recolección de la información a través de 26 submuestras de 14 días cada una, con lo cual los cuestionarios bisemanales de recolección de la información no se corresponden con los días calendarios.

10 Esta distinción es importante, ya que existen días en que los informantes no realizan gastos y este tipo de información debe quedar reflejada en los formularios de gastos. Los días sin gastos se consideran como un registro válido y si un informante declara que no realizó ningún gasto, dicho día se considera "día con registro" (donde el gasto del día es igual a cero").

bre libretas completas de gastos individuales, sino que se imputó la no respuesta parcial de libretas de gastos individuales pero no la no respuesta total de gastos individuales.

Si bien existen diversas metodologías desarrolladas por diferentes países respecto a la imputación de libretas de autollenado, se decidió no imputar libretas de gastos individuales completas, ya que las EPFs anteriores no realizaron dicha imputación y para realizar este tipo de imputaciones, debe desarrollarse un modelo que busque reflejar las preferencias de los consumidores. Este estudio sobre los patrones de consumo plantea una próxima línea de investigación para el equipo, lo que en definitiva es requisito previo para la imputación de libretas completas.

El equipo técnico de la VII EPF se planteó la necesidad de probar distintas formas de imputación de los gastos diarios de los integrantes del hogar, en total tres. El primer método es el ajuste por Factor de no Respuesta (FNR). El segundo método es el Ajuste por peso diario, siendo el tercero la imputación por el método Hot-Deck. El desarrollo y explicación de cada método corresponde al punto IV.

Por otro lado, la experiencia internacional y nacional en la imputación de ingresos es más variada, ya que más encuestas consultan a los hogares sobre este tema y cada una de ellas plantea una solución a la no respuesta. La revisión sobre las experiencias internacionales se concentra en las encuestas símiles a la EPF, mientras que la revisión de la nacional se extiende a otras encuestas. La experiencia internacional en la imputación de ingresos en encuestas similares a la EPF muestra una variedad de métodos usados y en general se resalta la necesidad de corrección por no respuesta debido a su link con el gasto.

Australia. Los ingresos se imputan con la definición de donantes que coinciden en un set de variables, como espacio geográfico, sexo, edad, situación laboral, así como el nivel del gasto. Entonces una vez determinado el conjunto de donantes para cada caso, el valor imputado se elige aleatoriamente de éste. La desventaja de este tipo de elección por lo tanto es que no es completamente replicable, ya que probablemente no se elegirá siempre el mismo valor (ABS, 2012).

Canadá. En este caso las personas tienen la posibilidad de autorizar a la encuesta acceder a la declaración de impuestos sobre sus ingresos, de donde se extrae esta información. Si esta vía para consignar la información no está disponible se determina un donante y en este caso no explicita las variables jerárquicas (Statistics Canada, 2013).

Estados Unidos. Los ingresos de la encuesta análoga se imputan a partir del 2004. En este caso se utiliza la imputación múltiple para asignar el valor del ingreso cuando este dato sea faltante para el hogar. Para las versiones anteriores al 2004, simplemente se dejaba una variable de alerta para tipificar que ese hogar no respondió información de ingresos (Bureau of Labor Statistics [BLS] Estados Unidos, 2008; Paulin y Ferraro, 1994).

Chile. Las encuestas revisadas en las que se imputa la información faltante son:

La **Encuesta Suplementaria de Ingresos (ESI)** que se realiza cada año, históricamente ha reportado una tasa de no respuesta de ingreso por debajo del 10%[11]. El método de imputación utilizado en la NESI es una adecuación del Hot-Deck adaptado para la encuesta con las variables disponibles en su cuestionario. Se basa en una matriz que se aplica a nivel de las personas que reciben ingresos y busca a una persona o a un grupo de personas donantes con información de ingresos que compartan características similares a las personas que no tengan la información. Para reemplazar el valor se utiliza la mediana de los ingresos del grupo donante.

La **Encuesta de Caracterización Socioeconómica Nacional (Casen)**, realiza un proceso similar de corrección por no respuesta en los conceptos de sueldos y salarios, trabajo independiente o jubilaciones y pensiones, pero no informa acerca de su respectivo monto. La metodología utilizada también es una variante del Hot-Deck, pues busca un set de características tanto del trabajo como de la persona para encontrar un grupo de personas similares que tengan ingreso. Lo denominan cruce sistémico y se consideran siete variables: categoría ocupacional, región, parentesco (jefe de hogar; no jefe de hogar), sexo, nivel educacional (años y rangos) y rama de actividad económica (recodificada a nivel de gran división).

En esta encuesta, a diferencia de la anterior, se utiliza la media del grupo de donantes. Aunque tanto la media como la mediana son estadísticos de tenden-

11 En la tasa reportada se incluyen los casos de rechazo no responde y los no sabe, que se originan porque el informante desconoce la información. Entre los asalariados, por ejemplo, para el año 2012 se reportó un 5,2%, mientras que para los trabajadores independientes un 5,9%.

cia central, puede existir una diferencia entre el uso de una u otra en la determinación del valor a imputar. La principal diferencia entre usar la media o la mediana para hacer la imputación de dato con respecto a un grupo de donantes, es la distribución en el grupo definido como donante. Mientras más homogéneo sea el grupo menor será el impacto entre usar la media y la mediana, la segunda consideración será el tamaño del grupo de donantes.

La **Encuesta Financiera de hogares (EFH)** que realiza el Centro de Microdatos de la Universidad de Chile por encargo del Banco Central de Chile, también presenta una alternativa para la imputación de ingresos diferente y ajustada a sus propios datos. En este caso utilizan la imputación múltiple[12], con 30 repeticiones. Este proceso no sólo se aplica a ingresos, pues esta encuesta también consulta sobre activos, patrimonio y pasivos de propiedad del hogar.

La revisión bibliográfica de la información disponible en los cuestionarios y el avance tecnológico dieron paso a la decisión de analizar y probar diferentes métodos de imputación para los ingresos. Tanto en la revisión internacional como en la nacional se distingue el énfasis en los ingresos del trabajo y, en una menor medida, en el ingreso de las jubilaciones, razón por la que se eligen éstas como las dos fuentes afectas a la corrección por no respuesta en la VII EPF. El ajuste de los ingresos laborales por las variables sociodemográficas disponibles en la encuesta es acorde a la teoría de capital humano y del mercado laboral.

12 La descripción de cómo funciona el método de imputación múltiple será abordado en el Capítulo V sección B.4.

III. METODOLOGÍA DE ANÁLISIS

A continuación se presentan las fases de tratamiento de la información utilizadas por el equipo técnico de la encuesta para la aplicación de métodos de imputación. El proceso implementado para la VII EPF se conformó por tres fases, las mismas que dan forma al presente documento. El objetivo de cada fase es el mismo para ambos conceptos, gastos e ingresos.

DIAGRAMA 2: Proceso de imputación en la VII EPF para gastos individuales

	Preparación **Objetivo:** conocer y cuantificar lo que se debe imputar	**Aplicación** **Objetivo:** aplicar las metodologías de imputación seleccionadas para la evaluación	**Análisis** **Objetivo:** conocer los resultados obtenidos de cada método
GASTO	1.- Definición del universo: ¿Cuántos? 2.- Clasificar y priorizar la no respuesta. ¿Quiénes? 3.- Cuantificar y caracterizar la no respuesta en función a la clasificación. ¿Cómo son? 4.- Aplicar prueba Little. 5.- Determinación de las variables más correlacionadas con la variable a imputar ¿Se pueden utilizar los diferentes métodos de imputación planteados?	1.- Ajuste por factor de no respuesta. 2.- Ajuste por peso diario. 3.- Imputación Hot-Deck	1.- Análisis de resultados. 2.- Análisis de los estadísticos de tendencia central, concentración y de distribución. 3.- Análisis de la estructura de gastos. 4.- Toma de decisión sobre el método a utilizar
INGRESO		1.- Métodos econométricos - Regresión Heckman 2.- Imputación múltiple, 3.- Imputación Hot-Deck, 4.- Expectation Minimization (EM)	1.- Análisis de resultados. 2.- Cálculo de los estadísticos para la toma de decisión: A) Coeficiente de variación, B) Raíz del error cuadrático medio C) Error absoluto medio 3.- Toma de decisión sobre el método a utilizar

La primera fase la denominamos de preparación. En esta fase se presentan los detalles de lo que será imputado. En el caso de gasto se trabaja sobre las características de los informantes que responden su LGI de forma parcial, ya que sobre ellos se realizará la imputación de gastos. Para los ingresos, en tanto, se trabaja sobre aquellos informantes ocupados cuyos períodos de referencia de su actividad laboral principal y de ingresos del trabajo correspondan al período de referencia de la encuesta, sin que hayan declarado los montos. En seguida, se describe la no respuesta en base a las variables con mayor correlación con el valor de las variables de interés, buscando describir el mecanismo de la no respuesta detrás de la información de la encuesta. En este mismo sentido, esta fase incluye la prueba de aleatoriedad de la misma. El resultado de la prueba es el supuesto principal para la aplicación de algunos de los métodos de imputación implementados en la siguiente fase.

La segunda fase se trata de la aplicación empírica, donde se utilizan los distintos modelos de imputación y se analiza el ajuste de éstos a los datos de la VII EPF. En el diagrama se especifican los métodos aplicados para cada concepto. En cada sección de esta fase, y para gasto e ingreso, se explica teóricamente el método de imputación y la implementación del método en la encuesta.

Finalmente, la tercera fase consta del análisis y un proceso de toma decisión, donde basándose en los insumos obtenidos en las anteriores fases y otros estadísticos de ajuste, se recomienda el método a usar en las imputaciones de no respuesta para los datos particulares.

La metodología de imputación planteada en este documento pretende guiar la toma de decisión y no establecer un único método correcto para cualquier caso, ya que los datos recopilados de los hogares harán variar los resultados de cada fase, por lo que se deberá utilizar el proceso completo para cada nuevo set de datos.

Antes de continuar, es importante señalar que el proceso de imputación en la implementación se realizó primero en las variables de ingresos, pues para el método Hot-Deck en las variables de gastos diarios se utiliza la variable de ingresos imputados en la fabricación de la matriz de exigencia para la elección del vecino cercano.

IV. GASTOS. LIBRETA DE GASTOS INDIVIDUALES (LGI)

La imputación de gastos en la VII EPF se realiza sobre la Libreta de Gastos Individuales. Esta libreta es entregada a todos los integrantes del hogar de quince años y más. En ella cada informante debe completar, día a día, el detalle de todos los bienes y servicios adquiridos durante la permanencia de la libreta en el hogar (aproximadamente quince días). Tiene por objetivo registrar todos los gastos en productos y servicios de consumo final que realice cada integrante del hogar de quince años o más. Además se busca registrar en esta libreta el consumo de productos de los hogares provenientes del autosuministro[13]. En general quedan registrados en esta libreta todos los gastos de carácter intramensual que realizan los integrantes del hogar.

La imputación se centra en la LGI, dado que en ella se registra la información sobre los gastos individuales de las personas que forman el hogar. Es una libreta de suma importancia para la aplicación de este estudio, ya que los gastos capturados en ella representan sobre el 48% del total de los gastos de la encuesta (sin considerar el arriendo imputado) y asimismo se tiene información precisa sobre los días que no presentan información. Además está ampliamente documentado el hecho que los informantes no contestan o entregan información parcial sobre sus gastos en las libretas de gastos individuales de las distintas EPF. El resto de las libretas utilizadas en la EPF recogen información sobre el gasto a nivel hogar y en general presentan muy baja no respuesta[14].

En la Libreta de Gastos Individuales se revisan los validadores de los días de registro para cada integrante del hogar[15]. Se establece para cada persona la cantidad de días con registro (gasto mayor o igual a cero) y días sin registro (no existe información para el día en cuestión, no se sabe si hubo o no gasto)[16].

Considerando que la imputación de gastos se realiza solamente sobre la libreta de gastos individuales, a continuación se describe brevemente el funcionamiento de dicha libreta.

La muestra mensual es dividida en dos quincenas por razones operativas[17]. Los gastos registrados en cada instrumento de levantamiento de la información son ajustados al mes de referencia con el fin de presentar el gasto promedio mensual de los hogares representados en la muestra.

Al tener dos quincenas con equivalente cantidad de hogares obtenidas de una muestra dependiente, los gastos de cada submuestra son ajustados al mes de referencia obteniéndose el gasto promedio mensual de los hogares según se observa en el Diagrama 3.

DIAGRAMA 3: Ejemplo de períodos de referencia quincenales de los meses de noviembre y diciembre

Quincena 1	Quincena 2
01/11 – 15/11	16/11 – 30/11

Noviembre 2011

Quincena 3	Quincena 4
01/12 – 15/12	16/12 – 31/12

Diciembre 2011

13 Se entiende por autosumistro aquellos productos que provienen de alguna empresa, comercio, minimarket, kiosko, etc. de algún miembro del hogar, por lo tanto han sido adquiridos de forma gratuita o semigratuita. Ejemplos de autosuministro podemos encontrarlos en hogares que son dueños de una tienda comercial adosada al hogar.

14 La LGI junto a la libreta de gastos del hogar (LGH, 27% del gasto total aproximadamente) y la libreta de gastos del recuerdo (LGR, 19,6% del gasto total aproximadamente) acumulan sobre el 95% del gasto total de la encuesta (sin considerar el gasto en arriendo imputado).

15 Los validadores se revisan para que sean consistentes con la información registrada, es decir, no puede haber días con el código “día sin gasto” o “día sin registro”, mientras que todos los días que no presentan gastos deben tener el código de “día sin gasto” o “día sin registro”.

16 Esta información permite determinar si una encuesta cumple con los mínimos de calidad para ingresar en la base final de gastos de la VII Encuesta de Presupuestos Familiares.

17 Dependiendo del mes de referencia, la quincena de referencia posee 14, 15 o 16 días. En términos operativos se habla de la quincena de referencia aunque no sean exactamente quince días.

La libreta de gastos individuales busca recoger todos los gastos diarios (de consumo frecuente) que realicen los miembros del hogar como gastos en pasajes de transportes, alimentos, servicios personales, cargas de bencina, etc. Tal y como se explicó anteriormente, no se deben registrar los gastos de elevada cuantía y baja periodicidad de adquisición o aquellos gastos que presenten regularidad en el pago[18]. Por ejemplo, si se compra un diario, éste debe ser registrado en la LGI, pero el pago de una suscripción mensual a un diario debe ser registrado en la LGH, ya que dicha libreta recoge los gastos que presentan cierta periodicidad en el gasto.

La situación ideal es que todos los integrantes del hogar de quince años o más llenen a cabalidad sus libretas de gastos individuales, pero en la práctica de este tipo de encuestas es frecuente que los informantes no contesten todos los días que les son solicitados. Al respecto, la libreta de gastos individuales puede presentar tres situaciones según el llenado y la constancia de los distintos informantes:

i. **El informante aporta información de gasto sobre todos los días consultados.** Este es el caso ideal, por lo que se cuenta con información sobre el gasto referente a todos los días de la quincena de registro. En este caso el informante ha respondido la libreta de forma completa.

ii. **El informante no informa gasto para algunos de los días consultados.** La segunda situación es tener una la libreta con información parcial. Se entiende por información parcial aquellos casos en que las libretas presentan al menos un día con registro (ya sea día con gasto cero o día con gasto).

iii. **El informante no informa gasto para ningún día consultado.** La tercera situación corresponde al caso en el que una libreta de gastos es rechazada (el rechazo puede ser explícito o implícito). Esto puede ocurrir para cada informante del hogar al comienzo de la quincena, durante el transcurso o al final de la quincena[19].

La necesidad de imputar los gastos diarios de los integrantes del hogar surge por la subestimación del gasto que se obtendría derivada de los días de aquellos informantes que presentan un registro parcial de gastos individuales. Si los informantes no registran todos los gastos diarios que realizan, se afectan a la baja las estimaciones poblacionales del gasto por hogar.

Dado que se está trabajando sobre la estructura de gastos de las personas, para realizar la imputación de los gastos individuales debe considerarse el rol que juegan los gustos y preferencias de cada sujeto que será imputado, lo que responde a componentes subjetivos que deben ser incluidos para acercarse de la forma más adecuada posible a los distintos patrones de consumo de los distintos informantes. Es por esta razón que la mayoría de los métodos de imputación buscan imputar los días faltantes con información de la misma persona o con los datos de un vecino cercano que comparta características sociodemográficas comunes.

A. FASE DE PREPARACIÓN

Antes de describir los distintos métodos de imputación de la libreta de gastos individuales, es importante caracterizar la no respuesta e identificar qué porcentaje de datos faltantes respecto al total de los datos serán imputados. Además esta sección entrega evidencia sobre las variables bajo las cuales la no respuesta se comporta de forma aleatoria como requisito para aplicar imputaciones. También se presentan las variables que más se correlacionan con la variable a imputar (gasto individual diario), ya que dichas variables deben ser consideradas al momento de explicar la estructura de gastos de los informantes pues están relacionadas con la variable de interés.

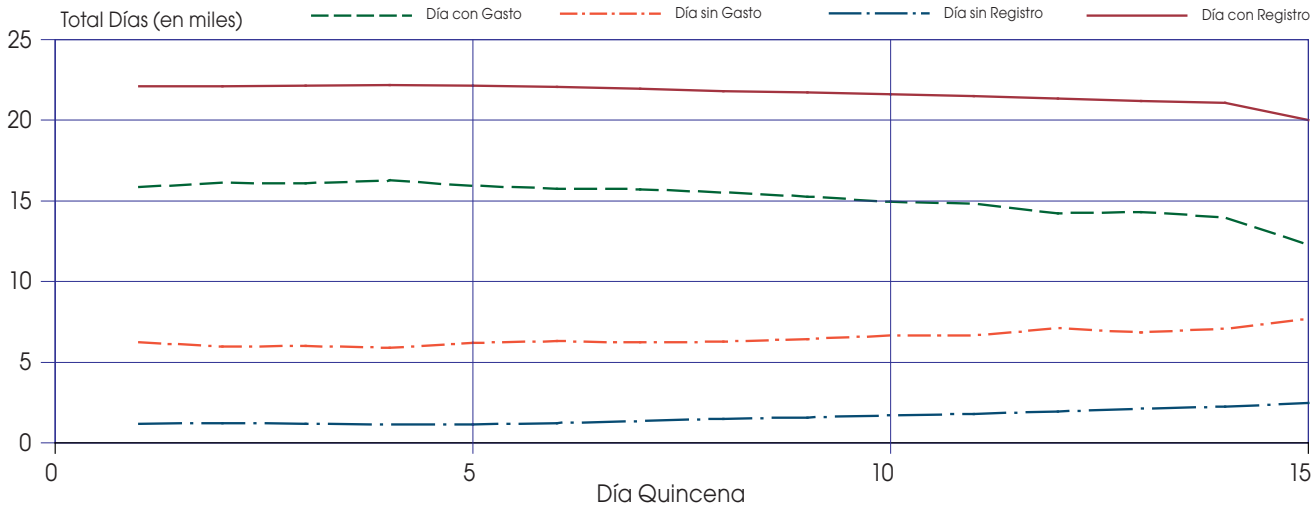
18 Esta distinción la realiza el encuestador de hogares en el momento de revisar la información completada por los integrantes del hogar. Estos gastos deben quedar registrados en la LGR y la LGH. Para una explicación más detallada de los instrumentos, consultar la metodología y los manuales de trabajo de campo de la encuesta disponibles en www.ine.cl/epf.

19 Algunos informantes retuvieron la libreta durante la quincena completa de levantamiento y la devolvieron vacía a pesar de las insistencias realizadas por los encuestadores. Muchas veces es difícil acceder de forma directa a todos los integrantes del hogar, por lo que este tipo de rechazo se considera rechazo implícito, es decir, el investigador de hogares no tuvo nunca una negativa directa del informante a no participar en la encuesta.

1. Características generales de la no respuesta en la Libreta de Gastos Individuales

A continuación, se realiza un análisis estadístico de las distintas variables a ser consideradas para la imputación. La revisión estadística comienza analizando cómo se distribuye la respuesta global de días sin registro a través de la quincena de levantamiento de la información, para después centrarse en libretas y características personales de los informantes de la Libreta de Gastos Individuales (LGI). Cabe destacar que las LGI son aplicadas a individuos de quince años o más, por lo tanto, el análisis de la no respuesta parcial de estas libretas es sobre este grupo de informantes.

Gráfico 1: Tipos de Registros. Libreta de Gastos Individuales



FUENTE: VII EPF-INE.

El Gráfico 1 muestra el comportamiento de los registros a través de la quincena de levantamiento de la información. Se representan en el gráfico el total de días con gastos, el total de días sin gastos, el total de días sin registros y el total de días con registros; esta última categoría corresponde a la agregación de los días con gasto y los días con gasto igual a cero. Se puede observar con el transcurso del paso de la quincena que la cantidad de días con registros va decayendo, mientras que la cantidad de días sin registros aumenta; esto es un comportamiento frecuente en este tipo de encuestas donde el cansancio de llenado del instrumento hace que la cantidad de información entregada por los informantes decaiga con el paso de la quincena.

El Gráfico 1 considera sólo los quince días de todas las quincenas, eliminando el último día de aquellas quincenas que tuvieron una cantidad de días mayor a quince. Esto se realiza para homogeneizar en el análisis el número de días por quincena[20]. El gráfico muestra una marcada caída para el día quince, esto está influido en parte por la inclusión en el análisis

20 El último día de las quincenas operativas de 16 días ha sido excluido del análisis pues distorsiona el análisis del llenado de la quincena al contar con un día más. De las 24 submuestras de levantamiento, existen siete submuestras de 16 días, dieciséis submuestras de 15 días y una de 14 días en febrero.

de la segunda quincena de febrero que tuvo catorce días de levantamiento y además por el retiro de LGI que se realizaba a partir del último día de levantamiento de la quincena.

Se puede apreciar en el gráfico también un aumento en los últimos días de llenado del instrumento de la cantidad de días sin gasto. Como hipótesis plausible, esta subida de días sin gasto puede atribuirse al cansancio de los informantes en el llenado del instrumento, quienes en vez de registrar los gastos de sus últimos días declaran directamente "día sin gasto" en vez de "día sin registro". Sin embargo, ante esta situación se acepta como verdadera la información entregada por los informantes.

En el Anexo B es posible apreciar la misma información presentada en el Gráfico 1 pero diferenciada por el tipo de quincena. En los gráficos, se aprecia claramente la caída del registro de los días sin gasto sobre todo en el último día de la quincena de referencia acompañada de la subida de los días sin registro.

Se puede observar también el comportamiento del registro diario más errático del segundo cuadro, ya que tan sólo agrupa siete quincenas.

Para efectos de la imputación de gastos de la VII EPF, la imputación se realiza sobre la respuesta parcial de aquellas libretas contempladas entre las libretas contestadas, por lo tanto las libretas de gastos individuales con no respuesta total no fueron imputadas. Para determinar el porcentaje exacto de libretas no contestadas, debemos excluir del cálculo aquellas libretas de gastos individuales de los miembros del hogar que presentan códigos de entrega de libretas de gastos individuales 84, 85, 86 y 87 contenidos en la hoja de ruta[21], dado que dichos códigos corresponden a miembros del hogar mayores de quince años que, por las razones contenidas en cada código, no estaban en condiciones de contestar la LGI (el significado de cada uno de estos códigos es detallado en la tabla que aparece a continuación).

Tabla 1: Categorización particular de respuestas de libretas de gastos individuales

Códigos de Entrega	LGI No Contestadas	LGI Contestadas*	Personas sin LGI	Total
81. Entrega de forma personal.	518	13.624	-	14.142
82. Entrega de forma indirecta (a través de un miembro del hogar).	2.606	9.694	-	12.300
83. No entrega, menor de 15 años (todo menor de 15 años debe llevar este código).	-	-	7.851	7.851
84. No entrega, persona con capacidades mentales no aptas para contestar el estudio, Ej.: Síndrome de Down, Alzheimer.	-	-	284	284
85. No entrega, persona con capacidades físicas no aptas para contestar el estudio, Ej.: postrado.	-	-	270	270
86. No entrega, persona con síndrome de adicción.	-	-	53	53
87. No entrega, miembro del hogar no ubicable en el periodo de referencia, Ej.: motivos vacacionales, trabajos fuera del hogar entre otros.	-	-	769	769
			TOTAL	**35.669**

*Se consideran contestadas aquellas LGI que tengan al menos un día de registro.
Fuente: VII EPF

La Tabla 1 muestra la caracterización de libretas de gastos individuales para el total de personas presentes en la encuesta, es decir, las personas de los hogares que cumplen con los requisitos de calidad de la grilla técnica. Se puede apreciar en la tabla que 7.851 informantes no debían contestar la encuesta porque eran menores de quince años, este registro no es un problema, ya que dichos informantes no tienen asociadas LGI. Del total de personas, 27.818 informantes declararon tener quince años o más y de éstos, 1.376 no contestaron la LGI porque presentaban los códigos 84, 85, 86 u 87, por lo tanto no correspondía que contestaran la LGI y no se contabilizan como no respuesta de LGI (estos casos representan aproximadamente un 5% del total de personas de quince años o más).

21 Los códigos utilizados corresponden a los códigos de asignación de LGI contenidos en la Hoja de Ruta de la encuesta. Ver Anexo A. Para una explicación más detallada sobre la hoja de ruta de la encuesta, consultar el Tomo II del Manual del Trabajo de Campo de la VII EPF, disponible en http://www.ine.cl/epf/

Dado que un hogar puede estar formado por varios informantes de quince años y más, es posible que existan distintas combinaciones en el llenado de LGI al interior de los hogares. El análisis y los procesos de imputación de gastos individuales se realizan a nivel persona y no a nivel hogar. Para ejemplificar esta situación supongamos un hogar con tres informantes de quince años o más. Mientras que uno contestó su LGI de forma completa, otro puede haber contestado una semana y el tercero ningún día. Estos tres casos son tratados de forma diferente según se explicará en cada método de imputación de gastos[22].

La Tabla 2 considera la respuesta para aquellos códigos para los que debía responderse la LGI, es decir el código 81 (entrega de forma personal) y el código 82 (entrega de forma indirecta). Los análisis y cuadros que se realizan a partir de aquí son realizados considerando sólo los códigos 81 (entrega de forma personal) y 82 (entrega de forma indirecta) de entrega de libretas de gastos individuales consignados en el módulo de control de LGI de la Hoja de Ruta de la encuesta[23].

Considerando sólo las LGI que debían ser contestadas, 23.318 personas (88,2%) contestaron al menos un día de registro, mientras que 3.124 (11,8%) personas no contestaron la LGI. Tal como se mencionó con anterioridad, las 3.124 libretas de personas que no contestaron sus LGI no fueron incluidas en la imputación de gastos individuales.

Tabla 2: Detalle de LGI que debían ser contestadas

Códigos de Entrega	LGI No Contestadas		LGI Contestadas*		Total
81. Entrega de forma personal.	518	3,7%	13.624	96,3%	14.142
82. Entrega de forma indirecta (a través de un miembro del hogar).	2.606	21,2%	9.694	78,8%	12.300
TOTAL	**3.124**	**11,8%**	**23.318**	**88,2%**	**26.442**

*Se consideran contestadas aquellas LGI que tengan al menos un día de registro.
Fuente: VII EPF

Cabe hacer la distinción entre las personas que recibieron directamente la LGI y aquellas personas que la recibieron de forma indirecta a través de un miembro del hogar. De las 14.142 personas que recibieron de forma personal la LGI, 518 (3,7%) no contestaron la libreta, mientras que de las 12.300 personas que recibieron la LGI de forma indirecta, 2.606 (21,2%) no contestaron dicha libreta. Esto pone de manifiesto la importancia de hacer a los miembros del hogar partícipes del estudio intentando entregarles la LGI de forma personal para explicar de forma directa la dinámica e importancia del llenado de la libreta, además de crear el compromiso activo de participación en la encuesta.

La Tabla 3 muestra la desagregación de la categoría "LGI contestadas" entre completas y parciales, esta distinción es relevante pues en la tabla se puede apreciar

Tabla 3: Detalle de LGI que debían ser contestadas (incluye desagregación de LGI contestadas)

Códigos de Entrega	LGI No Contestadas		LGI Contestadas				Total
			Completas		Parciales*		
81. Entrega de forma personal.	518	3,7%	10.948	77,4%	2.676	18,9%	14.142
82. Entrega de forma indirecta (a través de un miembro del hogar).	2.606	21,2%	7.099	57,7%	2.595	21,1%	12.300
TOTAL	**3.124**	**11,8%**	**18.047**	**68,3%**	**5.271**	**19,9%**	**26.442**

*Se consideran parciales aquellas LGI que tengan al menos un día de registro.
Fuente: VII EPF

22 Respecto a los requisitos para que un hogar cumpla con los mínimos de calidad exigidos por la grilla técnica, al menos uno de los informantes del hogar debía tener cuatro días de registro en su libreta de gastos individuales siendo al menos uno de estos días fin de semana.
23 Para ver el módulo de control de LGI de la Hoja de Ruta, ver Anexo A

que la mayoría de los informantes que reciben la libreta de forma personal contestan sus libretas de forma completa (77,4%), mientras que más de la mitad de los informantes que reciben la libreta de forma indirecta contestan también sus libretas de forma completa (57,7%).

Las tablas y análisis realizados de aquí en adelante son hechos considerando sólo la categoría de LGI contestadas, ya que sobre dichas libretas se realizarán los procesos de imputación, excluyendo así la categoría de LGI no contestadas en los análisis posteriores.

Las Libretas de Gastos Individuales del administrador de gastos del hogar[24] son fundamentales para el desarrollo del estudio ya que estas personas son las que realizan las compras del día a día al interior del hogar, entendidas éstas como las compras básicas de alimento, vestuario, calzado, electrodomésticos, bebidas, etc. La identificación del administrador de gastos del hogar permite realizar un seguimiento específico de la LGI de esa persona, asegurando de esa manera la mejor captura de los gastos individuales. Además, el gasto promedio muestral individual de los administradores de gastos del hogar representa aproximadamente 2,02 veces el gasto individual del resto de los informantes de la encuesta.

Del total de administradores de gastos del hogar, el 81,0% contestó la LGI de forma completa. Cabe mencionar que al interior de un hogar puede existir más de un administrador de gastos. Este porcentaje de LGI completadas de forma completa del administrador de gastos del hogar es notablemente más elevado que el porcentaje de respuestas de LGI completas del resto de los integrantes del hogar, el que alcanza un 73,7%.

La Tabla 5 muestra la misma información que la Tabla 4 desagregada por sexo del informante. Se puede apreciar la diferencia entre la cantidad de admi-

Tabla 4: Detalle de respuesta del administrador de gastos del hogar

	No contestadas	%	Completas	%	Parciales*	%	Total
Administrador de Gastos del Hogar	427	3,5%	9.535	78,2%	2.232	18,3%	12.194
Resto de los integrantes del hogar que deben contestar LGI	2.697	18,9%	8.512	59,7%	3.039	21,3%	14.248
Total	**3.124**	**11,8%**	**18.047**	**68,3%**	**5.271**	**19,9%**	**26.442**

*Se consideran contestadas aquellas LGI que tengan al menos un día de registro.
Fuente: VII EPF

Tabla 5: Detalle de respuesta del administrador de gastos del hogar por sexo

	Hombres							Mujeres						
	No contestadas	%	Completas	%	Parciales*	%	Total	No contestadas	%	Completas	%	Parciales*	%	Total
Administrador de Gastos del Hogar	209	6,0%	2.571	74,1%	691	19,9%	3.471	218	2,5%	6.964	79,8%	1.541	17,7%	8.723
Resto de los integrantes del hogar que deben contestar LGI	1.750	20,8%	4.865	57,8%	1.807	21,5%	8.422	947	16,3%	3.647	62,6%	1.232	21,1%	5.826
Total	**1.959**	**16,5%**	**7.436**	**62,5%**	**2.498**	**21,0%**	**11.893**	**1.165**	**8,0%**	**10.611**	**72,9%**	**2.773**	**19,1%**	**14.549**

*Se consideran contestadas aquellas LGI que tengan al menos un día de registro.
Fuente: VII EPF

24 El administrador de gastos del hogar es quien (es) regularmente realiza (n) las compras del hogar, evaluado en los últimos seis meses, independiente de quién las pague o entregue el dinero para hacerlo.

nistradores de gastos del hogar de sexo masculino versus la cantidad de administradoras de gastos del hogar de sexo femenino que contestó la LGI. Mientras que 3.262 informantes hombres son administradores de gasto, 8.505 informantes mujeres son administradoras de gasto (existen alrededor de 2,6 veces más administradoras de gasto).

La tasa de LGI contestadas de forma parcial es más baja para las mujeres, tanto para las que son administradoras de gastos del hogar (18,1% versus un 21,2% para los hombres), como para el resto de las integrantes del hogar (25,3% de libretas contestadas de forma parcial versus un 27,1 % de libretas parciales para el resto de los integrantes del hogar de sexo masculino). Es relevante observar que tanto para hombres y mujeres, los administradores de gastos del hogar participan más activamente del estudio que el resto de los integrantes del hogar. Para los hombres, la tasa de respuesta de LGI parciales de los administradores del hogar es de un 21,2% versus un 27,1% del resto de los integrantes hombres. Para las mujeres administradoras de gastos, la tasa de respuesta parcial de LGI es de 18,1% versus el 25,3% del resto de las integrantes del hogar. La tasa de LGI contestadas de forma parcial por los hombres es mayor en 4,4 puntos porcentuales a la tasa de LGI contestadas de forma parcial por las mujeres (25,1% versus 20,7% respectivamente), esta situación deja de manifiesto que una vez que se decide participar en el llenado de las libretas de gastos individuales, las mujeres presentan una mayor disposición a completar el llenado de libretas de gastos individuales que los hombres.

La Tabla 6 muestra la situación de las libretas contestadas de gastos individuales según tramos etarios de cinco años. Las dos situaciones detalladas co-

Tabla 6: Categorias de respuesta de la LGI por tramos etarios (cada 5 años)

EDAD TRAMOS	Situación de LGI				
	LGI Completa	%	LGI Parcial*	%	Total
15-19	1.798	73,3%	655	26,7%	2.453
20-24	1.808	71,9%	708	28,1%	2.516
25-29	1.520	74,3%	526	25,7%	2.046
30-34	1.471	78,4%	405	21,6%	1.876
35-39	1.520	79,1%	401	20,9%	1.921
40-44	1.618	78,7%	439	21,3%	2.057
45-49	1.687	78,0%	477	22,0%	2.164
50-54	1.675	78,7%	454	21,3%	2.129
55-59	1.296	78,9%	346	21,1%	1.642
60-64	1.062	79,6%	272	20,4%	1.334
65-69	927	80,8%	220	19,2%	1.147
70-74	700	83,1%	142	16,9%	842
75-79	471	79,4%	122	20,6%	593
80-84	321	81,7%	72	18,3%	393
85 o más	172	84,3%	32	15,7%	204
No Responde	1	100,0%	0	0,0%	1
Total	**18.047**	**77,4%**	**5.271**	**22,6%**	**23.318**

*Se consideran parciales aquellas LGI que tengan al menos un día de registro.
Fuente: VII EPF

rresponden a “LGI completa” (el informante contestó la totalidad de días de la quincena) y “LGI parcial” (el informante al menos contestó un día de registro de la LGI).

Se puede apreciar en la Tabla 6 que las categorías de respuesta de libretas de gastos individuales varían a través de los distintos grupos etarios que responden la LGI. Los grupos de informantes más jóvenes tienden a presentar mayores niveles de libretas contestadas de forma parcial. El tramo etario de 15 a 19 años tiene una tasa de LGI contestadas de forma parcial del 26,7%, mientras que el tramo etario que le sigue (20-24) presenta una tasa de LGI contestadas de forma parcial de 28,1%, siendo éstos los mayores niveles de respuesta parcial de libretas de gastos individuales. En la tabla se puede observar que a medida que aumenta la edad de los participantes del estudio, la tasa de respuesta de LGI completas aumenta, alcanzando su peak de participación en el grupo de 85 años o más donde la participación en el llenado de LGI completas alcanza un 84,3% de libretas.

La Tabla 7 muestra la desagregación de LGI contestadas por tramos etarios y sexo. En la tabla se observa sistemáticamente que las mujeres independiente del tramo etario al cual pertenezcan, presentan mayor disposición para participar en el estudio a través del llenado total de las libretas de gastos individuales[25].

Tabla 7: Categorías de respuesta de la LGI por sexo y tramos etarios (cada 5 años)

EDAD TRAMOS	Situación de LGI Hombres					Situación de LGI Mujeres				
	LGI Completa	%	LGI Parcial	%	Total	LGI Completa	%	LGI Parcial	%	Total
15-19	816	72,4%	311	27,6%	**1.127**	982	74,1%	344	25,9%	**1.326**
20-24	810	70,3%	342	29,7%	**1.152**	998	73,2%	366	26,8%	**1.364**
25-29	634	70,4%	267	29,6%	**901**	886	77,4%	259	22,6%	**1.145**
30-34	627	77,1%	186	22,9%	**813**	844	79,4%	219	20,6%	**1.063**
35-39	610	75,1%	202	24,9%	**812**	910	82,1%	199	17,9%	**1.109**
40-44	627	76,5%	193	23,5%	**820**	991	80,1%	246	19,9%	**1.237**
45-49	633	73,5%	228	26,5%	**861**	1.054	80,9%	249	19,1%	**1.303**
50-54	661	75,3%	217	24,7%	**878**	1.014	81,1%	237	18,9%	**1.251**
55-59	531	76,1%	167	23,9%	**698**	765	81,0%	179	19,0%	**944**
60-64	411	76,0%	130	24,0%	**541**	651	82,1%	142	17,9%	**793**
65-69	413	79,9%	104	20,1%	**517**	514	81,6%	116	18,4%	**630**
70-74	296	84,6%	54	15,4%	**350**	404	82,1%	88	17,9%	**492**
75-79	196	77,5%	57	22,5%	**253**	275	80,9%	65	19,1%	**340**
80-84	107	78,1%	30	21,9%	**137**	214	83,6%	42	16,4%	**256**
85 o más	64	86,5%	10	13,5%	**74**	108	83,1%	22	16,9%	**130**
No Responde	-	-	-	-	**-**	1	100%	0	0,0%	**1**
Total	**7.436**	**74,9%**	**2.498**	**25,1%**	**9.934**	**10.611**	**79,3%**	**2.773**	**20,7%**	**13.384**

*Se consideran parciales aquellas LGI que tengan al menos un día de registro.
Fuente: VII EPF

25 Esta situación sólo se revierte en los tramos etarios de 70-74 años y de 85 o más años.

Tabla 8: Categorías de respuesta de la LGI por condición de actividad económica

Condición de Actividad Económica	LGI Completa	%	LGI Parcial*	%	Total
No responde	1	33,3%	2	66,7%	**3**
Trabaja durante la semana de referencia	10.028	76,0%	3.173	24,0%	**13.201**
Desocupado o Inactivo durante la semana de referencia	8.018	79,3%	2.096	20,7%	**10.114**
Total	**18.047**	**77,4%**	**5.271**	**22,6%**	**23.318**

*Se consideran parciales aquellas LGI que tengan al menos un día de registro.
Fuente: VII EPF

La Tabla 8 muestra la relación entre respuestas de la LGI y las categorías de la condición de actividad económica. A nivel general, se aprecia una pequeña diferencia en la tasa de respuesta de libretas de gastos individuales de forma parcial entre las personas que en el período de referencia se encontraban trabajando y aquellas que estaban desocupadas o inactivas. Mientras un 24,0% de las personas que declaran trabajar durante la semana de referencia contestan la LGI de forma parcial, un 20,7% de las personas que declaran ser desocupadas o inactivas durante la semana de referencia completaron la LGI de forma parcial.

Tabla 9: Categoría de respuesta de la LGI por sexo y condición de actividad económica

	Hombres					Mujeres				
Condición de Actividad Económica	LGI Completa	%	LGI Parcial*	%	Total	LGI Completa	%	LGI Parcial*	%	Total
No responde	1	50,0%	1	50,0%	**2**	0	0,0%	1	100%	**1**
Trabaja durante la semana de referencia	5.064	74,0%	1.782	26,0%	**6.846**	4.964	78,1%	1.391	21,9%	**6.355**
Desocupado o Inactivo durante la semana de referencia	2.371	76,8%	715	23,2%	**3.086**	5.647	80,4%	1.381	19,6%	**7.028**
Total	**7.436**	**74,9%**	**2.498**	**25,1%**	**9.934**	**10.611**	**79,3%**	**2.773**	**20,7%**	**13.384**

*Se consideran parciales aquellas LGI que tengan al menos un día de registro.
Fuente: VII EPF

Si se desagrega la información de la Tabla 8 por sexo como se aprecia en la Tabla 9, podemos observar que al comparar nuevamente a niveles generales la condición de actividad económica, se observa para ambos sexos que las personas que declaran haber trabajado durante la semana de referencia tienen una tasa LGI parciales mayor que las que declaran haber estado desocupadas o inactivas durante la semana de referencia. Esta situación se corrobora en términos generales al observar qué ocurre al desagregar la condición de actividad económica por tramos etarios, esta situación se puede observar en la Tabla 10.

Tabla 10: Categorías de respuesta de la LGI por condición de actividad económica y tramos etarios (cada 5 años)

EDAD TRAMOS	Situación de LGI CAE=Trabaja durante la semana de referencia					Situación de LGI CAE=Desocupado o Inactivo durante la semana de referencia				
	LGI Completa	%	LGI Parcial	%	**Total**	LGI Completa	%	LGI Parcial	%	**Total**
15-19	209	64,1%	117	35,9%	**326**	1.589	74,7%	538	25,3%	**2.127**
20-24	773	67,7%	369	32,3%	**1.142**	1.035	75,3%	339	24,7%	**1.374**
25-29	1.040	71,9%	406	28,1%	**1.446**	480	80,0%	120	20,0%	**600**
30-34	1.166	78,1%	327	21,9%	**1.493**	305	79,6%	78	20,4%	**383**
35-39	1.160	77,9%	329	22,1%	**1.489**	360	83,3%	72	16,7%	**432**
40-44	1.226	77,9%	348	22,1%	**1.574**	392	81,2%	91	18,8%	**483**
45-49	1.266	77,7%	363	22,3%	**1.629**	421	78,7%	114	21,3%	**535**
50-54	1.215	77,0%	362	23,0%	**1.577**	460	83,5%	91	16,5%	**551**
55-59	876	77,5%	254	22,5%	**1.130**	419	82,2%	91	17,8%	**510**
60-64	535	77,8%	153	22,2%	**688**	527	81,6%	119	18,4%	**646**
65-69	338	81,1%	79	18,9%	**417**	589	80,7%	141	19,3%	**730**
70-74	149	80,1%	37	19,9%	**186**	551	84,0%	105	16,0%	**656**
75-79	53	72,6%	20	27,4%	**73**	418	80,4%	102	19,6%	**520**
80-84	13	72,2%	5	27,8%	**18**	308	82,1%	67	17,9%	**375**
85 o más	9	69,2%	4	30,8%	**13**	163	85,3%	28	14,7%	**191**
No Responde	-	-	-	-	**-**	1	100%	-	-	**1**
Total	**10.028**	**76,0%**	**3.173**	**24,0%**	**13.201**	**8.018**	**79,3%**	**2.096**	**20,7%**	**10.114**

*Se consideran parciales aquellas LGI que tengan al menos un día de registro.
Nota: Se excluye del cuadro los 3 informantes que no declararon condición de actividad económica.
Fuente: VII EPF

La Tabla 10 muestra que al observar la información antes descrita por tramos etarios, el análisis se mantiene pues en general los informantes que trabajan durante la semana de referencia presentan mayores tasas de respuesta parcial de LGI que el resto de la población y les cuesta más completar las encuestas de gastos individuales (salvo el grupo etario entre los 65-69 años).

En el Anexo D se presenta la misma información que la Tabla 10 desagregada por sexo de los informantes. Al mirar la comparación por condición de actividad económica para los diferentes sexos y tramos etarios, se aprecia que las mujeres que declaran trabajar durante la semana de referencia presentan una mayor proporción de libretas LGI completas en relación con las que se declararon desocupadas o inactivas durante la semana de referencia. Esta información difiere del comportamiento de los hombres, quienes en general presentan mayores tasas de LGI completas en los tramos etarios de menor edad para los que declaran trabajar durante la semana de referencia en comparación con los que se declararon desocupados o inactivos, mientras que en los tramos de mayor edad, la situación tiende a invertirse.

A continuación, se presenta el total de registros de días de los informantes que debían contestar las libretas de gastos individuales. Este análisis es relevante porque permite observar la cantidad y porcentaje de días de registros individuales que serán imputados respecto al total de días de gastos individuales que las libretas debieron haber tenido. Es importante resaltar que si bien la unidad de análisis con que se ha trabajado hasta ahora son las personas,

Tabla 11: Caracterización de los días de respuesta para aquellas libretas de gastos individuales de los hogares que superan grilla de calidad

Gastos Individuales*	Total de Días	Porcentajes por tipo de registro
Días con registro	**330.802**	**93,0%**
días con gasto	230.718	64,8%
días sin gasto	100.084	28,1%
Días sin registro**	**25.034**	**7,0%**
Total esperado de días de registro	**355.836**	**100,0%**

*Libretas de LGI con días con registro>0
**Si no tiene días de registro, se asume que rechazó la libreta
Fuente: VII EPF

la unidad de análisis con que se trabajará para la imputación de LGI parciales son los días de registro para cada informante.

La imputación de gastos individuales se realiza sobre el 7% del total de los días de registros diarios de los informantes. Este 7% corresponde aproximadamente a un 5,9% del total de los días de registro esperado de los hogares[26].

26 Esta cifra aproximada es el producto de 7% (% de días sin registro) * 83,8% (% de libretas de gastos individuales contestadas).

2. Test de aleatoriedad de Little

Para asegurar que la no respuesta parcial de gastos individuales se distribuye de forma aleatoria, se realizó el test de Little con el fin de determinar que los datos que serán imputados no estén sesgados por alguna característica de los informantes. De esta forma el análisis de la no respuesta parcial de libretas de gastos individuales se realizó a nivel de zona y estrato socieconómico de la muestra[27], controlándose por tramos de ingreso autónomo[28] de las personas y por sus niveles educativos. Los análisis de la aleatoriedad de la no respuesta a través del test de Little arrojaron que no se rechaza la hipótesis nula de que la no respuesta se distribuye de forma aleatoria.

La variable testeada de gasto toma el valor del gasto de la persona si responde todos los días, mientras que si tiene al menos un día sin contestar toma el valor missing. Esta variable se testeó a nivel de personas mayores de quince años que respondieron al menos un día de la libreta de gastos individuales.

Tabla 12: Test de aleatoriedad de Little para la no respuesta parcial de la libreta de gastos individuales

Gasto Individual	Estrato Económico								
	Bajo			Medio			Alto		
	Nacional	RM	RR	Nacional	RM	RR	Nacional	RM	RR
Chi Cuadrado	0,48	1,35	1,71	1,49	3,47	6,34	1,36	1,55	2,83
Valor P	0,79	0,51	0,43	0,83	0,48	0,18	0,85	0,46	0,59

Fuente: VII EPF

La Tabla 12 muestra los resultados de la prueba de Little para testear la aleatoriedad de la no respuesta para el gasto diario de las personas. Dado que los valores p-value del test son mayores a 0,05, tanto a nivel de zonas como a nivel de agregado nacional para los diferentes estratos económicos, a un nivel de confianza del 0,95%, no podemos rechazar la hipótesis nula de aleatoriedad de la no respuesta asumida por el test de little, lo cual nos permite aplicar distintos métodos de imputación con mayor libertad.

Cabe mencionar que el gasto se comporta de forma no aleatoria por otras variables testeadas como sexo, edad, administrador de gastos del hogar, entre otras. Estas variables también son consideradas a la hora de formar clusters de imputación, ya que son variables que se mantienen más rígidas en la matriz de imputación Hot-Deck, puesto que en dicho método se busca crear cluster de informantes con características lo más parecidas posible desde el punto de vista de la explicación del gasto.

27 El estrato económico toma valores 1, 2 y 3. Corresponde a una variable de post- estratificación asignada a las viviendas del marco a partir de información de Censo 2002 (INE Chile, 2003).

28 Cada tramo del ingreso autónomo de las personas de $100.000.

3. Variables correlacionadas con gasto

Para realizar grupos (cluster) de personas que compartían características comunes para la imputación Hot-Deck se tuvo que determinar qué variables se correlacionaban más con el gasto individual registrado en las libretas de gastos individuales. Dado que se busca imputar el gasto individual de los informantes, se estudiaron las correlaciones entre distintas variables sociodemográficas y la variable gasto diario total por persona. Para establecer las correlaciones con el gasto individual, se consideraron sólo aquellos integrantes que contestaron el 100% de los días de registro, ya que este grupo se considera como modelo para determinar las variables más correlacionadas con el gasto (first best).

Para realizar los análisis, la variable gasto diario total por persona se trabajó en niveles absolutos y en logaritmos con el fin de suavizar los efectos generados por los valores extremos. El utilizar la variable en logaritmo suaviza el efecto de los valores extremos, por lo tanto, los modelos en general logran un mejor ajuste cuando la variable gasto diario total se encuentra expresada en dichos términos. Dado que se obtuvieron mejores estimaciones de las correlaciones al utilizar la variable transformada a logaritmo natural, los resultados que se muestran a continuación consideran las correlaciones con la variable de interés expresada en dicha unidad.

Según la teoría, para las correlaciones no existe un valor exacto a partir del cual se pueda definir que existe una correlación lineal entre dos variables. Se debe tener en cuenta la interpretación que se realice sobre los datos estudiados. A continuación, se detalla la interpretación de Bisquerra (1987) de los coeficientes:

Tabla 13: Interpretación de los coeficientes de correlación r (según Bisquerra)

0,80 < r	correlación muy alta
0,60 < r < 0,79	correlación alta
0,40 < r < 0,.59	correlación moderada
0,20 < r < 0,39	correlación baja
r < 0,19	correlación muy baja

Se consideró la mayor cantidad de variables para estudiar las distintas correlaciones con la variable gasto individual de la persona. Las variables consideradas se listan a continuación:

Var. Cuantitativas	Var. Categóricas	
- Gasto individual de la Persona	**Ordinales**	**Nominales**
- Edad	- Estrato Económico (bajo, medio, alto)	- Sexo
- Escolaridad	- Tramos de Ingreso del hogar	- Parentesco
- Número de Personas en el Hogar	- Tramos de Ingreso Autónomo de la Persona	- Estado Civil
		- Sustentador Principal del Hogar
		- Adm. de Gastos del Hogar
		- Ocupación (CIUO)
		- Condición de Actividad Económica (CAE)

Fuente: VII EPF

Dado que se utilizaron variables continuas y discretas (ordinales y nominales) para correlacionar con la variable gasto individual de la persona (que es una variable continua), se utilizaron distintos tipos de correlaciones dependiendo del tipo de variables que se analizaba.

Pese a que la variable ocupación tiene una desagregación a dos dígitos de codificación, de 28 categorías en las que existe un orden establecido a través de las competencias que requiere cada labor realizada, se utiliza la desagregación a un dígito para las correlaciones. Esta variable se trabaja como categórica, dado que no es posible asumir un orden creciente entre los distintos niveles respecto a las correlaciones con el ingreso.

Las correlaciones con las variables continuas son especificadas a continuación:

Tabla 14: Correlaciones de Pearson entre variables de interés (en logaritmos)

Correlaciones	ln Gasto Individual de la Persona	Tramos de Ingreso del hogar	Tramos de Ingreso Autónomo de la Persona	Número de Personas en el Hogar	Edad	Escolaridad	Estrato Económico (bajo, medio, alto)
ln Gasto Mensual de la Persona	1						
Tramos de Ingreso del hogar	0,324	1					
Tramos de Ingreso Autónomo de la Persona	0,395	0,404	1				
Número de Personas en el Hogar	-0,098	-0,141	0,175	1			
Edad	0,213	0,268	-0,050	-0,258	1		
Escolaridad	0,302	0,193	0,315	-0,088	-0,235	1	
Estrato Económico (bajo, medio, alto)	0,206	0,360	0,175	0,007	0,006	0,351	1

La Tabla 14 muestra los coeficientes de correlación para las variables cuantitativas y ordinales antes enumeradas[29]. Se puede observar que la correlación más alta que se da con el gasto individual de la persona es con la variable "tramos de ingreso autónomo de la persona" (0,4). Esta variable es considerada posteriormente en la generación de los cluster, privilegiándose sobre la variable de "tramos de ingreso del hogar". Si bien esta es la correlación más alta de la tabla, es considerada una correlación moderada según la interpretación sugerida por Bisquerra.

Las correlaciones de las variables tienen los signos que teóricamente se esperaría que tengan, ya que a un mayor nivel de ingreso, escolaridad, estrato económico y edad, el ingreso de las personas debe aumentar. La variable "número de personas en el hogar" presenta una correlación muy baja, por lo que su interpretación puede no ser concluyente.

Si bien la correlación con la edad es positiva, es interesante tener en consideración que las correlaciones del gasto individual con los tramos etarios muestra valores negativos para los tramos etarios de menos de 35 años. Esta situación puede ser apreciada en el Anexo C que muestra las correlaciones de cada tramo etario (variable convertida en dummy para cada tramo etario) y el nivel de gasto individual de la personas. Para los tramos de 25 años y más, las correlaciones toman valores positivos hasta los 70 años o más en que vuelven a ser negativas. El anexo presenta los datos para el total de la población que contestó la LGI de manera completa diferenciada por tramos etarios y además por sexo. Esto es avalado por la teoría del ciclo de la vida, en los años de mayor actividad laboral, la edad presenta una correlación positiva con el gasto individual, sin embargo a medida que las personas cumplen una determinada edad, esta correlación vuelve a ser negativa. Por esta misma razón la correlación entre edad y tramos del ingreso autónomo de la persona presenta una correlación negativa (aunque muy baja).

29 Para las variables cuantitativas se utiliza la correlación de Pearson, mientras que para las variables ordinales se utiliza la correlación de Spearman.

Tabla 15: Correlaciones de punto biserial entre variables de interés

Correlación de punto biserial	ln Gasto Individual de la Persona
Sexo	**0,1107**
Parentesco. Jefe de hogar	0,2037
Parentesco. Cónyuge, conviviente o pareja	0,1842
Parentesco.Hijos e hijas de la pareja	-0,3521
Parentesco. Hermano	-0,0335
Parentesco. Padre o madre	-0,0413
Parentesco. Otro pariente	-0,1193
Estado Civil. Casado o conviviente	0,2647
Estado Civil. Soltero	-0,316
Estado Civil. Separado, divorciado y anulado	0,0609
Estado Civil. Viudo	-0,0013
Sustentador Principal del Hogar	0,2017
Adm. de Gastos del Hogar	0,393
CIUO	
Miembros del Poder Ejecutivo y de los Cuerpos Legislativos y Personal Directivo de la A.P. y de Empresas	0,1429
Profesionales, Científicos e Intelectuales	0,1233
Técnicos y Profesionales de Nivel Medio	0,0339
Empleados de Oficina	0,0672
Trabajadores de los Servicios y Vendedores de Comercios y Mercados	0,0005
Agricultores y Trabajadores Calificados Agropecuarios y Pesqueros	0,0143
Oficiales, Operarios y Artesanos de Artes Mecánicas y de Otros Oficios	-0,0267
Operadores de Instalaciones y Máquinas y Montadores	-0,0229
Trabajadores no Calificados	-0,0449
Otros Grupos no Identificados	0,2697
Inactivos, desempleados y no clasificados	-0,2582
CAE	
Desempleado o inactivo	-0,2582
Trabajador Dependiente	0,1995
Trabajador Independiente	0,0853

Fuente: VII EPF

La Tabla 15, muestra los coeficientes de correlaciones del punto biserial para las variables nominales consideradas en el análisis. El sexo es una variable dicotómica que toma el valor 1 si es femenino y 0 si es masculino, se observa que para el gasto individual (contrario a lo que pasa en ingreso como se aprecia en el apartado de correlaciones con el ingreso), existe una correlación positiva entre ser de sexo femenino y el gasto individual.

Otra relación importante que puede ser apreciada en la tabla precedente es la correlación que existe

entre el gasto individual por persona y la variable administrador de gastos del hogar. En los análisis hechos con anterioridad se explicó que la variable administrador de gastos del hogar es fundamental para el estudio y esta idea es reforzada al analizar la correlación con la variable de interés, ya que muestra una alta correlación en comparación al resto de las variables.

Estas relaciones son consideradas en la aplicación de la matriz del Hot-Deck para determinar las variables que no se relajan nunca en la matriz o aquellas que se van relajando de forma paulatina.

Las variables que mejor ayudaban a explicar el gasto individual fueron:

- Sexo
- Estrato económico
- Tramos del ingreso de la persona autónomo (tramos de 100.000) e Ingreso del hogar (tramos de 300.000)
- Nivel educativo de la persona (niveles y en años)
- Edad (en años y por tramos de 5 y 10 años)
- Ocupación, CIUO y CISE
- Administrador de gastos del hogar

Además se agregaron variables de temporalidad (quincena y mes) y desagregación geográfica[30] (manzana, comuna, región y macro-zona) propias del diseño muestral.

B. FASE DE APLICACIÓN: IMPUTACIÓN DE GASTOS

Existe una amplia variedad de métodos estadísticos para el proceso de imputación. Se debe considerar un tope máximo de datos a imputar, ya que se podría desvirtuar la muestra con valores de imputación muy elevados[31]. Estos niveles de imputación deben tenerse presentes al momento de analizar la cantidad de información que se requiera imputar.

La imputación de gastos individuales se realiza para las libretas de aquellos integrantes de quince años y más del hogar que presentan información faltante y además el hogar cumple con los mínimos de calidad de la encuesta[32]. Dado que los procesos probados realizan una imputación al interior del hogar, tal y como se mencionó anteriormente, puede que haya informantes al interior del hogar que requieran que se imputen sus gastos, mientras que otros puede que no necesiten dicha imputación. Esta combinatoria de casos hace que existan diversas combinaciones de imputación para cada informante al interior del hogar.

Ningún método de imputación es mejor per sé, ya que todos dependen de los datos. Dado lo anterior, se probaron distintos métodos que mantienen entre sí el supuesto base que la imputación con la misma persona (o una de características similares) es preferible en la imputación de gastos puesto que toma en cuenta las preferencias del individuo. Con el propósito de probar los resultados y el comportamiento estadístico de distintos métodos de imputación, se testearon tres distintas metodologías de imputación de gastos individuales. A continuación, se explica el detalle de cada una de las metodologías estudiadas.

1. AJUSTE POR FACTOR DE NO RESPUESTA (FNR)

Este procedimiento consiste en ajustar la información declarada por los informantes compensando la subdeclaración de gastos. El ajuste se realiza diferenciando dos categorías de libretas:

- Libretas que poseen un registro de información diaria igual o superior a seis días, donde los gastos se ajustan a la quincena de referencia aplicando un algoritmo de ajuste (Faq) que varía según el número de días con registro. Cabe señalar que se consideran días con registro aquellos en los que efectivamente existe algún gasto declarado o el informante declara no haber realizado gasto (gasto=0).

Si n° de días con registro >=6

$$F_{aq} = \frac{\text{n° de días de la quincena } q_n}{\text{n° de días con registro dentro de la quincena } q_n}$$

30 El detalle de las comunas encuestadas en la VII EPF se encuentra disponible en el Anexo G.

31 Si bien no existen criterios objetivos para establecer un máximo de omisiones que deben aceptarse, Medina et al (2007), pág. 12, establece que en aquellas bases de datos en donde la omisión de variables de interés se sitúe por sobre el 25%, la modelación puede considerarse poco útil desde un punto de vista práctico, sobre todo si los resultados se utilizarán para apoyar el diseño o evaluación de políticas públicas. En el texto se recomienda no imputar datos en situaciones en que la omisión en una o más variables alcance porcentajes superiores al 20%.

32 Para una explicación más detallada sobre los mínimos de calidad de la encuesta y la grilla técnica, consultar página 106 de la Metodología de la encuesta de la VII EPF, disponible en http://www.ine.cl/epf/

- Libretas que poseen un registro de información diaria inferior a seis días, los gastos se ajustan a la quincena de referencia, llevando los días declarados de la primera semana a la siguiente semana, multiplicando el gasto por una constante[33], igual a 2.

$F_{aq} = 2$ Si n° de días con registro <6

Siendo Faq = Factor de ajuste quincenal para aquellas libretas que poseen menor cantidad de registro de gastos individuales.

Tabla 16: Cantidad de días con registro para todos los informantes que contestaron la LGI

Cantidad de días con Registro	Frecuencia absoluta (ni)	Porcentaje (pi)	Porcentaje acumulado (Pi)	Porcentaje acumulado descendente
1	206	0,9	0,9	100
2	156	0,7	1,6	99,1
3	139	0,6	2,2	98,5
4	197	0,8	3,0	97,9
5	162	0,7	3,7	97,0
6	186	0,8	4,5	96,3
7	233	1,0	5,5	95,5
8	226	1,0	6,5	94,5
9	245	1,1	7,5	93,6
10	296	1,3	8,8	92,5
11	309	1,3	10,1	91,2
12	478	2,1	12,2	89,9
13	618	2,7	14,8	87,9
14	1.806	7,8	22,5	85,2
15	12.880	55,2	77,8	77,5
16	5.181	22,2	100,0	22,2
Total	**23.318**	**100**		

Fuente: VII EPF

La Tabla 16 muestra la cantidad de días con registro para el total de las personas que contestaron la LGI de forma completa o parcial. Se aprecia que para la mayoría de las personas, el registro de la LGI se concentra por sobre los trece días. Si se agregan los 14, 15 y 16 días de registro, se agrega por sobre el 85% de personas que contestaron la LGI. Al considerar diez o más días de registro, esta estimación sube a más del 92% de los informantes. Esta situación pone en manifiesto que la mayoría de las personas que contesta la LGI entrega información sobre la mayor parte de los días que les son solicitados.

La siguiente tabla sólo considera el número de días con registro de las personas que contestaron la LGI de forma parcial, es decir, las personas sobre las que se realiza la imputación parcial de gastos individuales.

33 Esta constante es arbitraria y se basa en el supuesto de que no es posible ajustar menos de seis días a la quincena, ya que si se se hiciera, la estructura de gastos intramensual de los informantes puede quedar distorsionada. Así, solamente se realiza el supuesto de que en ambas semanas se realizan gastos similares.

Tabla 17: Cantidad de días con registro para los informantes que contestaron la LGI de forma parcial

Cantidad de días con Registro	Frecuencia absoluta (ni)	Porcentaje (pi)	Porcentaje acumulado (Pi)	Porcentaje acumulado descendente
1	206	3,9	3,9	100
2	156	3,0	6,9	96,1
3	139	2,6	9,5	93,1
4	197	3,7	13,2	90,5
5	162	3,1	16,3	86,8
6	186	3,5	19,8	83,7
7	233	4,4	24,3	80,2
8	226	4,3	28,6	75,7
9	245	4,7	33,2	71,5
10	296	5,6	38,8	66,8
11	309	5,9	44,7	61,2
12	478	9,1	53,8	55,3
13	618	11,7	65,5	46,3
14	1.164	22,1	87,6	34,5
15	656	12,5	100,0	12,5
Total	**5.271**	**100**		

Fuente: VII EPF

En la Tabla 17 se aprecia que para las personas que contestan la LGI de forma parcial, la mayoría contesta más de cinco días de registro y tan sólo un 16,3% contestan cinco días o menos, mientras que un 83,7% contestan más de cinco días. A partir de los seis días de registro, comienza a subir la cantidad de informantes que declaran más días en sus LGI, lo que justifica la generación de dos categorías de ajuste diferentes ya que las personas que contestan cinco o menos días son consideradas con una estructura de gastos insuficiente para ajustar dichos gastos a la quincena de levantamiento[34].

De las 5.271 personas que contestaron su LGI de forma parcial, a 1.718 personas (32,6%) de ellas tan sólo les falta contestar un día de registro. A más de la mitad de estos informantes (2.816 personas que representan un 53,4% de este subgrupo de informantes) sólo les faltó entre dos y tres días para completar su LGI, por lo tanto, las personas con sus LGI parciales tienen pocos días sin registro.

34 Considerando que muchos países ajustan los gastos diarios al mes o año de referencia a partir de un periodo de registro de una semana, no es un criterio poco exigente exigir un mínimo de cinco días de registro para llevar los gastos de la semana a la quincena.

Tabla 18: Cantidad de días imputados para todos los informantes que contestaron la LGI

Cantidad de días Ajustados	Frecuencia absoluta (ni)	Porcentaje (pi)	Porcentaje acumulado (Pi)	Porcentaje acumulado descendente
0	18.047	77,4	77,4	100
1	1.924	8,3	85,7	22,6
2	761	3,3	88,9	14,4
3	632	2,7	91,6	11,1
4	543	2,3	94,0	8,4
5	452	1,9	95,9	6,1
6	256	1,1	97,0	4,1
7	227	1,0	98,0	3,0
8	247	1,1	99,0	2,0
9	183	0,8	99,8	1,0
10	46	0,2	100,0	0,2
Total	**23.318**	**100**		

Fuente: VII EPF

La Tabla 18 muestra la cantidad de días ajustados para el total de los informantes que respondieron su LGI. Se aprecia en la tabla que, dado los criterios establecidos, como máximo se imputan 10 días de gasto para un informante. Se aprecia además que para la mayoría de los informantes se imputaron pocos días en el proceso de ajuste.

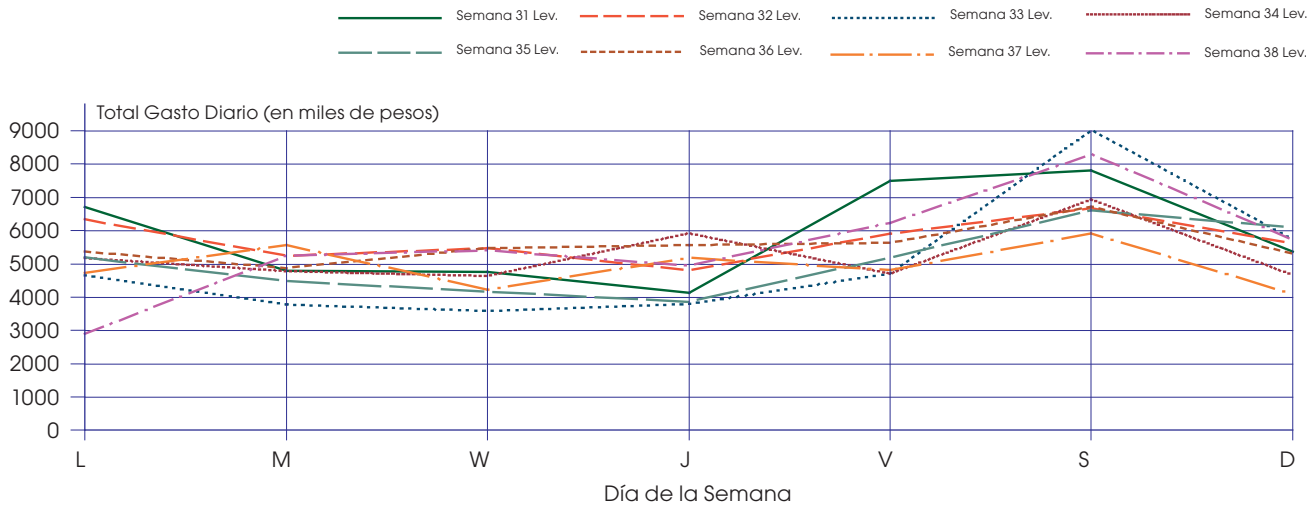
2. Ajuste por peso diario

El método de imputación de la no respuesta parcial a través del ajuste por peso diario corresponde a un método de imputación en el cual se realiza un proceso de imputación en dos etapas. El supuesto subyacente a este modelo es que los días de la semana (lunes a domingo), dependiendo de la distribución de los días calendarios tienen una estructura y peso diferente. El modelo busca agregar el componente de peso diario a cada día faltante de los informantes e incorpora la idea de que, por ejemplo, un día sábado tiene una estructura de gasto distinta a la de un día laboral ordinario.

A modo de ejemplo, el Gráfico 2 presenta una muestra de ocho semanas del levantamiento oficial de la encuesta. Dichas semanas se encuentran entre el día lunes 28 de mayo de 2012 y el día domingo 22 de julio de 2012[35].

35 Semanas seleccionadas al azar.

Gráfico 2: Gasto Diario Según Semanas (desde el 28/05/2012 hasta el 22/07/2012)



FUENTE: VII EPF-INE.

El Gráfico 2 presenta en el eje de las ordenadas el monto de gasto total realizado por todos los informantes para cada día seleccionado, mientras que en el eje de las abscisas se presentan los siete días de la semana. Las series[36] representan desde la semana 31 del levantamiento oficial de la encuesta hasta la semana 38. Se puede apreciar en el gráfico que las semanas tienen un comportamiento que sigue ciertos patrones, por ejemplo el día sábado representa un monto de gasto elevado en comparación con el resto de los días de la semana. Además, el gasto diario total de cada día de la semana depende de otras variables, como por ejemplo el día de pago[37] y la existencia de días feriados.

En el Gráfico 2, el elevado monto de gasto del día viernes de la semana 31 del levantamiento oficial de la encuesta se explica dado a que correspondió al primer viernes del mes coincidiendo con ser el día posterior a la fecha de pago de la mayoría de los informantes. Los días lunes 02/07/2012 y 16/07/2012, correspondientes a las semanas 36 y 38 respectivamente, fueron días festivos en el país. El día lunes de la semana 38 presenta un bajo nivel de gasto, sin embargo el día lunes de la semana 36 no presenta una pronunciada caída del gasto dado que corresponde a un día festivo no muy alejado a la fecha de pago de la mayoría de los informantes (el viernes de la semana 35).

Volviendo al método, el ajuste que se realiza a través del peso del gasto que tiene cada día de la semana en respecto al gasto total. Es importante que sea ajustado mes a mes, ya que el peso diario varía dependiendo de la cantidad de días festivos, fechas de los días de pago, entre otros.

En la primera etapa del método de imputación por peso diario, se determina el monto a imputar según el promedio de gasto diario de la persona ajustado por la ponderación del gasto diario mensual. En la segunda etapa se distribuye el gasto imputado según la estructura de gasto del hogar al cual pertenece la persona.

36 La semana 31 comprende desde el día lunes 28/05/2012 hasta el día domingo 03/06/2012, mientras que la semana 38 comprende desde el día lunes 16/07/2012 hasta el día domingo 22/07/2012.

37 Asociado normalmente al último día hábil de cada mes.

En la primera etapa del proceso se obtiene el peso del día de la semana con respecto al promedio del gasto mensual. A continuación, se describen los pasos a seguir en la primera etapa del proceso. Se comienza calculando el monto promedio total del gasto semanal:

$$\overline{Gto\_D}_{m} = \frac{\sum_{i=1}^{N} Gto\_D_{mti}}{N\_D} \quad (1)$$

donde m=mes.

t=día de la semana.

i= gasto total reportado por el informante i.

N_D= número de días t por cada semana en cada m.

En el primer paso, descrito por la ecuación 1, se obtiene el promedio de gasto total para cada día de la semana. Se obtiene agregando el gasto diario de todos los informantes para cada día de la semana y dividiendo dicha cifra por la cantidad de días en el mes (por ejemplo si hay 4 lunes en el mes, la sumatoria del gasto de todos los informantes en el día lunes se divide por 4. Si hubiesen sido 5 lunes, la sumatoria se divide por 5).

$$\overline{Gto\_total}_{m} = \frac{\sum_{t=1}^{7} \overline{Gto\_D}_{m}}{7} \quad (2)$$

donde $\overline{Gto\_total_m}$ = gasto promedio total diario en el mes m

El segundo paso (ecuación 2) consiste en sumar el gasto diario de cada día de la semana en cada mes y eso dividirlo por 7 para obtener el promedio de gasto total ($\overline{Gto\_total_m}$) para cada mes. Finalmente, para obtener el peso diario de cada día se realiza la siguiente operación:

$$P_{mt} = \frac{\overline{Gto\_D}_{mt}}{\overline{Gto\_total}_{m}} \qquad donde \sum_{t=1}^{7} P_{mt} = 7 \qquad donde\ m = \{1,2,...,12\} \quad (3)$$

donde P_{mt}= ponderador diario del día t en el mes m.

Una vez calculado el gasto agregado diario promedio para cada mes descrito en la ecuación 2, se calcula el ponderador diario para cada mes (P_{mt}). Para ello, se divide para cada uno de los días de la semana el gasto promedio agregado de cada día ($\overline{Gto\_D_{mt}}$) por el gasto promedio diario agregado ($\overline{Gto\_total}$), esto entrega un ponderador diferente para cada uno de los meses el que indica una aproximación al peso estacional que cada día debe tener en cada mes. La sumatoria de los ponderadores de cada día de la semana suma 7. Si el día posee un ponderador mayor a 1, dicho día aporta más que la media al gasto total, mientras que si es menor a 1, dicho día aporta menos que la media al gasto total.

Una vez obtenido el ponderador de cada día de la semana, se calcula el gasto promedio diario para cada uno de los integrantes respecto a los días contestados de la encuesta (no se considera en el cálculo los días sin registro, pero sí los días con gasto igual a cero) según el siguiente detalle:

$$\overline{G_i} = \frac{gasto\ de\ todos\ los\ días\ con\ registro}{número\ de\ días\ con\ registro} \quad (4)$$

donde = gasto promedio diario de los días con registro de la persona i.

Finalmente, cada monto de gasto faltante se obtiene:

$$G_{it} = \overline{G}_i \cdot P_{mt} \quad (5)$$

donde = gasto imputado para la persona i en el día t.

Si a la persona le falta un día martes, se imputa el monto de su gasto promedio multiplicado por el ponderador estacional del día martes correspondiente al mes de levantamiento en el cual fue encuestado el informante.

Una vez imputados los montos de gasto, éstos son desagregados según el peso de cada producto en la estructura de gastos del hogar:

Peso para los productos consumidos por el hogar

G_{it} * Peso en el producto 01.1.1.01.01
Peso en el producto 01.1.1.02.01
Peso en el producto 01.1.1.01.02

3 Imputación Hot-Deck

Al igual que el método de imputación a través del peso diario, el método de imputación por Hot-Deck es un método de imputación en dos etapas. En la pri-

mera etapa se imputa el monto de gasto individual, mientras que en la segunda etapa dicho monto es repartido según la estructura de gasto de la persona o del hogar[38].

El dato faltante del gasto diario de una persona se completa con el gasto del mismo día de otra semana de la misma persona o individuo o dentro de un grupo. La búsqueda de los donantes se basa en una matriz construida según las variables correlacionadas con el gasto individual, considerando además la estacionalidad y la ubicación geográfica. Tomando en cuenta las características de los donantes, los valores faltantes pueden ser imputados con el gasto diario del mismo informante, el gasto de un día de otro individuo o el promedio del gasto de un grupo de donantes que compartan características comunes.

Para encontrar un donante cercano, se establecen variables jerárquicas que van siendo flexibilizadas en caso de no encontrar donantes de la variable a imputar con el fin de encontrar un donante o un grupo de donantes que compartan las características establecidas en la matriz. La lista de variables jerárquicas vinculadas con las variables correlacionadas son las siguientes[39]:

Temporalidad. Quincena, Mes, Día de la Semana.

Sexo

Estrato socioeconómico

Ingreso. Disponible autónomo por persona, disponible del hogar: se construyeron tramos para estas variables. Para el ingreso disponible autónomo por persona se utilizaron 12 tramos cada 100.000 pesos de ingreso autónomo hasta los 600.000 de ingreso y luego los tramos se ampliaron a 200.000, mientras que a nivel de hogar se utilizaron 12 tramos de 300.000, cada uno hasta los 2.400.000 que después se ampliaron a 600.000 cada uno.

Nivel educativo (escolaridad) de la persona y nivel análisis

Edad: Además de la variable edad, se construyen rangos de edad cada cinco años para el primer tipo de tramo (EDAD_R1) y de diez años para el segundo tramo (EDAD_R2).

Ocupación: se utilizó la variable que corresponde a la Clasificación Internacional Uniforme de Ocupaciones de 1988 (CIUO) agregada a un dígito y la Clasificación Internacional de la Situación en el Empleo (CISE) resumida a dos categorías (trabajador dependiente e independiente).

Administrador de gastos del hogar

Espacio geográfico: Manzana, Comuna, Región, Macro-zona.

Identificador de la persona e identificador del hogar.

La matriz para la generación de los distintos *cluster* quedó definida según el siguiente detalle:

De esta forma es posible imputar los montos diarios

Tabla 19: Matriz de exigencia para la elección del vecino cercano para la imputación de gastos individuales

1	2	3	4	5	6	7
Quincena	Quincena	MES	Quincena	Mes		
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico
ING_trP1	ING_trP1	ING_trP1	ING_trP1	ING_trP1	ING_trP1	ING_trP1
Escolaridad	Escolaridad	Escolaridad	Escolaridad	Nivel Análsis (Esc. Niveles)	Nivel Análsis (Esc. Niveles)	Nivel Análsis (Esc. Niveles)
Edad	Edad	Edad	Edad	Edad	Edad(Tramo1. c/5 años)	Edad (Tramo1. c/5 años)
CIUO_N1	CIUO_N1	CIUO_N1	CIUO_N1	CIUO_N1	CIUO_N1	CIUO_N1
Cise_Análisis	Cise_Análisis	Cise_Análisis	Cise_Análisis	Cise_Análisis	Cise_Análisis	Cise_Análisis
RHIP03 (adm. De Gastos)	RHIP03 (adm. De Gastos)	RHIP03 (adm. De Gastos)	RHIP03 (adm. De Gastos)	RHIP03 (adm. De Gastos)	RHIP03 (adm. De Gastos)	RHIP03 (adm. De Gastos)
ID_MANZANA	ID_MANZANA	RPC	DIR_REGION	DIR_REGION	DIR_REGION	Macro
Día de la Semana	Día de la Semana	Día de la Semana	Día de la Semana	Día de la Semana	Día de la Semana	Día de la Semana
Hogar						
Persona						

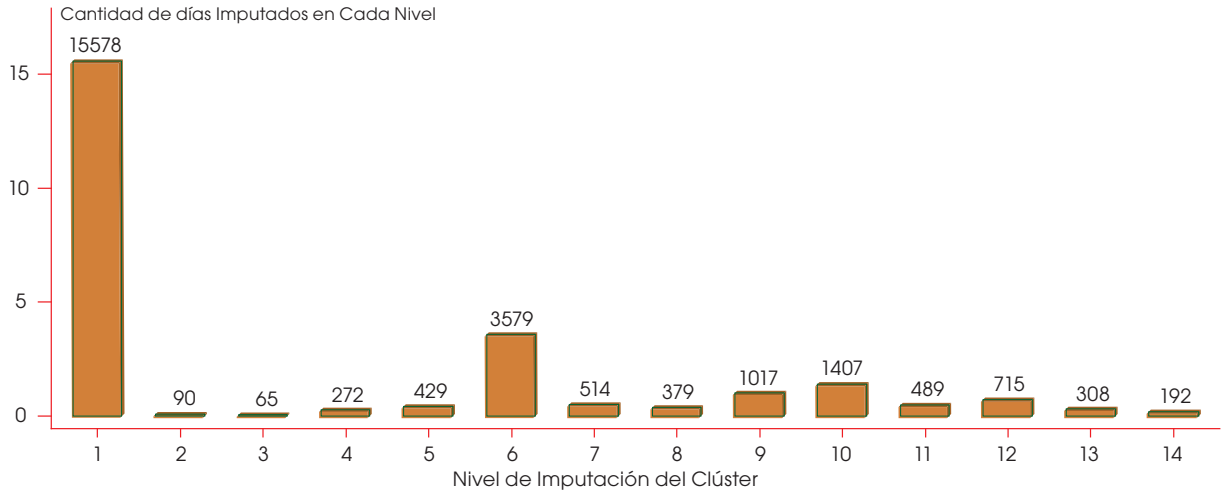
38 Dependiendo de la etapa en que se realice la imputación. Este procedimiento es explicado más adelante.
39 Las definiciones de estas variables se encuentran en el Anexo J.

8	9	10	11	12	13	14
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico		
ING_trP1	ING_trP1	ING_trP1	ING_trP1	ING_trP1	ING_trP1	ING_tr1
Nivel Análsis (Esc. Niveles)	Nivel Análsis (Esc. Niveles)	Nivel Análsis (Esc. Niveles)	Nivel Análsis (Esc. Niveles)	Nivel Análsis (Esc. Niveles)	Nivel Análsis (Esc. Niveles)	
Edad (Tramo1. c/5 años)	Edad (Tramo2. c/10 años)	Edad (Tramo2. c/10 años)	Edad (Tramo2. c/10 años)	Edad (Tramo2. c/10 años)	Edad (Tramo2. c/10 años)	Edad (Tramo2. c/10 años)
CIUO_N1	CIUO_N1					
		Cise_Análisis				
RHIP03 (adm. De Gastos)	RHIP03 (adm. De Gastos)	RHIP03 (adm. De Gastos)	RHIP03 (adm. De Gastos)	RHIP03 (adm. De Gastos)	RHIP03 (adm. De Gastos)	RHIP03 (adm. De Gastos)
Macro	Macro	Macro	Macro			
Día de la Semana	Día de la Semana	Día de la Semana	Día de la Semana	Día de la Semana	Día de la Semana	Día de la Semana

para los informantes que presenten información faltante según el grupo de donantes que comparten características sociodemográficas comunes.

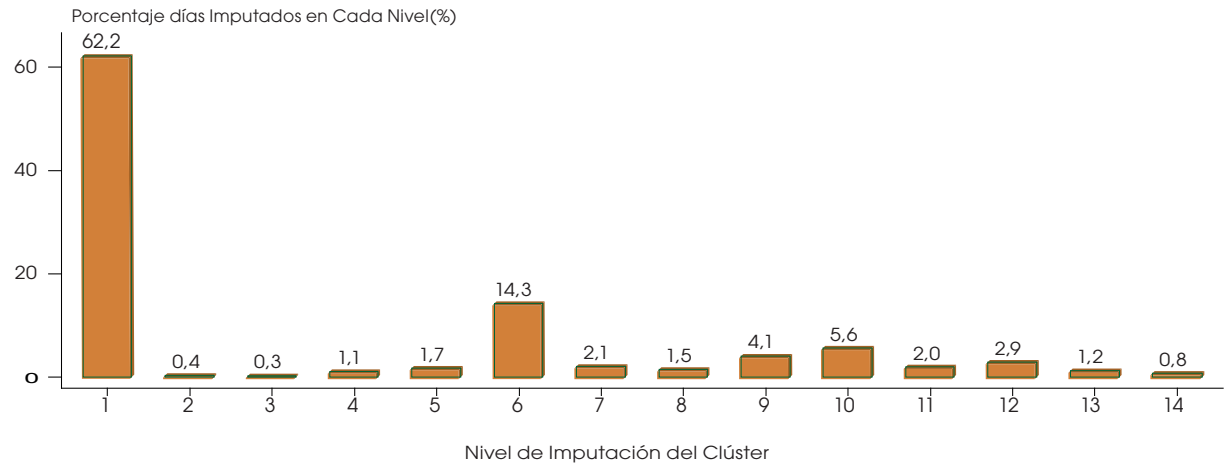
El Gráfico 3 muestra la cantidad total de días imputa-

Gráfico 3: Total de días Imputados en Cada Nivel. Libreta de Gastos Individuales



FUENTE: VII EPF-INE.

Gráfico 4: Porcentaje de días Imputados en Cada Nivel. Libreta de Gastos Individuales



FUENTE: VII EPF-INE.

dos en cada nivel de cluster, mientras que el Gráfico 4 muestra el porcentaje total de días imputados en cada nivel de cluster. Como se puede apreciar, la mayor parte de la imputación se produce en el primer cluster, es decir, más de un 60 % de los días son imputados con el monto de gasto hecho por la misma persona en el día correspondiente, a la otra semana (se imputa un lunes faltante con el lunes de otra semana con registro de la misma persona).

En el segundo nivel de exigencia de la matriz de transición se imputan 90 días correspondientes a un 0,4 % de los días a imputar. La distribución de los distintos niveles de imputación a lo largo de los distintos cluster se observa en los gráficos. Resalta la imputación que se realiza en el nivel 6, la que es explicada en parte por la relajación del criterio de temporalidad (quincena y mes) y en parte por el cambio de la variable edad a tramos.

Dado que en cada grupo se realizan imputaciones de días sin registro según los vecinos cercanos, es importante distinguir la cantidad de donantes que en promedio tiene cada grupo en cada nivel de imputación.

Tabla 20: Cantidad de cluster por cada nivel y número promedio de personas en cada cluster

Nivel	Cantidad de Cluster	Número de Personas promedio en cada Cluster
1	163.226	1
2	162.512	1,004
3	161.595	1,010
4	160.384	1,018
5	155.589	1,049
6	107.674	1,516
7	94.248	1,732
8	90.027	1,813
9	75.866	2,152
10	61.040	2,674
11	47.474	3,438
12	23.842	6,846
13	12.663	12,890
14	2.520	64,772

Fuente: VII EPF

La Tabla 20 muestra la cantidad promedio de personas en cada cluster. Si observamos el primer nivel de cluster formados, podemos apreciar que el número promedio de personas en el cluster es de 1, ya que cada cluster está conformado sólo por un informante que con sus días con registro de una semana puede donar los gastos a sus días sin registro de la otra.

Los primeros niveles de imputación muestran que el promedio de personas es muy cercano a 1. Tan solo en el nivel 9 superamos el promedio de 2 personas por cluster, lo que permite concluir que en general los cluster son pequeños y están integrados por personas que cumplen con características sociodemográficas similares, permitiendo así imputar por la media de gasto. Esta metodología busca reproducir una estructura de gastos para los días faltantes de los informantes de una persona (o grupo de éstas) que comparte la mayor cantidad de características comunes.

Antes de pasar al apartado de análisis de los distintos modelos evaluados en la imputación, cabe mencionar que todos estos modelos son aplicados antes del ajuste de los gastos de la quincena al mes de referencia[40]. Este procedimiento se realiza así porque se requiere primero completar los días sin registro de las libretas parciales de los informantes para luego expandir dicho gasto al mes de referencia.

40 Para una explicación detallada sobre el algoritmo de ajuste temporal, consultar el apartado “ajuste al mes de referencia: factores de ajuste temporal”, página 106 de la Metodología de la encuesta de la VII EPF, disponible en http://www.ine.cl/epf/

C. FASE DE ANÁLISIS: EVALUACIÓN DE LAS METODOLOGÍAS DE IMPUTACIÓN, REGLAS DECISIÓN

A continuación se realizan análisis descriptivos del gasto de los hogares una vez que han sido imputados los días faltantes a través de los distintos métodos de imputación testeados para la no respuesta parcial de LGI. Estos análisis buscan evaluar la distribución de los datos cuando ya han sido imputados respecto al conjunto de informantes que contestaron el 100% de los días, quienes son considerados modelo a seguir de los datos imputados.

En la Tabla 21 se pueden observar estadísticos básicos para el gasto individual muestral obtenido en las libretas de gastos individuales de los distintos individuos participantes en la VII EPF. El gasto promedio registrado en la tabla corresponde al gasto quincenal promedio declarado por los informantes en la LGI. Se agregan además algunos estadísticos básicos respecto a dicho gasto diario.

Tabla 21: Comparación de resultados. Datos muestrales por persona

Tipos de Métodos	Obser.	Promedio	Mínimo	Máximo	SD	CV	p25	p50	p75
Sólo los que contestan 100%	18.047	88.636	0	1.907.164	782	1,186	24.177	57.650	114.609
Ajuste por no respuesta	23.318	88.664	0	2.279.217	709	1,221	23.329	56.488	113.960
Ajuste por Peso diario	23.318	86.086	0	2.127.269	705	1,250	22.230	54.216	110.213
Imputados por Hot-Deck	23.318	89.817	0	2.129.269	705	1,199	24.870	57.876	114.809

Fuente: VII EPF

El número de personas que respondieron el 100% de los días de registro corresponde a 18.047, mientras que 5.271 personas entregan información parcial de la LGI. Si se observa el promedio de gasto individual para cada método, vemos que no existen grandes diferencias en el monto. El promedio del gasto diario que más se aproxima a los informantes que respondieron el 100% de los días es el método de ajuste por no respuesta, mientras que el método Hot-Deck es el que mantiene una estructura más similar al tener un CV más cercano.

Los estadísticos básicos no presentan diferencias entre los distintos métodos de imputación[41], sin embargo también es importante observar cómo se distribuyen los gastos entre los distintos deciles de ingreso y de gasto de la población para determinar si los métodos de imputación modifican la estructura de gastos de los diferentes deciles.

Tabla 22: Comparación de resultados. Datos muestrales decilizados a partir del ingreso disponible del hogar*. Participación en el gasto total de los hogares por grupos de deciles

	1	2	3	4	5	6	7	8	9	10
Sólo los que contestan 100%	2,83	3,92	4,94	5,63	6,77	8,08	9,85	12,05	16,99	28,95
Ajuste por no respuesta	2,73	3,84	4,89	5,63	6,58	8,17	9,76	12,42	17,28	28,69
Ajuste por Peso diario	2,83	3,90	4,93	5,65	6,74	8,02	9,86	12,00	17,05	29,01
Imputados por Hot-Deck	2,83	3,92	4,95	5,63	6,79	8,07	9,85	12,08	16,99	28,89

*grupo de decil de hogares por ingreso disponible del hogar, no incluye arriendo imputado
Fuente: VII EPF

La Tabla 22 muestra la participación en el gasto de cada uno de los deciles de ingresos de la población. Los datos se presentan sin expandir a la población para observar de manera más directa los

41 Esto ratifica la teoría que al tener pequeños porcentajes de imputación de datos (en éste se está imputando un 7% de los días de registro de la LGI), los distintos métodos de imputación, en general, no producen diferencias estadísticas muy significativas.

cambios en la distribución muestral que pueden producir los distintos métodos de imputación. Se aprecia que, en general, los tres métodos de imputación no presentan grandes diferencias en la distribución del gasto a través de los distintos deciles.

Al observar esta distribución, vemos también que el método que más se acerca a la distribución del gasto en los diferentes deciles de los informantes que contestaron el 100% de los días de registros es el método Hot-Deck.

Tabla 23: Comparación de resultados. Datos muestrales decilizados a partir del gasto de consumo final del hogar*. Participación en el gasto total de los hogares por grupos de deciles

	1	2	3	4	5	6	7	8	9	10
Sólo los que contestan 100%	1,41	2,62	3,69	4,74	6,01	7,57	9,62	12,55	17,64	34,14
Ajuste por no respuesta	1,45	2,70	3,76	4,80	6,06	7,56	9,55	12,44	17,41	34,27
Ajuste por Peso diario	1,44	2,68	3,73	4,78	6,03	7,52	9,50	12,42	17,44	34,46
Imputados por Hot-Deck	1,45	2,71	3,76	4,83	6,09	7,59	9,58	12,42	17,37	34,20

*grupo de decil de hogares por gasto total de consumo final del hogar, no incluye arriendo imputado
Fuente: VII EPF

La Tabla 23 muestra la participación en el gasto de cada uno de los deciles de gastos de la población. A diferencia de la Tabla 22, esta tabla muestra la información decilizada por gasto y no por ingreso, es por ello que aumenta la diferencia entre la participación del gasto del último decil en comparación al primer decil. Para construir la tabla, se realizaron las cuatro decilizaciones correspondientes para cada método de imputación de gastos individuales.

En la tabla se puede apreciar que al observar las diferentes participaciones en el gasto total de cada decil, la estructura prácticamente se mantiene por todos los métodos con respecto a la estructura que tiene el grupo de hogares que contestó el 100% de registro de gastos individuales. Dado que se está trabajando sobre el gasto individual, con esta tabla se corrobora la idea de que ningún método de imputación de días faltantes de la libreta de gastos individuales genera grandes diferencias sobre la estructura de la participación de los distintos deciles en el gasto total.

Además de observar la participación de cada uno de los deciles de la población en el gasto total, es importante comparar el impacto que los distintos métodos de imputación pueden tener en las estimaciones finales expandidas al total de la población objetivo de la encuesta sobre el gasto en las distintas divisiones que tienen cada uno de los distintos métodos.

Tabla 24: Comparación de resultados. Datos expandidos. Gasto promedio mensual, según divisiones de CCIF para el total de capitales regionales (Excluye arriendo imputado)

DIVISIÓN	FNR		Peso Diario		Hot Deck	
	Gasto	%	Gasto	%	Gasto	%
	807.409	100	794.600	100	813.884	100
1	150.439	18,63	146.426	18,43	154.127	18,94
2	13.200	1,63	12.706	1,60	13.377	1,64
3	35.412	4,39	34.281	4,31	35.745	4,39
4	108.806	13,48	107.753	13,56	108.362	13,31
5	55.245	6,84	54.742	6,89	55.612	6,83
6	50.657	6,27	50.129	6,31	51.044	6,27
7	132.228	16,38	129.705	16,32	133.021	16,34
8	39.327	4,87	39.344	4,95	39.327	4,83
9	54.522	6,75	53.744	6,76	54.729	6,72
10	63.955	7,92	63.955	8,05	63.955	7,86
11	33.846	4,19	32.988	4,15	34.268	4,21
12	69.772	8,64	68.826	8,66	70.317	8,64

Fuente: VII EPF

Si se observan las diferentes estructuras de gastos obtenidas con los diferentes métodos de imputación, vemos que ella varía muy poco mientras que el monto mensual de gasto de los hogares varía entre 794.600 y 813.884.

Una vez realizados estos análisis, la determinación final del método escogido para la imputación de días faltantes de la libreta de gastos individuales fue el método de ajuste por no respuesta. Esta decisión fue tomada considerando que el método de ajuste por no respuesta es el que menos cambia el promedio de gasto individual respecto a los que contestan el 100% de los días. Además, no presenta grandes diferencias al estudiar los distintos estadísticos revisados ni presenta grandes cambios en la participación de los diferentes deciles en la estructura de gastos. Junto con ser un método fácil de entender, es un método utilizado por otros países para ajustar este tipo de gastos y fue el aplicado en la VI EPF.

A modo de conclusión de los distintos métodos testeados para la imputación de días faltantes de gastos individuales, para el nivel de información faltante en la encuesta los métodos presentaron estimaciones y estructuras similares, por lo tanto la utilización de un método u otro, a este nivel de imputación, reproducen resultados muy similares. Respecto a la imputación de libretas completas de gastos individuales, se deben hacer estudios a futuro para desarrollar metodologías y evaluar el impacto de la aplicación de modelos que imputen libretas completas de gastos individuales, ya que en esta versión de la encuesta se decidió no innovar respecto a este tema específico y no realizar imputación de libretas completas de gastos individuales o gastos en otras libretas.

V INGRESOS DE LA ACTIVIDAD LABORAL PRINCIPAL Y DE JUBILACIONES

El segundo tema abarcado en el presente informe de imputación de la EPF se trata de los ingresos, variables que son de suma importancia para el análisis de los hogares y su bienestar. Como se mencionó anteriormente, se analizaron cuatro métodos para enfrentar la no respuesta en ingresos laborales y de jubilaciones de la VII EPF, con características distintas entre sí tanto respecto al supuesto de aleatoriedad en la no respuesta, como en el modelo y distribución subyacente. No se tomó una decisión a priori respecto al método a utilizar para poder contrastar el comportamiento de cada uno ellos con respecto a los datos reales de la encuesta. Para presentar los resultados y conclusiones se conservó el proceso mostrado en el punto II, Metodología de análisis.

A FASE DE PREPARACIÓN:

En la fase de preparación de los ingresos se proporciona la evidencia estadística y características de la no respuesta en sus diferentes niveles (variable, módulo y libreta) para así demostrar que ésta se distribuye de forma aleatoria, como requisito para ser sujeto de imputación.

1. Características generales de la no respuesta de ingresos laborales y jubilaciones

Para los ingresos del trabajo, el diseño de la libreta de ingresos incluye un trabajo de gabinete que debió realizar el investigador en función de la información registrada en la libreta de registro de personas del hogar (RPH); este flujo permite identificar la fuente de los ingresos del trabajo asociados a los miembros del hogar según la Clasificación Internacional de la Situación en el Empleo (CISE). El diagrama del Anexo E, ayuda a visualizar los saltos en las preguntas[42] y los dos primeros módulos que deberían ir contestados en la libreta de ingresos, de acuerdo a las respuestas entregadas en RPH.

Una primera mirada a los ocupados a la luz del CISE, permite diferenciar a dos subgrupos entre los ocupados, los dependientes y los independientes. Su principal diferencia es la característica de subordinación de los dependientes, de allí su nombre. Tradicionalmente, el clasificador diferencia a los independientes en dos grupos, empleadores y cuenta propias, mientras que a los dependientes se les denominan asalariados[43]. En la Tabla 25 se reportan los ingresos promedios y el coeficiente de variación, considerando los datos observados.

42 Corresponde al diseño de la libreta para levantamiento.
43 Los familiares no remunerados no son considerados en la estadística descriptiva pues, como su denominación lo señala, no reciben remuneración por lo que no se le debe imputar ingresos.

Tabla 25: Detalle de la no respuesta del ingreso según fuente

Categoría en la ocupación principal	N Teórico	% Pob	Promedio Observado	N obs.	CV	Datos faltantes	
						#	%
Ocupados	**15.057**	**100,0%**	**586.677,0**	**13.758**	**1,520**	**1.299**	**8,63%**
- Dependientes	**11.544**	**76,7%**	**618.273,2**	**10.513**	**1,302**	**1.031**	**8,93%**
Asalariados	10.952	72,7%	626.447,0	10.014	1,304	938	8,56%
Honorarios	592	3,9%	454.240,1	499	1,024	93	15,71%
- Independientes	**3.513**	**23,3%**	**484.313,0**	**3.245**	**2,319**	**268**	**7,63%**
Negocios por cuenta propia	2.727	18,1%	410.536,8	2.516	2,463	211	7,74%
Profesionales Independientes	786	5,2%	738.937,0	729	1,917	57	7,25%
Jubilados	**3.331**		**241.735,9**	**3.187**	**1,110**	**144**	**4,32%**

Fuente: VII EPF

En primer lugar, se dimensionó la no respuesta mostrando también los porcentajes desagregados en las cuatro categorías de empleo, además del grupo que percibe ingresos de jubilaciones y/o pensiones de vejez. La tasa de no respuesta para el ingreso del trabajo es del 8,6% (dependientes más independientes) y la mitad 4,3% para los jubilados.

Asimismo, se deberá considerar que la población ocupada con período de referencia esté en concordancia con la libreta de ingresos que muestra que el 72,7% de la población se concentra en la categoría de asalariados. En el opuesto está el grupo de honorarios con apenas el 3,9%. Estos porcentajes, sin embargo, son acompañados de proporciones invertidas en la tasa de no respuesta. Por otro lado, los trabajadores independientes muestran porcentajes similares en la no respuesta, aunque el grupo de quienes trabajan en negocios propios es 3,5 veces más grande que el de profesionales independientes.

La condición de percepción de ingresos por jubilación es independiente de su condicional laboral, es decir, es factible que un perceptor de jubilaciones o pensiones de vejez declare estar activo en la fuerza laboral y perciba un ingreso por ese trabajo. La comparación muestra que cerca del 25% de los perceptores de jubilaciones trabaja y percibe un ingreso por un trabajo que corresponde al período de referencia.

En segundo lugar, se realizó un análisis de la estadística descriptiva disponible para quienes no respondieron la libreta de ingresos versus aquellas personas de las que se tiene información completa. Las diferencias entre los conjuntos con datos faltantes y completos revelan obstáculos en la imputación; sin embargo se debe recordar que la prueba de aleatoriedad en la no respuesta es el paso siguiente para proceder con la imputación, ya que se debe verificar el tipo de mecanismo seguido por la no respuesta.

Tabla 26: Algunas características sociodemográficas de quienes trabajan

Fuente de ingresos laborales	Completos		Incompletos		TOTAL	
	Hombre	Mujeres	Hombre	Mujeres	Hombre	Mujeres
Asalariados	55,62%	44,38%	56,18%	43,82%	55,67%	44,33%
Honorarios	41,08%	58,92%	45,16%	54,84%	41,72%	58,28%
Negocios por cuenta propia	53,70%	46,30%	50,71%	49,29%	53,47%	46,53%
Profesionales independientes	66,94%	33,06%	56,14%	43,86%	66,16%	33,84%
Escolaridad promedio por fuente y sexo de la persona						
Asalariados	12,250	12,700	11,920	11,990	12,230	12,640
Honorarios	14,410	14,340	13,620	13,470	14,280	14,210
Negocios por cuenta propia	10,770	10,590	11,280	10,600	10,810	10,590
Profesionales independientes	13,180	14,460	13,690	13,680	13,220	14,380
Edad promedio por fuente y sexo de la persona						
Asalariados	40,150	40,090	38,520	37,760	40,010	39,890
Honorarios	39,630	37,890	34,950	30,800	38,840	36,850
Negocios por cuenta propia	49,200	46,860	47,300	46,060	49,060	46,800
Profesionales independientes	47,830	44,320	42,090	46,360	47,480	44,520

Fuente: VII EPF

La Tabla 26 debe leerse por fila, pues pretende mostrar la similitud de los grupos con información de ingresos faltante y el grupo que respondió completamente. Por ejemplo, se observa que para los dependientes el porcentaje de mujeres con datos faltantes es menor (para los asalariados es medio punto menor mientras que para los honorarios es menor en cerca de 4pp) en comparación con los independientes, donde las mujeres tienen más presencia en el grupo que no respondió la información de ingresos (hasta 10pp de diferencia para los profesionales independientes).

La variable escolaridad, por otro lado, no presenta casi diferencia, con menos de un año en el promedio. Este equilibrio también se evidencia en la desagregación por sexo. Sin embargo, se puede apreciar que los años promedio de edad varían entre las fuentes, siendo los grupos de Honorarios (3,9% de la masa) y Profesionales (5,2%) los que reportaron mayor escolaridad, aproximadamente 14 en promedio.

La edad, no obstante, muestra una diferencia constante en todas las fuentes, ya que la edad promedio es menor para el grupo que no contestó. Incluyendo al análisis el sexo, esto se mantiene, exceptuando las mujeres profesionales que evidencian una diferencia de dos años superior en el grupo que no respondió el monto de su ingreso.

Tabla 27: Algunas características sociodemográficas de quienes perciben jubilaciones

Jubilados	PROMEDIO		
	Completos	Incompletos	TOTAL
MUJERES	50,20%	50,00%	50,20%
ESCOLARIDAD	9,86	9,64	9,85
EDAD	71,32	72,44	71,37
Fuente: VII EPF			

Para el grupo de quienes reciben ingresos por jubilación, las características de quienes responden el monto que perciben y a quienes les falta el dato son muy similares, además tomando en cuenta que sólo el

4,25% de este grupo no respondió el monto de la jubilación, se considera que existe información suficiente y de buena calidad para imputar.

2. Test de aleatoriedad de Little

Para continuar el análisis de no respuesta y de aleatoriedad de la misma, se recurre al test de Little, cuya hipótesis nula es que la "no respuesta" se distribuye de forma aleatoria. La aplicación del test en los datos de ingresos de la VII EPF proporcionaron los siguientes valores p para las desagregaciones elegidas (Tabla 28). La prueba se realizó por estrato económico del diseño de la muestra y zona geográfica, controlando por sexo, y para los ingresos del trabajo la clasificación de la ocupación (CIUO88).

Tabla 28: Prueba de Little para aleatoriedad en la no respuesta

Fuente de Ingresos del Trabajo	Nacional	Estrato Económico					
		Bajo		Medio		Alto	
		RM	RR	RM	RR	RM	RR
Dependientes	**0,20**	**1,00**	**0,94**	**0,39**	**0,42**	**0,38**	**0,42**
Asalariados	0,36	0,90	0,97	0,33	0,44	0,65	0,47
Honorarios	0,00	0,03	0,54	0,24	0,01	0,05	0,82
Independientes	**0,31**	**0,05**	**0,12**	**0,43**	**0,38**	**0,92**	**0,32**
Negocios por cuenta Propia	0,66	0,08	0,51	0,83	0,77	0,76	0,25
Profesionales Independientes	0,16	0,60	0,02	0,04	0,43	0,76	0,40
Jubilados	0,96	0,41	0,06	0,71	0,65	0,55	0,39

Fuente: VII EPF

La lectura de la prueba se relaciona con el mecanismo de no respuesta. Si la probabilidad de que se registre el ingreso depende del valor de los ingresos de cada categoría ocupacional y del sexo de la persona (por ejemplo, las personas con ingresos elevados no los declara), los datos no serán ni MCAR ni MAR. Si se produce esta situación, no hay ningún método adecuado para imputar. Algunos métodos de imputación, dependiendo de los porcentajes de no respuesta, pueden llevar a resultados insesgados, tales como la imputación múltiple y los métodos de máxima verosimilitud (Schafer & Graham, 2002).

La lectura de la Tabla 28 se complementa con la Tabla 26 donde se observa que el grupo de profesionales independientes es el que tiene mayor diferencia entre los porcentajes de mujeres (aproximadamente 10%), no obstante tienen un porcentaje cercano al 7% de no respuesta y el valor p del test de aleatoriedad supera en casi todas las categorías el 0.01, por lo que no se rechaza la hipótesis de aleatoriedad con el 99% de confianza[44]. Diferente es la situación, por ejemplo, del grupo de honorarios, cuyo conjunto con datos completos muestra una diferencia de aproximadamente 6 años en las edades. Esto, acompañado de su tasa de no respuesta del 15,7%, afecta el resultado de la prueba Little, por lo tanto, se recomienda evaluar a los honorarios como parte del grupo asalariados ya que mantienen el mismo tipo de

44 En general las pruebas de hipótesis conllevan un valor p de comparación que se denomina α, el que se asocia comúnmente con 0,05 o 0,01. Este indicador de la regla de decisión caracteriza la probabilidad de tomar una decisión equívoca, con respecto a rechazar una hipótesis cuando ésta debió ser aceptada. Sin embargo, también existe la posibilidad de cometer el denominado error tipo 2, que es aceptar una hipótesis que se debió rechazar; por eso la conclusión es que no se rechaza la hipótesis de la prueba y no se dice que se acepta.

relación laboral, subordinación y dependencia con su empleador.

Las agregaciones de las fuentes de ingreso por dependientes e independientes tienen resultados más robustos con referencia a las pruebas de aleatoriedad de la no respuesta, ya que aglutinan grupos más grandes y mantienen al interior características de su situación laboral referidas a la subordinación. Bajo esta mirada, analizando en la desagregación geográfica bajo la cual la muestra es representativa (Región Metropolitana y resto de regiones) y considerando el estrato económico usado en el diseño muestral, se obtienen valores p más altos. Esto fue utilizado posteriormente para guiar la forma en cómo se plantean los métodos de imputación. Para los dependientes, el valor p casi en todas las desagregaciones supera el 0,4 y aun viendo la totalidad geográfica, el valor p es 0,2 es decir, con amplio margen no se rechaza la aleatoriedad. Para el caso de los independientes, el comportamiento de los resultados está en el otro sentido, ya que en las desagregaciones no se rechaza la hipótesis con menor margen que en el nivel nacional, es decir, en las desagregaciones en el estrato más bajo el valor es más próximo al 0,05 mientras que el nacional bordea el 0,3.

Para el caso de los ingresos de las jubilaciones y pensiones de vejez, los resultados de las pruebas de aleatoriedad fueron más estables en las desagregaciones y en el total nacional.

3. Variables correlacionadas con el ingreso laboral y las jubilaciones

La teoría económica sobre formación del capital humano y la teoría el ciclo de vida, revelan relaciones de los ingresos laborales con características demográficas de las personas. Por ejemplo, se espera que el ingreso aumente con la edad, sin embargo llegará un punto de inflexión que hará que disminuya el promedio de ingresos para los grupos de mayor edad. También que el ingreso aumente con el nivel de educación, siendo mayor para personas que posean estudios universitarios y de postgrado, en comparación con los que tienen un menor nivel de estudios. Es así que en esta sección mostramos primero las correlaciones de las variables demográficas de características de su ocupación.

Se debe recordar que un coeficiente de correlación nos proporciona tres características principales: la existencia o no de una relación entre las variables estudiadas, la dirección de la relación y el grado de esta relación.

Las correlaciones entre variables continuas y categóricas se evalúan de diferente forma. Debido a que la correlación es una relación lineal, que podría considerar que una variable categórica ordenada es una variable continua de números enteros. La correlación entre una variable continua y una discreta se calcula con la fórmula punto biserial, mientras que la correlación entre dos variables dicotómicas se calcula con el coeficiente phi.

Dentro de las variables, es de especial interés cómo entendemos la ocupación, pues cuando está desagregada tiene 28 categorías, pudiendo tomarla entonces como una variable continua ya que tiene un orden (OIT, 1991). Este orden se basa en dos conceptos: empleos o tareas cumplidas y competencias o conocimientos (formales y/o basados en experiencia). Sin embargo, al considerar esta variable como continua se supondría que la distancia entre una ocupación y otra es la misma, por lo que se prefiere utilizar la desagregación a 1 dígito para las regresiones y mantener su condición de categórica.

Dicotómicas	Continuas	Categóricas
• Sexo	• Escolaridad	• Desagregación geográfica (Macrozona, Región)
• Estrato socioeconómico	• Edad	• Ocupación (CIUO)*
• Desagregación geográfica(Zona)		

En todos los casos la lectura es semejante[45], es decir, en valor absoluto mientras más alto el escalar más fuerte es la relación, y el signo revelará la dirección de ésta. La determinación de cortes de valor del coeficiente para designar relaciones altas, moderadas o bajas se encuentra en la Tabla 13, de la página 31.

45 El valor máximo en la correlación biserial es 0,7879.

Tabla 29: Correlación entre variables de interés

Correlación	Ingreso laboral (ln)	Edad	Escolaridad	Sexo	Zona
Ingreso laboral (ln)	1				
Edad	0,0308	1			
Escolaridad	0,5028	-0,2087	1		
Sexo	-0,2223	-0,0296	0,0449	1	
Zona	0,0238	0,0159	0,0453	0,0115	1

Fuente: VII EPF

La matriz expuesta en la Tabla 29 muestra los coeficientes de correlación de acuerdo a las consideraciones previamente explicadas. La correlación más alta se da entre el ingreso laboral y la escolaridad (0,5) y es positiva, tal como se espera. La variable sexo es dicotómica y toma el valor igual a 1 cuando se trata de una mujer, por eso la correlación negativa es acorde a la evidencia sobre discriminación salarial. La variable zona toma el valor 1 para la Región Metropolitana y su magnitud es acorde al test de aleatoriedad.

El coeficiente de la edad muestra una asociación baja (0,03) con el ingreso pero con el signo correcto. La magnitud de esta relación probablemente tiene que ver con la correlación entre la edad y la escolaridad, que es más alta (-0,2), aunque con el signo no esperado. Para analizar este caso, se calcularon las correlaciones por intervalos de edad y se determinó que el motivo del signo entre 25 y 44 años la relación es positiva; sin embargo pasada la edad de 44 la correlación es negativa. Este cambio de signo en la correlación a medida que la edad aumenta nos indica un cambio en los rendimientos educacionales, razón por la que en las regresiones se incluyen las series de edad y educación elevadas al cuadrado.

Tabla 30: Correlación entre la ocupación e ingreso

CIUO \ Ingreso Laboral	Correlación	%	N
1.- Miembros del poder ejecutivo, legislativo y personal directivo.	0,2714	3,90	536
2.- Porfesionales científicos e intelectuales.	0,4073	14,67	2018
3.- Técnicos y profesionales de nivel medio.	0,1312	12,33	1696
4.- Empleados de oficina.	0,0137	9,91	1363
5.- Trabajadores de los servicios y vendedores.	-0,1692	13,42	1846
6.- Agricultores y trabajadores calificados	-0,0502	1,05	144
7.- Oficiales, operarios y artesados de artes mecánicas y otros.	-0,1085	13,32	1833
8.- Operadores de instalaciones y máquinas	0,0354	7,81	1075
9.- Trabajadores no calificados.	-0,3861	22,76	3132
10.- Otros grupos no identificados.	0,0899	0,84	115

Fuente: VII EPF

La Tabla 30 muestra los coeficientes del punto biserial de correlación para los grandes grupos y la distribución de la población ocupada. Como se observa, la correlación más fuerte y con signo positivo se da en los grupos más calificados.

B. FASE DE APLICACIÓN: IMPUTACIÓN DE INGRESOS

En la fase de aplicación se prueban distintos métodos de imputación de forma paralela y, dependiendo del módulo en la Libreta de Ingresos, se seleccionan las variables a considerar. Los métodos a evaluar son: Hot-Deck, ecuaciones de Mincer (con corrección de sesgo de selección Heckman), imputación múltiple y de máxima verosimilitud con EM.

Ingresos del trabajo principal: Los ingresos que están sujetos a imputación son aquellos que corresponden al período de referencia. No se incluyen a las personas auto clasificadas como personal no remunerado y que hayan rechazado responder la libreta de ingresos, así como a las que confirmaron el no recibir ingresos (ni en especie) por sus tareas laborales. Las fuentes laborales cuyo período de referencia sea distinto, quedan registradas en otra partida, ya sea de un trabajo asalariado o independiente, y son denominadas ingreso bruto por otros trabajos asalariados o ingreso bruto por otros trabajos independientes.

Ingresos por jubilaciones: Se imputa a quienes declararon percibir una jubilación, independientemente si están trabajando. Las variables son restringidas, pero la evidencia empírica en Chile y otros países muestra un mejor ajuste por menor dispersión en los datos. Se cuentan con suficientes variables sociodemográficas que ayudan a realizar una adecuada imputación.

A continuación, se presenta la adaptación de cada método de imputación para los datos del proyecto VII EPF.

1. Imputación Hot-Deck

El proceso de esta metodología es simple. El dato faltante de una persona se completa con la información de otro individuo o de un grupo, con los cuales se comparta un set determinado a priori de características. En el caso de que haya un grupo de donantes, se imputa el promedio de la variable de este grupo[46].

La búsqueda de los donantes se basa en una matriz construida para cada fuente de ingreso laboral. Al elaborar estos niveles se exige un grado de cercanía o similitud en función de un set de variables, llamadas también jerárquicas, las que, en caso de no encontrar un donante, se flexibilizan hasta encontrar un donante con algún grado de compatibilidad.

El procedimiento señala que se debe comenzar con la elección de las variables y con la determinación del rango que limita cada categoría. Éstas son[47]:

- **Sexo.**
- **Edad:** Se construyen además rangos cada 5 años para el primer tipo de rango (EDAD_R1): 15 a 24, 25 a 34, 35 a 44, 45 a 54 y más de 55. El segundo rango es cada 10 años (EDAD_R2) y finalmente otro que tiene (EDAD_R3): 18 a 24, 25 a 54 y más de 55 años.
- **Nivel educativo (escolaridad) de la persona:** Se calcula la escolaridad a partir de las preguntas RHED01, RHED02 (curso y nivel) y RHED03 (concluyó el nivel) que pertenecen a la RPH. Luego se construyen rangos: 0 años (sin educación formal), 1 a 8 (primaria incompleta), primaria completa (nivel terminado), de 9 a 12 (secundaria incompleta), Bachiller (nivel terminado), 13 a 16 (estudios superiores), universitario (nivel completo)
- **Ocupación:** Codificados a 2 y 1 dígito de la CIUO.
- **Estrato socioeconómico.**
- **Espacio geográfico**: Manzana, Comuna, Región y Macrozona.

Es así que la definición de vecino cercano depende por categorías de la situación en el empleo, ya que se dispone de diferente información. El mecanismo para ir flexibilizando los criterios sigue la teoría de capital humano y de la ecuación de Mincer, siendo entonces las variables más importantes el sexo, la escolaridad y la edad; por lo que si bien las categorías son más amplias en los últimos niveles de cada fuente, su restricción en la definición de vecino cercano no se anula completamente. Como segundo criterio de flexibilización en los niveles está el anhelo de mantener *cluster* de tamaño controlado.

Por ejemplo, la siguiente matriz muestra 14 niveles de exigencia para la definición de vecino cercano (Tabla 31). Se puede observar que los cambios entre niveles primero sacrifican la cercanía geográfica, partiendo de la más pequeña, que es la manzana hacia la Macrozona. A continuación, las variables continuas toman valores de categoría y representan rangos, lo que reduce la cantidad de cluster formados y amplía la cantidad de vecinos por *cluster*.

46 El proyecto VII EPF seleccionó el promedio del grupo de donantes con las mismas características, otras encuestas en el INE, como la NESI utiliza la mediana como valor a imputar. Cabe destacar, que durante el proceso de construcción de los criterios se analizó el tamaño de los clúster o grupos formados, los que resultaron muy pequeños por el tamaño de la muestra, proporcionando así resultados muy similares utilizando ambos estadísticos (media y mediana).

47 Definición de variables ver en el Anexo J.

Tabla 31: Matriz de exigencia para el vecino de asalariados

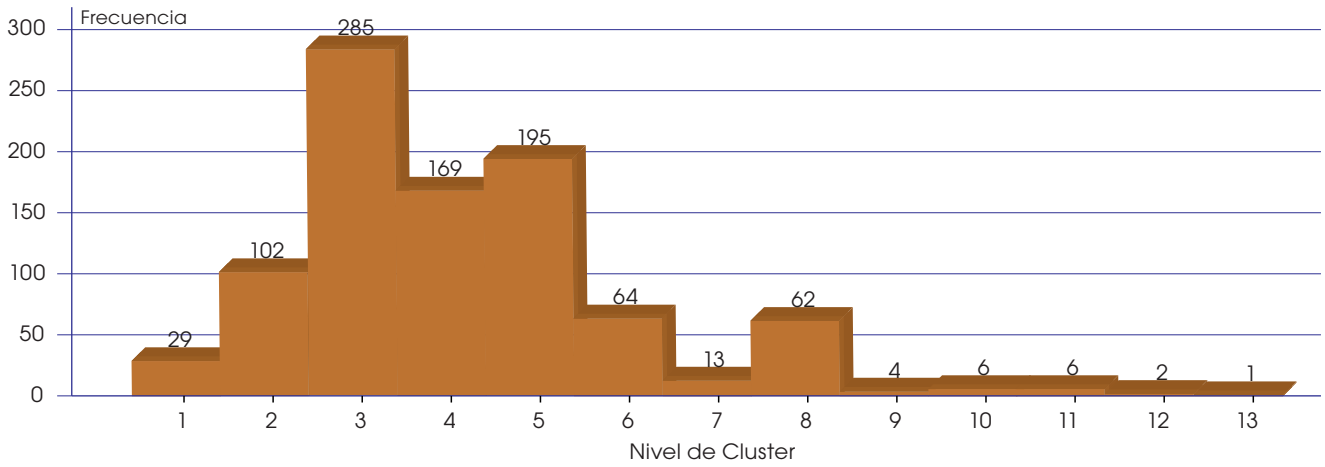
1	2	3	4	5	6	7
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico
Escolaridad	Escolaridad	Escolaridad	Rango 1 Escolaridad	Rango 1 Escolaridad	Rango 1 Escolaridad	Rango 1 Escolaridad
Edad R1	Edad R1	Edad R1	Edad R1	Edad R2	Edad R2	Edad R2
CISE desagregado	CISE desagregado	CISE desagregado	Privado vs Publico	Privado vs Publico	Privado vs Publico	Privado vs Publico
CIUO (2d)	CIUO (2d)	CIUO (2d)	CIUO (2d)	CIUO (1d)	CIUO (2d)	CIUO (1d)
Manzana	Comuna	Región	Región	Región	Macrozona	Macrozona

8	9	10	11	12	13	14
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo	Sexo
Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	-	-	-
Rango 1 Escolaridad	Rango 1 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad
Edad R2	Edad R2	Edad R3	Edad R3	Edad R3	Edad R3	Edad R3
CISE desagregado	Privado vs Publico	Privado vs Publico	-	Privado vs Publico	-	-
-	-	-	CIUO (1d)	-	CIUO (1d)	-
Macrozona	Macrozona	Macrozona	-	-	-	-

Fuente: VII EPF-INE

Para la construcción de cada una de las matrices (según la fuente de ingreso), se trató de maximizar la cantidad de imputaciones sin agrandar el número de personas por cluster. El tamaño máximo de un cluster donde se realizó una imputación fue de 93 personas en el nivel 8, como ejemplo entre los asalariados.

Gráfico 5: Nivel en el que se encuentra el donante para el grupo de asalariados



FUENTE: VII EPF-INE.

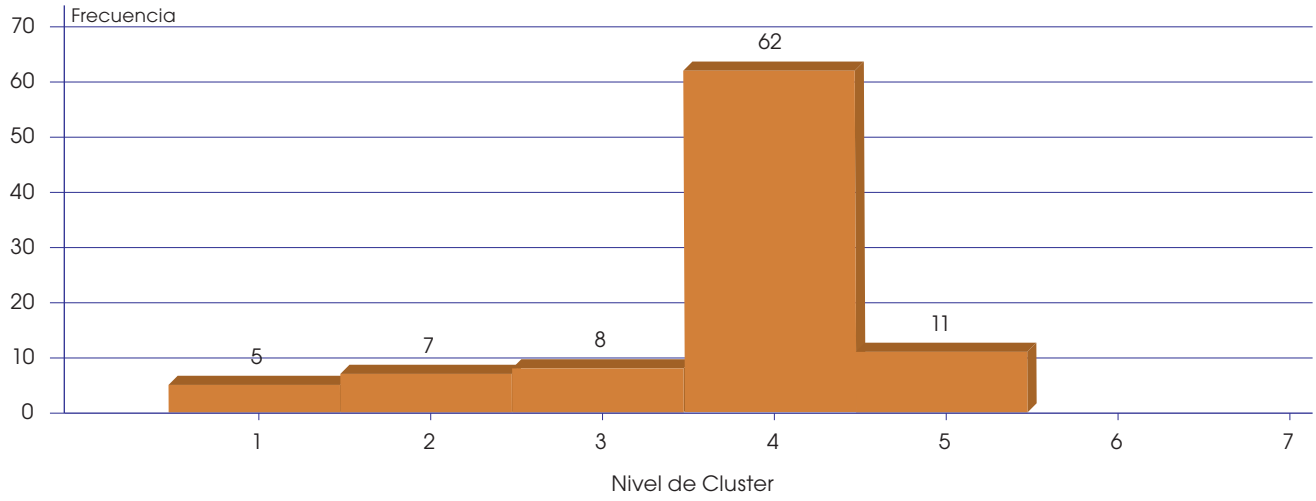
El Gráfico 5 muestra la frecuencia por nivel en que se imputaron los casos con no respuesta para los asalariados. El 77,7% de los casos fueron imputados hasta el nivel 5, lo que muestra que a pesar de la exigencia en la definición de vecino existen donantes. El tamaño de los cluster va creciendo a medida que el nivel de exigencia va disminuyendo. El 40,7% de los cluster utilizados para imputar un dato faltante tenía un solo donante; de aquí se desprende la justificación de utilizar el valor promedio de lo observado, mientras que el acumulado, representa un 80%, tiene hasta cinco observaciones.

Tabla 32: Matriz de exigencia para el vecino de honorarios

1	2	3	4	5	6	7
Sexo	Sexo	Sexo	Sexo	Sexo	Sexo	-
Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico
Escolaridad	Escolaridad	Escolaridad	Rango 1 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad
Edad R1	Edad R1	Edad R1	Edad R1	Edad R2	Edad R3	-
CIUO (1d)	CIUO (1d)	CIUO (1d)	-	-	-	-
Comuna	Región	Macrozona	-	-	-	-

Fuente: VII EPF-INE

Gráfico 6: Nivel en el que se encuentra el donante para el grupo de honorarios



Fuente: VII EPF-INE.

Para la categoría de honorarios la cantidad de niveles es menor debido a la baja cantidad de observaciones bajo esta modalidad de empleo. En esta categoría, se imputaron todos los casos con información faltante hasta el quinto nivel, no utilizando los niveles 6 y 7. Se debe destacar que sólo en el nivel 4 se imputan dos tercios de los casos. El tamaño de los cluster efectivamente utilizados para la imputación varía entre dos observaciones a máximo once casos por cluster; mientras que cerca del 54% de los casos imputados se encuentran en cluster con hasta tres observaciones para esta fuente de ingresos laborales.

Tabla 33: Matriz de exigencia para el vecino de quienes tienen negocios por cuenta propia

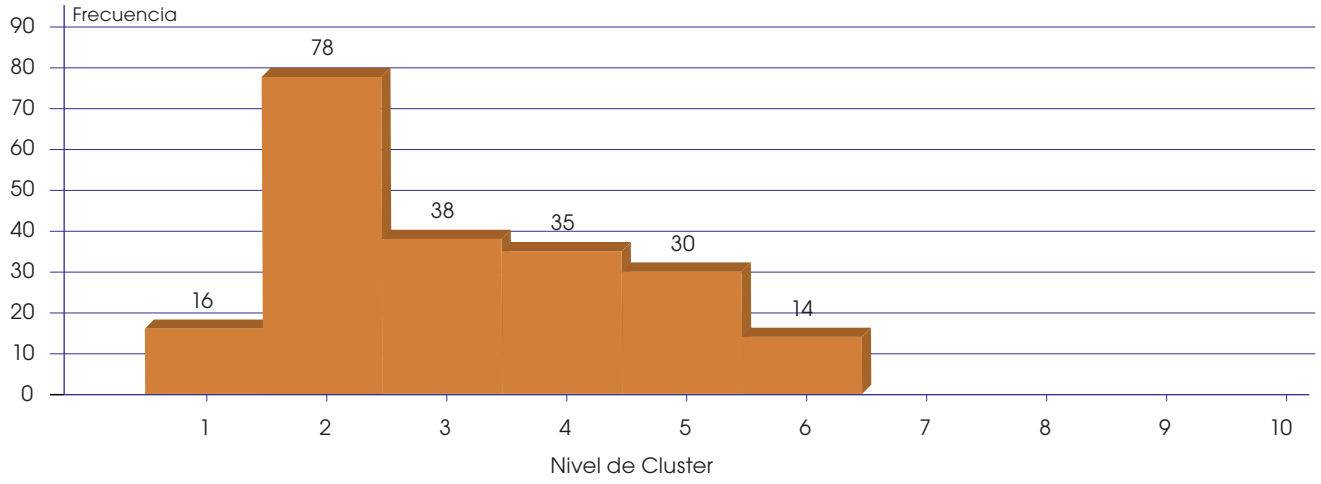
1	2	3	4	5
Sexo	Sexo	Sexo	Sexo	Sexo
Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico
Escolaridad	Escolaridad	Escolaridad	Rango 1 Escolaridad	Rango 2 Escolaridad
Edad R1	Edad R1	Edad R2	Edad R2	Edad R2
CIUO (2d)	CIUO (1d)	CIUO (1d)	CIUO (1d)	CIUO (1d)
Comuna	Región	Macrozona	Macrozona	-

6	7	8	9	10
Sexo	Sexo	Sexo	Sexo	Sexo
Estrato Económico	Estrato Económico	-	-	-
Rango 2 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad
Edad R3	Edad R3	Edad R3	Edad R3	Edad R3
-	-	CIUO (1d)	-	-
Macrozona	-	Macrozona	Macrozona	-

Fuente: VII EPF-INE

El grupo que trabaja es sus propios negocios representa el 18% de la masa que trabaja el mes de referencia de los ingresos, es por ello que para caracterizar mejor al vecino y disminuir la velocidad en que las restricciones de su definición se van diluyendo, el número de niveles vuelve a aumentar en comparación al grupo anterior.

Gráfico 7: Nivel en el que se encuentra el donante para quienes tienen negocios por cuenta propia



FUENTE: VII EPF-INE.

En esta categoría se imputaron todos los casos con no respuesta hasta el séptimo nivel, con porcentajes muy similares entre el tercer y quinto nivel, que en promedio representan el 16% (aproximadamente 34 casos cada nivel). El tamaño de los cluster varía entre 2 observaciones a máximo 79 casos por cluster utilizado (nivel 6) y cerca del 64% de los casos imputados se encuentran en clusters de dos o tres observaciones.

Tabla 34: Matriz de exigencia para el vecino de profesionales independientes

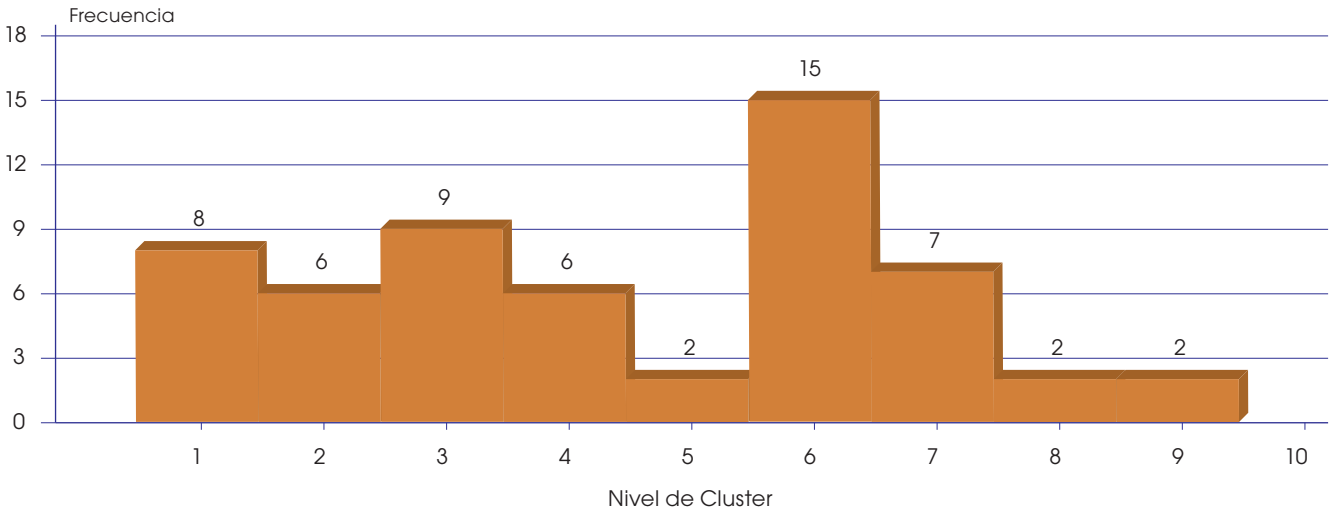
1	2	3	4	5
Sexo	Sexo	Sexo	Sexo	Sexo
Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico
Escolaridad	Escolaridad	Rango 1 Escolaridad	Rango 1 Escolaridad	Rango 2 Escolaridad
Edad R1	Edad R1	Edad R1	Edad R2	Edad R2
CIUO (2d)	CIUO (1d)	CIUO (1d)	CIUO (1d)	CIUO (1d)
Región	Macrozona	Macrozona	Macrozona	Macrozona

6	7	8	9	10
Sexo	Sexo	Sexo	Sexo	Sexo
Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	-
Rango 2 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad
Edad R2	Edad R2	Edad R3	Edad R3	Edad R3
CIUO (1d)	-	CIUO (1d)	-	CIUO (1d)
-	Macrozona	-	-	-

Fuente: VII EPF-INE

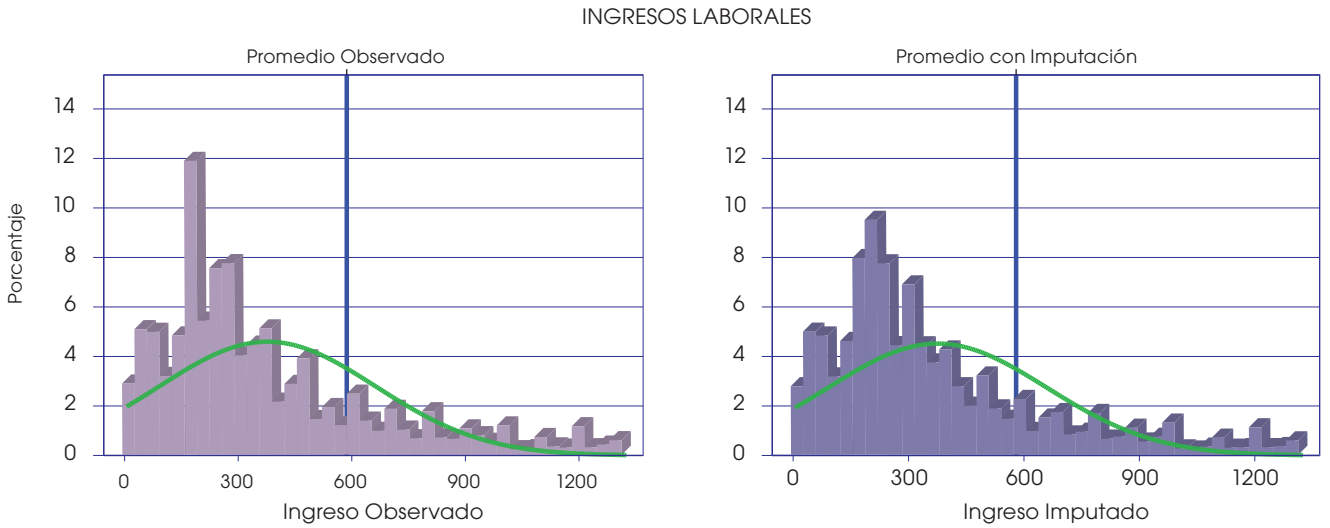
La matriz para aquellas personas que trabajan como profesionales independientes tiene la misma cantidad de niveles pero transita de diferente forma en cuanto a la exigencia de la definición de vecino cercano. Se utilizó hasta el octavo nivel para imputar todos los casos posibles. El tamaño de los cluster varía entre 2 observaciones a máximo 7 casos por cluster, en los que se realizó imputación (nivel 6 y 7) y el 60% de los casos imputados se encuentran en cluster de dos observaciones.

Gráfico 8: Nivel en el que se encuentra el donante para profesionales independientes



FUENTE: VII EPF-INE.

Gráfico 9: Comparación en la distribución de los datos observados e imputados por el método Hot-Deck



Promedio observado=586.677 - Promedio posterior a la imputación=579.506
Fuente: VII EPF-INE.

La distribución de los datos imputados mantiene la distribución original de datos: distribuciones asimétricas. Esto se puede evidenciar en las gráficas precedentes, donde el histograma de la izquierda está representando los datos observados, mientras que en el gráfico derecho están los datos imputados agregados para comparar el antes y después.

Tabla 35: Resultados muestrales de la imputación de ingresos del trabajo por método Hot-Deck

Categoría laboral	Obser.	Promedio	Mínimo	Máximo	SD	CV
Asalariados	10.014	626.447	36.000	19.000.000	817.078	1,304
- HD	10.952	618.535	36.000	19.000.000	808.740	1,308
Honorarios	499	454.240	22.000	4.000.000	465.242	1,024
- HD	592	437.502	22.000	4.000.000	442.692	1,012
Negocios por cuenta Propia	2.516	410.537	5.668	31.416.000	1.011.086	2,463
- HD	2.727	407.188	5.668	31.416.000	989.770	2,431
Profesionales Independientes	729	738.937	10.350	23.760.000	1.416.280	1,917
- HD	786	740.499	10.350	23.760.000	1.436.307	1,940

Fuente: VII EPF

En la Tabla 35 se puede evidenciar los principales puntos de una distribución antes y posterior a la imputación tipo Hot-Deck en cada categoría. Por ejemplo, para los asalariados el promedio del ingreso bajó un poco más de 2 mil pesos luego de la imputación, sin alterar los mínimos y máximos. Los coeficientes de variación se vieron alterados por centésimas. La categoría que más cambios tuvo fue la de honorarios.

Tabla 36: Matriz de exigencia para el vecino de jubilados

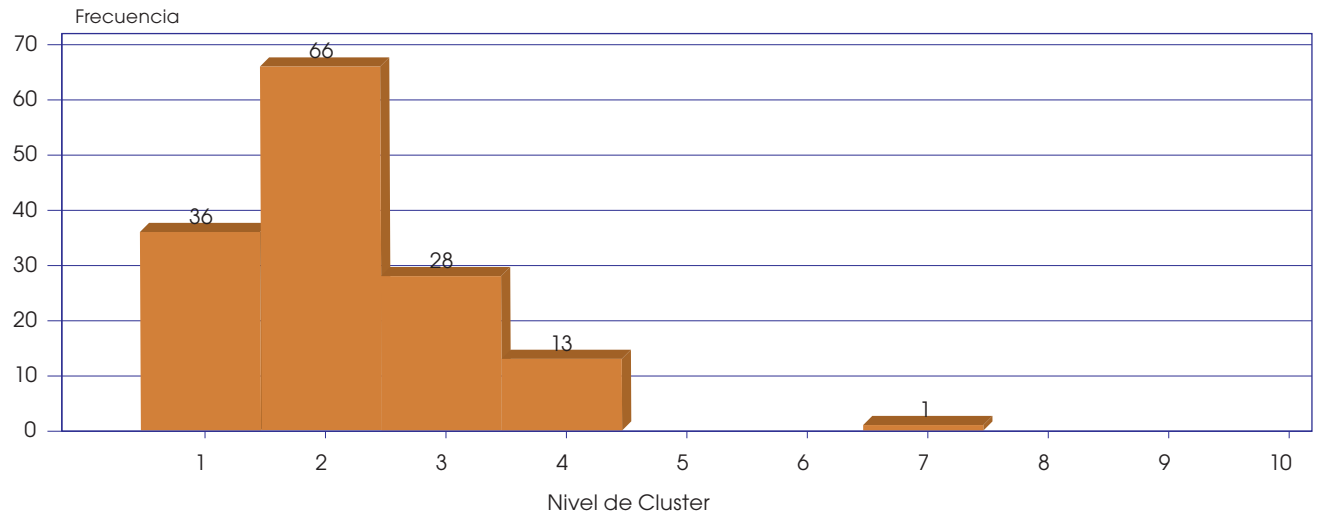
1	2	3	4	5
Sexo	Sexo	Sexo	Sexo	Sexo
Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico	Estrato Económico
Escolaridad	Escolaridad	Escolaridad	Rango 1 Escolaridad	Rango 1 Escolaridad
Edad R1	Edad R1	Edad R1	Edad R1	Edad R2
Manzana	Comuna	Región	Macrozona	Región

6	7	8	9	10
Sexo	Sexo	Sexo	Sexo	Sexo
Estrato Económico	Estrato Económico	Estrato Económico	-	-
Rango 1 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad	Rango 2 Escolaridad
Edad R2	Edad R3	Edad R3	Edad R3	Edad R3
Macrozona	Macrozona		Macrozona	-

Fuente: VII EPF-INE

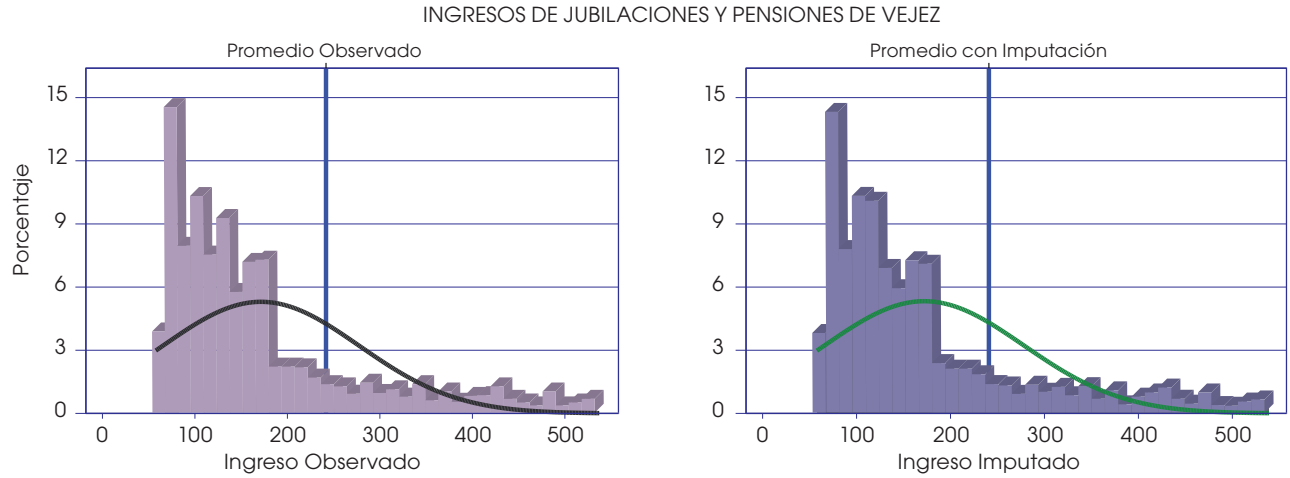
Finalmente, para las personas que reciben jubilaciones y pensiones de vejez se construyó una matriz de 10 niveles, reflexionando que se cuenta con un número menor de variables. Se imputaron el 70% de los casos en los dos primeros niveles y casi en su totalidad hasta el cuarto nivel.

Gráfico 10: Nivel en el que se encuentra el donante de jubilados



Fuente: VII EPF-INE.

Gráfico 11: Comparación en la distribución de las jubilaciones observadas e imputadas por el método Hot-Deck



Promedio observado= 241.736 - Promedio imputado=240.872
Fuente: VII EPF-INE.

El histograma no muestra alteraciones y, como se ve en la siguiente tabla, los principales estadísticos de distribución tampoco se alteraron.

Tabla 37: Resultados muestrales de la imputación de ingresos de jubilaciones por método Hot-Deck

Julaciones	Obser.	Promedio	Mínimo	Máximo	SD	CV
Observado	3.187	241.736	57.780	4.280.000	268.378	1,110
- HD	3.331	240.872	57.780	4.280.000	265.181	1,101

Fuente: VII EPF

2. Regresión Heckman

Bajo este método de imputación, la estimación de los ingresos no observados se realiza en base las observaciones completas usando un modelo de dos ecuaciones (dos etapas) y bajo el supuesto de que las personas que no respondieron ingresos tienen una distribución similar a los datos observados[48]. La primera ecuación corresponde a la síntesis de la participación en el mercado laboral, lo que requiere la construcción de una variable dicotómica, donde 1 participa y 0 no participa. Esta ecuación supone que la variable dicotómica depende del salario de reserva, lo que a su vez depende de características personales y de capital humano (Heckman, 1979).

La segunda ecuación corresponde a la estimación del salario como función de variables de capital humano y de la probabilidad de participar en el mercado laboral, la que fue estimada en la primera etapa. Corresponde a un modelo de ingresos tipo Mincer, donde el sesgo de selección está corregido. Una representación general de ambas ecuaciones es la siguiente:

$$Participación = \alpha_0 + \alpha_1 Edad + \alpha_2 Experiencia + \alpha_3 Experiencia^2 + \alpha_k .....$$

$$LN(INGRESO) = \beta_0 + \beta_1 Ed + \beta_2 Exp + \beta_3 Exp^2 + \beta_k ..... + \beta_{k+1}\lambda$$

48 Ver prueba de Little.

La especificación para cada grupo formado por la fuente laboral (dependiente e independiente) es diferente, primero debido a las características propias del grupo, pero también por la información disponible para cada uno. Una vez que se obtuvieron los parámetros de ambas ecuaciones α y β, se reemplaza la información de la matriz de las variables independientes (edad, experiencia, entre otras) pertenecientes a las personas de quienes no tenemos información de ingresos, entonces obtendremos valores predichos por el modelo (Perlbach de Maradona y Calderón, 1998, Belsley et al., 1980).

Los coeficientes se calculan con las observaciones para las que existe información completa. En el documento se presentan las tres regresiones finales, ya que el proceso de ajuste del modelo incorporó en primer lugar a todas las variables, las continuas, las categóricas que fueron incluidas seleccionando una de ellas como base y las dicotómicas. Las variables empleadas son[49]:

- **Edad.**
- **Sexo.**
- **Ocupación**: Solo a un dígito.
- **Macrozona.**
- **Estrato socioeconómico.**
- **Rango de escolaridad:** en algunos casos como variable continua y en otros mediante un rango. Éste se construyó considerando los siguientes tramos: 0 años (sin educación formal), 1 a 7, de 8 a 12; 13 a nivel técnico, universitario y superiores.
- **Dicotómicas:** el grupo de honorarios puede diferenciarse en algunas características específicas; es así que se introdujo una variable ficticia sola para este grupo (D.HONORARIO). Del mismo modo para los trabajadores independientes, se diferenció entre empleadores y los cuenta propia (DCISE). Y se interactuó las variables SEXO y escolaridad con el grupo de profesionales independientes, ya que esta relación es importante[50].

49 Definición completa de variables ver en Anexo J.
50 Ver punto 3 de la parte de ingresos: variables correlacionadas con el ingreso laboral y las jubilaciones

Tabla 38: Regresión Heckman para dependientes

Log likelihood =	**-13168,62**					
LN ingreso	**Coeficiente**	**Error Estándar**	**z**	**P>z**	**[Intervalo del 95%]**	
Tipo (Honorarios=1)	-0,501	0,029	-17,180	0	-0,558	-0,444
Sexo	-0,405	0,013	-30,860	0	-0,431	-0,379
Edad	0,067	0,003	23,010	0	0,061	0,073
Edad^2	-0,001	0,000	-20,350	0	-0,001	-0,001
Educación media (8 y 12 años)	0,274	0,025	11,000	0	0,225	0,323
Educación superior (13 a 18 años)	0,502	0,029	17,180	0	0,445	0,560
Educación superior (18 años y mas)	0,760	0,047	16,070	0	0,667	0,852
Norte	0,130	0,017	7,660	0	0,096	0,163
Centro	-0,110	0,014	-7,960	0	-0,137	-0,083
Personal directivo	1,803	0,044	40,790	0	1,717	1,890
Profesionales	1,268	0,025	50,580	0	1,219	1,318
Nivel medio	0,750	0,023	33,000	0	0,706	0,795
Empleados de oficina	0,469	0,022	21,460	0	0,426	0,512
Servicios y vendedores	0,126	0,021	5,860	0	0,084	0,168
Operarios y artesanos	0,295	0,022	13,190	0	0,251	0,339
Operadores de máquinas	0,393	0,025	15,750	0	0,344	0,442
Constante	10,871	0,064	168,720	0	10,745	10,997
Primera etapa						
Sexo (interactuada con Honorarios)	-0,170	0,084	-2,030	0,043	-0,334	-0,006
Edad	0,049	0,007	7,260	0	0,036	0,062
Edad^2	-5E-04	8E-05	-6,100	0	-0,001	0,000
Dicotómica Escolaridad superior	0,223	0,039	5,750	0	0,147	0,299
Sur	0,374	0,059	6,350	0	0,258	0,489
Constante	0,061	0,134	0,460	0,647	-0,202	0,325
LR test of indep. eqns. (rho = 0): chi2(1) = 23.97 Prob > chi2 = 0.0000						
/athrho	-0,647	0,062	-10,370	0	-0,769	-0,524
/lnsigma	-0,445	0,010	-44,390	0	-0,464	-0,425
rho	-0,569	0,042			-0,646	-0,481
sigma	0,641	0,006			0,629	0,654
lambda	-0,365	0,030			-0,423	-0,307

Fuente: VII EPF

Como vemos, todos los coeficientes son significativos al 95% (columna P>z) y la primera etapa es significativa (athrho[51]=-0.647 y valor p=0). Los signos de la estimación son los esperados; por ejemplo el ser mujer significa tener menos ingresos -0,40 (manteniendo todo lo demás constante). La escolaridad tiene un efecto positivo, al igual que la edad (esta última con rendimientos decrecientes).

51 Athrho es una trasformación de rho usada en la regresión, la que es una correlación entre los errores de la primera etapa (ecuación Probit) y la forma reducida de la segunda ecuación (o etapa), que explica la variable dependiente de interés.

Tabla 39: Regresión Heckman para quienes trabajan de forma independiente

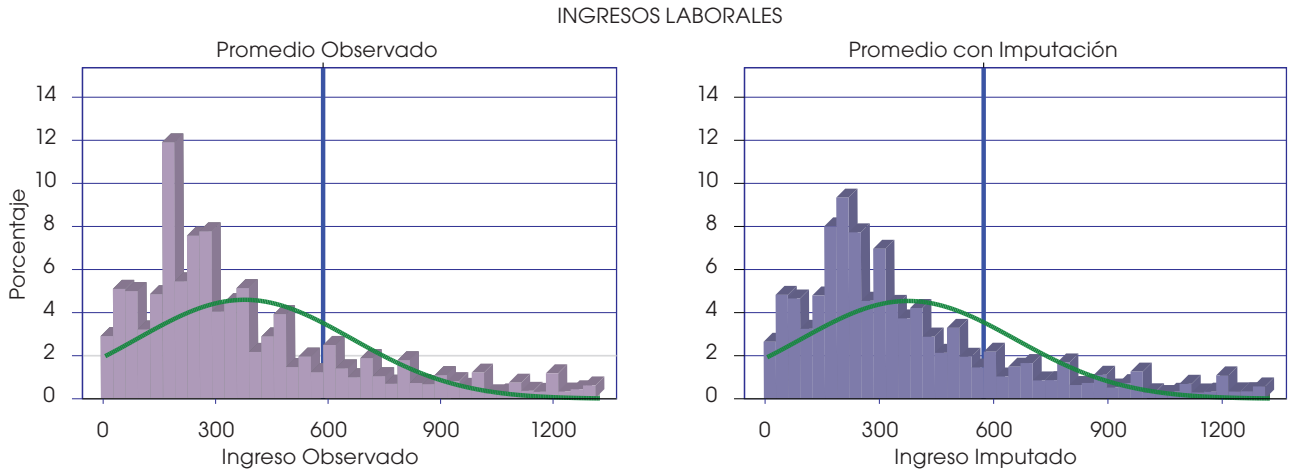
Log likelihood =	**-5331,933**					
LN ingreso	**Coeficiente**	**Error Estándar**	**z**	**P>z**	**[Intervalo del 95%]**	
DCISE	0,785	0,066	11,830	0	0,655	0,915
SEXO P	0,093	0,081	1,140	0,254	-0,067	0,252
SEXO	-0,668	0,040	-16,850	0	-0,746	-0,590
Edad	0,092	0,007	13,210	0	0,079	0,106
Edad^2	-0,001	0,000	-13,100	0	-0,001	-0,001
EDUE_P	0,270	0,059	4,560	0	0,154	0,386
Educación baja menos de 4	-0,363	0,129	-2,820	0,005	-0,616	-0,111
Educación baja (4 y 8 años)	-0,252	0,058	-4,360	0	-0,365	-0,139
Dicotómica escolaridad superior a 12	0,132	0,043	3,060	0,002	0,047	0,217
Norte	0,116	0,048	2,410	0,016	0,022	0,210
Centro	-0,089	0,040	-2,210	0,027	-0,167	-0,010
Personal directivo	0,906	0,074	12,320	0	0,762	1,050
Profesionales	0,836	0,068	12,260	0	0,703	0,970
Empleados de oficina	0,276	0,132	2,090	0,036	0,017	0,534
Servicios y vendedores	0,218	0,050	4,360	0	0,120	0,316
No calificados	-0,512	0,047	-11,010	0	-0,604	-0,421
Estrato Medio	0,069	0,039	1,770	0,077	-0,007	0,145
Estrato Alto	0,495	0,058	8,460	0	0,380	0,609
Constante	10,186	0,173	58,820	0	9,846	10,525
Primera etapa						
Edad	0,004	0,002	2,070	0,039	0,000	0,009
Zona	0,203	0,060	3,370	0,001	0,085	0,321
ciuo8	0,405	0,151	2,680	0,007	0,109	0,701
Constante	1,089	0,108	10,080	0	0,877	1,300
LR test of indep. eqns. (rho = 0): chi2(1) = 18.64 Prob > chi2 = 0.0000						
/athrho	-0,909	0,098	-9,270	0	-1,102	-0,717
/lnsigma	0,005	0,016	0,300	0,761	-0,027	0,037
rho	-0,721	0,047			-0,801	-0,615
sigma	1,005	0,016			0,973	1,038
lambda	-0,724	0,056			-0,834	-0,615

Fuente: VII EPF

Otro es el caso para los que trabajan de forma independiente. La Tabla 39 muestra el modelo final que se ajusta mejor a los datos de los ingresos de este tipo de trabajadores. Vemos que el test para determinar si la primera etapa es válida es satisfactoria (LR), sin embargo, tiene un coeficiente menor que para el caso de los dependientes.

Utilizando los parámetros de la segunda columna desde la izquierda de cada tabla se predijeron los ingresos de aquellas personas que no respondieron. A continuación, se muestran los principales puntos de la distribución de cada categoría según este método de imputación.

Gráfico 12: Comparación en la distribución de los datos observados e imputados por el método de regresión Heckman



Promedio observado=586.677 - Promedio imputado=573.735
Fuente: VII EPF-INE.

Como vemos el histograma muestra una distribución un tanto más plana, sin embargo en la Tabla 40 se evidencia que el grupo de honorarios es el que más reduce su ingreso promedio, bajando aproximadamente 25 mil pesos, en posición contraria, el grupo de profesionales independientes reduce su ingreso promedio en cerca de 7 mil pesos siendo el grupo que reduce más su coeficiente de variación

Tabla 40: Resultados muestrales de la imputación de ingresos del trabajo por método Heckman

Categoría laboral	Obser.	Promedio	Mínimo	Máximo	SD	CV
Asalariados	10.014	626.447	36.000	19.000.000	817.078	1,304
- Heckman	10.952	613.682	36.000	19.000.000	792.014	1,291
Honorarios	499	454.240	22.000	4.000.000	465.242	1,024
- Heckman	592	428.462	22.000	4.000.000	439.224	1,025
Negocios por cuenta Propia	2.516	410.537	5.668	31.416.000	1.011.086	2,463
- Heckman	2.727	399.425	5.668	31.416.000	976.519	2,445
Profesionales Independientes	729	738.937	10.350	23.760.000	1.416.280	1,917
- Heckman	786	731.310	10.350	23.760.000	1.377.041	1,883

Fuente: VII EPF

Tabla 41: Regresión Heckman para quienes reciben jubilaciones

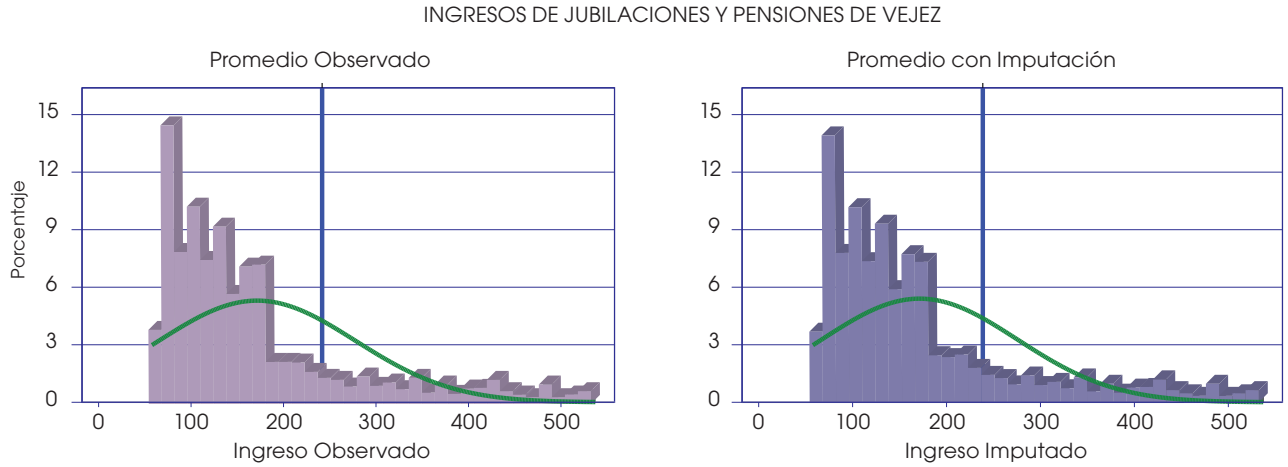
Log likelihood =-3574,942

LN ingreso	Coeficiente	Error Estándar	z	P>z	[Intervalo del 95%]	
Sexo	-0,2641	0,0230	-11,49	0	-0,309	-0,219
Edad	-0,0023	0,0014	-1,66	0,096	-0,005	0,000
Escolaridad	0,0526	0,0027	19,8	0	0,047	0,058
JU01B	0,1414	0,0300	4,71	0	0,083	0,200
Constante	11,6846	0,1051	111,2	0	11,479	11,891
Primera etapa						
Edad	-0,00888	0,00378	-2,35	0,019	-0,016	-0,001
JU01B	0,05614	0,07106	0,79	0,43	-0,083	0,195
Constante	2,17144	0,25938	8,37	0	1,663	2,680
LR test of indep. eqns. (rho = 0): chi2(1) = 169.40 Prob > chi2 = 0.0000						
/athrho	2,245	0,146	15,350	0	1,958	2,532
/lnsigma	-0,369	0,013	-28,120	0	-0,394	-0,343
rho	0,978	0,006			0,961	0,987
sigma	0,692	0,009			0,674	0,710
lambda	0,676	0,011			0,655	0,698

Fuente: VII EPF

El modelo de regresión para los jubilados es más simple ya que considera un número inferior de variables. La variable dicotómica JU01B representa a quienes no reciben ingresos laborales[52], por lo que su signo positivo se interpreta como un salario de reserva para no participar en el mercado laboral, coeficiente que denota no ser significativo pero mejora las estimaciones en su conjunto.

Gráfico 13: Comparación en la distribución de las jubilaciones observadas e imputadas por el método de regresión Heckman



Promedio observado=241.736 - Promedio imputado=238.623
Fuente: VII EPF-INE.

52 Es un aproximado de variable inactivo, pero no es igual debido a los períodos de referencia entre la declaración de la actividad laboral y la recopilación de los ingresos.

Tabla 42: Resultados muestrales de la imputación de ingresos por jubilaciones por el método Heckman

Julaciones	Obser.	Promedio	Mínimo	Máximo	SD	CV
Observado	3.187	241.736	57.780	4.280.000	268.378	1,110
- Heckman	3.327	238.623	57.780	4.280.000	263.258	1,103

Fuente: VII EPF

La distribución de los ingresos por jubilaciones no se altera con el método de imputación.

3. Imputación múltiple mediante regresión

El método IM aborda el tema de la no respuesta desde una perspectiva estadística, utilizando múltiples bases de datos donde cada una de ellas tiene un valor posible para la observación faltante. La estimación del salario en cada base se realiza con una función dependiente de variables de capital humano (tipo Mincer).

$$LN(INGRESO) = \theta_0 + \theta_1 Ed + \theta_2 Exp + \theta_3 Exp^2 + \theta_k ....$$

El siguiente paso es calcular el promedio de los ingresos de las 30 bases generadas[53], para así tener un solo dato por persona.

Un modelo plausible de los datos siguiendo la teoría del capital humano se resume en la siguiente ecuación Mincer: $LN(Y) = \beta X + \phi D + E$

Donde, Y: Ingreso total mensual, X: Variables de cuantía (edad, escolaridad, etc.) y D: conjunto de variables Dummy (ciudad, categoría del oficio CIOU 1 dígito, sexo).

A continuación se muestran las tres regresiones base para la imputación múltiple de tipo Mincer. Estas son las iniciales, ya que se utilizan regresiones gausianas (método de estimaciones robustas) para generar 30 valores predichos para cada observación.

Como se observa el estadístico de bondad de ajuste (R2 ajustado) es menor para los trabajadores independientes. Los modelos elegidos son similares en estructura a los modelos de la segunda etapa de las regresiones Heckman.

53 En la actualidad los programas estadísticos permiten realizar varias simulaciones rápidamente: STATA en su manual recomienda al menos 20, al igual que la experiencia del Banco Central de Chile en la Encuesta Financiera de Hogares 2011 (Banco Central, 2011), Schafer (1997), Alfaro y Fuenzalida (2009) y Stata Corporation (2009).

Tabla 43: Regresión base dependientes

N. Obs	10502
F(16, 10485)	821
Prob > F	0
R2 ajustado	0,5553

log Ingreso	Coeficiente	Error Est.	t	P>t	Intervalo de confianza	
					Inferior	Superior
Dummy Honorario	-0,701	0,043	-16,110	0	-0,786	-0,615
Sexo	-0,418	0,013	-31,580	0	-0,444	-0,392
Sexo*D Hon	0,300	0,056	5,320	0	0,189	0,411
Edad	0,074	0,003	26,550	0	0,068	0,079
Edad^2	-0,001	0,000	-23,140	0	-0,001	-0,001
Dicotómica Educación superior a 12	0,058	0,020	2,880	0,004	0,019	0,098
Escolaridad^2	0,003	0,000	18,920	0	0,002	0,003
Norte	0,120	0,017	7,200	0	0,087	0,153
Centro	-0,123	0,014	-9,060	0	-0,150	-0,096
Personal directivo	1,575	0,046	34,570	0	1,485	1,664
Profesionales	1,035	0,029	36,090	0	0,979	1,091
Nivel medio	0,668	0,022	29,800	0	0,624	0,712
Empleados de oficina	0,412	0,022	18,820	0	0,369	0,454
Servicios y vendedores	0,092	0,021	4,310	0	0,050	0,134
Operarios y artesanos	0,290	0,022	13,120	0	0,247	0,333
Operadores de máquinas	0,375	0,025	15,240	0	0,327	0,423
Constante	10,577	0,058	183,540	0	10,464	10,690

Fuente: VII EPF

Los dependientes contienen un número grande de observaciones, lo que combinado con las características tomadas para el modelo base que son consistentes con la teoría, evidencia semi-elasticidades con los signos esperados lo que lleva a un ajuste conveniente para la predicción de datos faltantes con este método. Por ejemplo, el signo negativo de la variable sexo muestra que los sueldos y salarios son inferiores para las mujeres, dejando todo lo demás constante. En añadidura al modelo Heckman, en este caso el sexo interactuado con una dicotómica para el subgrupo de honorarios es significativa y positiva, lo que indica que para las mujeres que trabajan como honorarios existe una brecha menor comparando con los sueldos de hombre.

En el caso del grupo de independientes, la ocupación pivote de comparación continúa siendo los trabajadores no calificados, de ahí que los coeficientes β asociados al oficio son positivos. En este caso, si bien el número de observaciones es menor, también se observa un estadístico de ajuste (r2 ajustado) favorable para la predicción de ingresos faltantes.

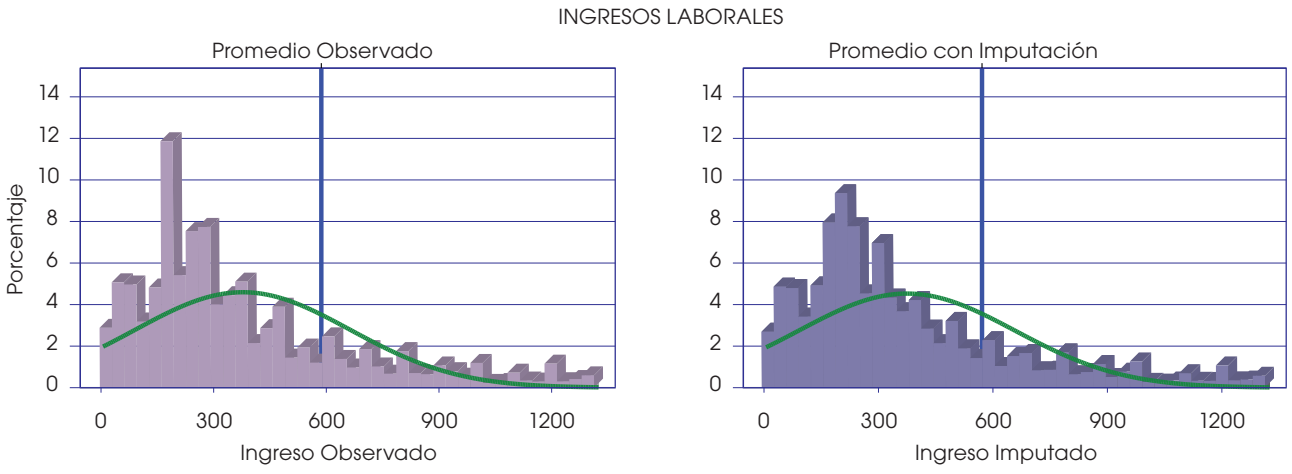
Tabla 44: Regresión base para quienes trabajan de forma independiente

N. Obs	3245
F(18, 3226)	165,45
Prob > F	0
R2 ajustado	0,48

log Ingreso	Coeficiente	Error Est.	t	P>t	Intervalo de confianza	
					Inferior	Superior
DCISE	0,112	0,082	1,360	0,17	-0,049	0,273
SEXO P	0,784	0,067	11,610	0	0,651	0,916
SEXO	-0,701	0,039	-17,840	0	-0,778	-0,624
Edad	0,096	0,007	13,660	0	0,082	0,110
Edad^2	-0,001	0,000	-13,460	0	-0,001	-0,001
EDUE_P	0,263	0,060	4,410	0	0,146	0,380
Educación baja menos de 4	-0,387	0,131	-2,960	0	-0,643	-0,130
Educación baja (4 y 8 años)	-0,244	0,058	-4,210	0	-0,358	-0,130
Dicotómica escolaridad superior a 12	0,133	0,044	3,050	0,002	0,048	0,218
Norte	0,143	0,047	3,030	0,002	0,051	0,236
Centro	-0,070	0,039	-1,780	0,075	-0,147	0,007
Personal Directivo de Empresas	0,904	0,075	12,070	0	0,757	1,051
Profesionales e Intelectuales	0,821	0,069	11,850	0	0,685	0,957
Empleados de Oficina	0,259	0,134	1,940	0,053	-0,003	0,521
Trabajadores de los Servicios	0,225	0,050	4,490	0	0,127	0,323
Trabajadores Calificados	-0,520	0,047	-11,090	0	-0,611	-0,428
Estrato Medio	0,067	0,039	1,720	0,085	-0,009	0,144
Estrato Alto	0,482	0,059	8,170	0	0,366	0,598
Constante	9,983	0,173	57,690	0	9,644	10,322

Fuente: VII EPF

Gráfico 14: Comparación en la distribución de los datos observados e imputados por el método de imputación múltiple



Promedio observado=586.677 - Promedio imputado= 571.405
Fuente: VII EPF-INE.

Tabla 45: Resultados muestrales de la imputación de ingresos del trabajo por método IM

Categoría laboral	Obser.	Promedio	Mínimo	Máximo	SD	CV
Asalariados	10.014	626.447	36.000	19.000.000	817.077,5	1,304
- Imputación Múltiple	10.948	611.691	36.000	19.000.000	791.970,6	1,295
Honorarios	499	454.240	22.000	4.000.000	465.242,2	1,024
- Imputación Múltiple	592	422.928	22.000	4.000.000	438.752,6	1,037
Negocios por cuenta Propia	2.516	410.537	5.668	31.416.000	1.011.085,9	2,463
- Imputación Múltiple	2.727	397.440	5.668	31.416.000	976.031,7	2,456
Profesionales Independientes	729	738.937	10.350	23.760.000	1.416.279,6	1,917
- Imputación Múltiple	786	725.666	10.350	23.760.000	1.375.266,2	1,895

Fuente: VII EPF

Tabla 46: Regresión base para quienes reciben jubilaciones

N. Obs	3173
F(4, 3168)	270,38
Prob > F	0,00
R2 ajustado	0,254

log Ingreso	Coeficiente	Error Est.	t	P>t	Intervalo de confianza	
					Inferior	Superior
Sexo	-0,321	0,022	-14,4	0	-0,364	-0,277
Edad	0,002	0,001	2,0	0	0,000	0,005
Dicotómica escolaridad superior a 12	-0,139	0,036	-3,9	0	-0,210	-0,068
Escolaridad^2	0,004	0,000	20,5	0	0,004	0,005
Constante	11,623	0,093	125,0	0	11,441	11,805

Fuente: VII EPF

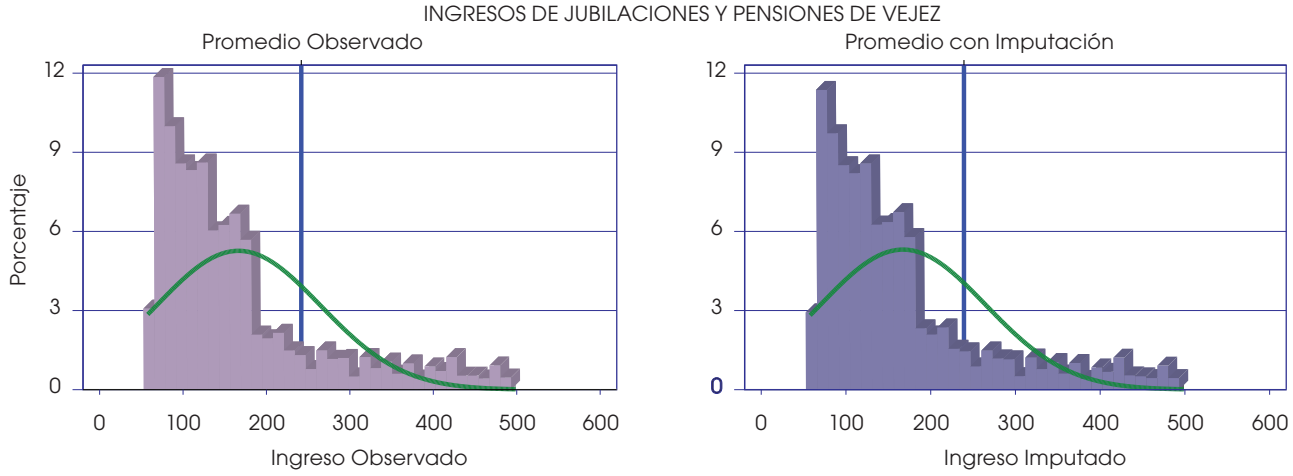
La regresión empleada para la imputación múltiple de los ingresos de jubilaciones es similar a la regresión planteada para Heckman y muestra un bajo ajuste con un R2 cercano al 25%. Este ajuste es bajo pues no se cuentan con características propias de los ingresos que generaron las jubilaciones, pero esto sucede en la práctica en las encuestas de corte transversal y este efecto negativo se contrarresta con los ingresos promedios y su dispersión incluyendo la imputación, como se observa en la siguiente tabla. El ingreso promedio baja en menos de 2.500 pesos manteniendo de cerca el coeficiente de variación.

Tabla 47: Resultados muestrales de la imputación de ingresos por jubilación por método IM

Julaciones	Obser.	Promedio	Mínimo	Máximo	SD	CV
Observado	3.187	241.736	57.780	4.280.000	268.378	1,110
- IM	3.327	239.276	57.780	4.280.000	263.335	1,101

Fuente: VII EPF

Gráfico 15: Comparación en la distribución de las jubilaciones observadas e imputadas el método de imputación múltiple



Promedio observado=241.736 - Promedio imputado=239.268
Fuente: VII EPF-INE.

4. Máxima verosimilitud con EM (Expectation Minimization)

El método EM es un procedimiento que maximiza la verosimilitud de forma iterativa y que en cada repetición se incorpora más información. Los pasos en la iteración se alternan entre realizar una expectativa (E), lo que crea una función de expectativa sobre la log-verosimilitud evaluado con los parámetros estimados en (E) y se maximiza (M). Los nuevos parámetros se utilizan para determinar la distribución de la variable latente en la siguiente etapa E.

El modelo de verosimilitud es:

$$L(\theta; X) = p(X \mid \theta) = \sum_{z} p(X, Z \mid \theta)$$

Paso E: $$Q\left(\theta \mid \theta^{(t)}\right) = E_{Z|X,\theta^{(t)}}\left[\log L(\theta; X, Z)\right]$$

Paso M: $$\theta^{(t+1)} = \arg\max_{\theta} Q\left(\theta \mid \theta^{(t)}\right)$$

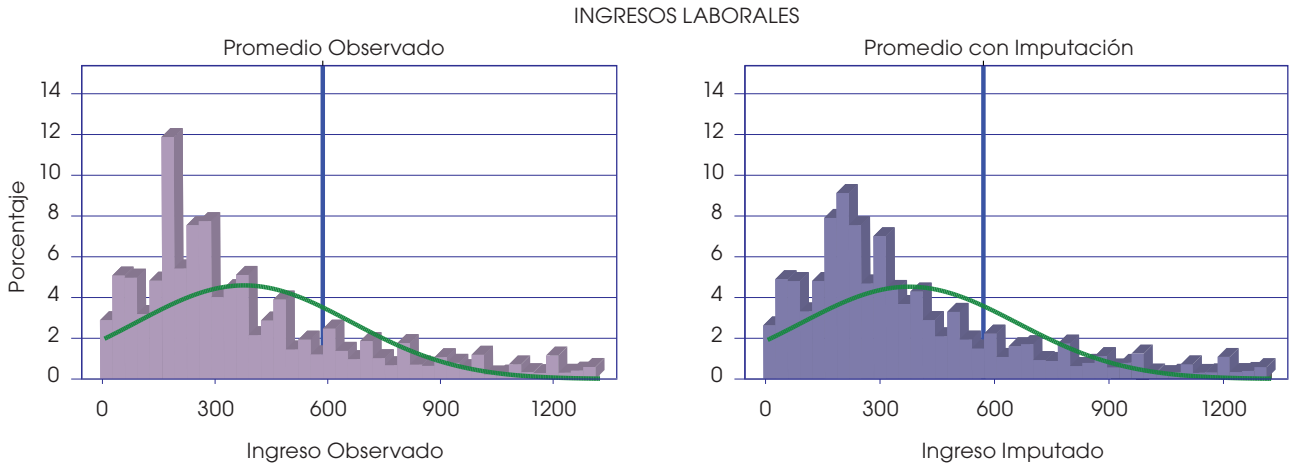
Las covariables observadas son X, la variable latente no observada es Z (los ingresos laborales y los de jubilación) y θ son los parámetros que inicialmente toman un valor aleatorio. El supuesto principal es que todas las variables en el análisis tienen una distribución normal multivariada.

Como se explicó anteriormente se debe elegir un set de covariables (X) que ayude a plantear la distribución de la variable latente. Para poder evidenciar el uso de esta metodología se optó por comparar dos set de covariables. Es decir, esto permite que el nivel de exigencia para la distribución normal multivariada se adecúe a los datos observados y mientras menos covariables menor exigencia y por lo tanto, un menor ajuste. Con ambos sets de restricciones esporádicamente se obtienen estimaciones inferiores a los mínimos observados para los casos faltantes, siendo éste el principal problema con este método. En estos casos se optó por no asignarle el valor predicho y asumir la falta del valor.

a) EM Restringido:

- SEXO
- EDAD
- ESTRATO ECONÓMICO
- ESCOLARIDAD
- TIPO DE OCUPACIÓN (AE07)
- CIUO (1 dígito)
- ZONA

Gráfico 16: Comparación en la distribución de los datos observados e imputados por el método EM restringido



Promedio observado=586.677 - Promedio imputado=570.876
Fuente: VII EPF-INE.

El Gráfico 16 muestra las fuentes agregadas del ingreso proveniente del trabajo y está concentrado hasta el percentil 90 para mostrar mejor los datos. Después de la imputación se ve un histograma un poco más plano pero con la misma asimetría.

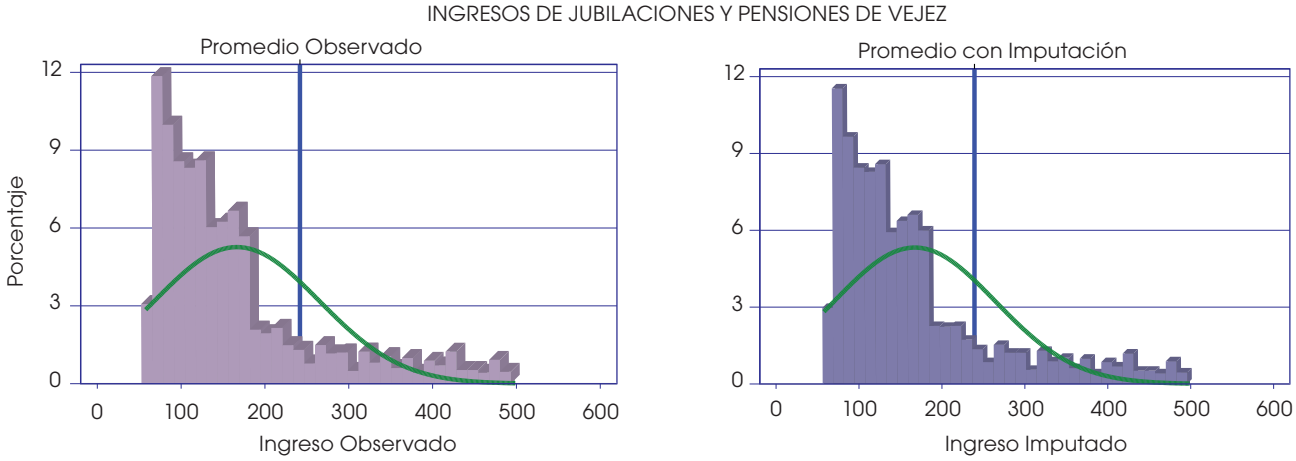
Los ingresos estimados mediante este método reducen el promedio, variando en 19 mil pesos siendo el diferencial entre promedios superior para los honorarios. Los cambios en la distribución se pueden evidenciar en el coeficiente de variación que disminuye para todas las fuentes, exceptuando los honorarios, en cuyo caso subió de 1.024 a 1.034.

Tabla 48: Resultados muestrales de la imputación de ingresos del trabajo por método EM Restringido

Categoría laboral	Obser.	Promedio	Mínimo	Máximo	SD	CV
Asalariados	10.014	626.447	36.000	19.000.000	817.077,5	1,304
- EM restingido	10.937	611.285	36.000	19.000.000	788.555,3	1,290
Honorarios	499	454.240	22.000	4.000.000	465.242,2	1,024
- EM restingido	592	421.527	22.000	4.000.001	436.030,2	1,034
Negocios por cuenta Propia	2.516	410.537	5.668	31.416.000	1.011.085,9	2,463
- EM restingido	2.721	398.439	5.668	31.416.030	976.713,2	2,451
Profesionales Independientes	729	738.937	10.350	23.760.000	1.416.279,6	1,917
- EM restingido	786	722.070	10.350	23.760.017	1.371.329,5	1,899

Fuente: VII EPF

Gráfico 17: Comparación en la distribución de las jubilaciones observadas e imputadas por el método EM restringido



Promedio observado=241.736 - Promedio imputado=239.229
Fuente: VII EPF-INE.

Para el caso de las jubilaciones, la distribución parece idéntica. Tanto el Gráfico 17 como en la Tabla 49, se muestra la evidencia del cambio disminuido en la distribución que realiza una imputación que considera la máxima verosimilitud condicionada a las características observadas.

Tabla 49: Resultados muestrales de la imputación de ingresos por jubilación por método EM restringido

Julaciones	Obser.	Promedio	Mínimo	Máximo	SD	CV
Observado	3.187	241.736	57.780	4.280.000	268.378	1,110
- EM restingido	3.327	239.229	57.780	4.280.000	263.316	1,101

Fuente: VII EPF

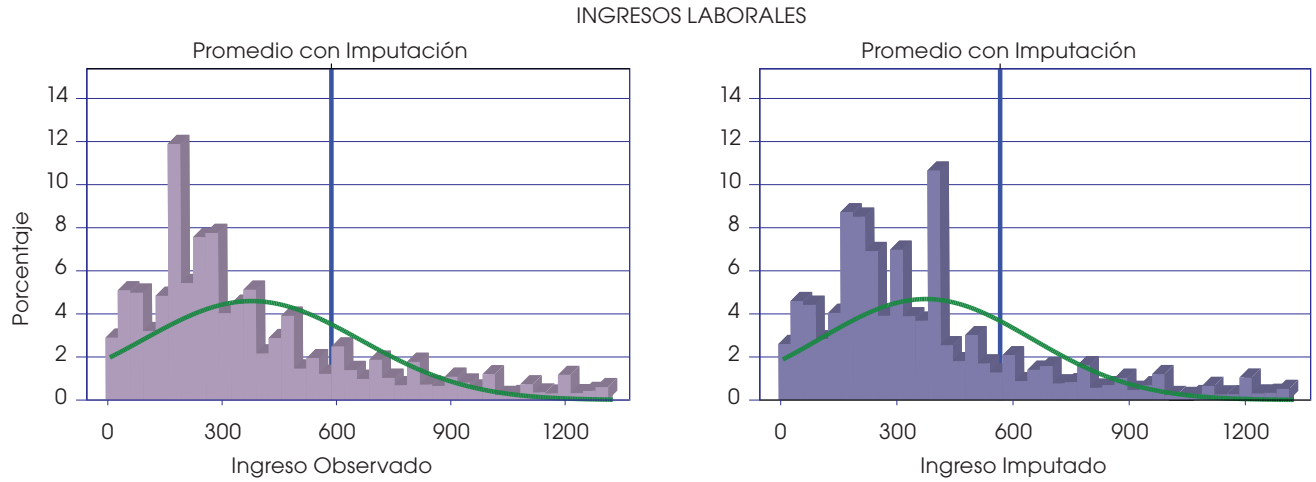
b) EM no restringido:

Como medida de comparación también se consideró un modelo con menos restricciones en la predicción de la distribución. Se consideraron dos covariables, las mismas que son utilizadas para la prueba Little para determinar el mecanismo de no respuesta, éstas son:

- ESTRATO ECONÓMICO
- ZONA

Se observa que la imputación sin considerar las características de las personas y de sus trabajos modifica el comportamiento del grupo. Esto se evidencia en el histograma con el comportamiento cerca de la cota de los 300 mil pesos.

Gráfico 18: Comparación en la distribución de los datos observados e imputados por el método EM



Promedio observado=586.677 - Promedio imputado=566.742
Fuente: VII EPF-INE.

Con ayuda de la Tabla 50, se evidencia una mayor caída en los promedios, siendo los grupos más pequeños los más afectados (honorarios y profesionales independientes) con una baja de 26 mil pesos aproximadamente. En este mismo sentido se realiza la comparación con la imputación EM restringida y los mismos grupos son los que muestran mayor diferencia de 10 mil pesos, aunque para el caso de honorarios es más alto el promedio no controlado. En promedio los ingresos disminuyen en aproximadamente 20 mil pesos, en comparación al grupo control.

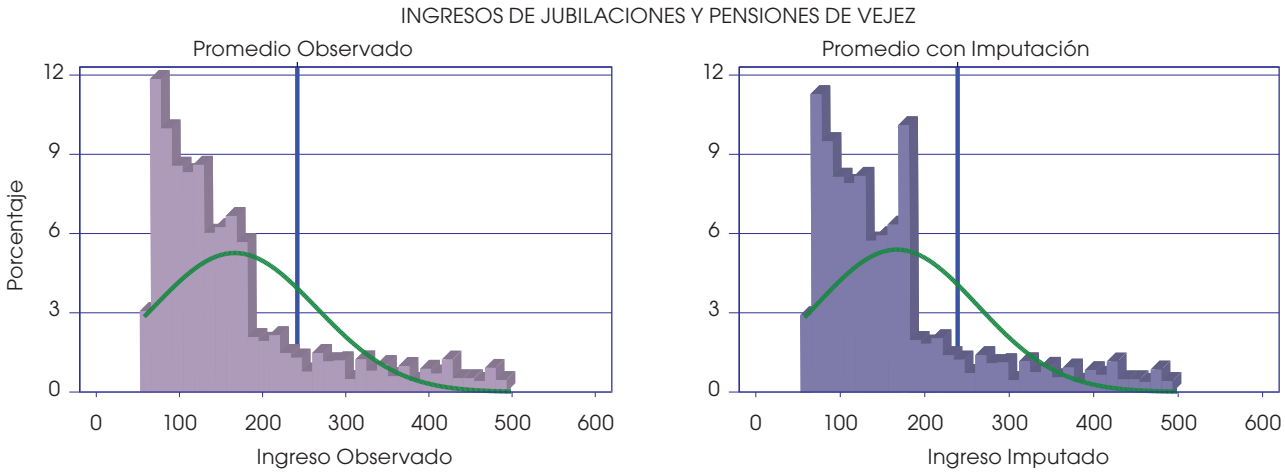
Tabla 50: Resultados muestrales de la imputación de ingresos del trabajo por método EM

Categoría laboral	Obser.	Promedio	Mínimo	Máximo	SD	CV
Asalariados	10.014	626.447	36.000	19.000.000	817.077,5	1,304
- EM	10.937	607.322	36.000	19.000.000	784.241,3	1,291
Honorarios	499	454.240	22.000	4.000.000	465.242,2	1,024
- EM	592	430.014	22.000	4.000.001	430.748,2	1,002
Negocios por cuenta Propia	2.516	410.537	5.668	31.416.000	1.011.085,9	2,463
- EM	2.721	392.660	5.668	31.416.030	974.248,3	2,481
Profesionales Independientes	729	738.937	10.350	23.760.000	1.416.279,6	1,917
- EM	786	711.662	10.350	23.760.017	1.367.380,0	1,921

Fuente: VII EPF

Considerando nuevamente la zona (RM o resto de las regiones) y el estrato económico del diseño muestral, se realizó la imputación de los ingresos por jubilaciones y pensiones de vejez. Mediante este ejercicio se puede evidenciar que cuando no se controla por más co-variables, la distribución final se ve afectada. En este caso se evidencia un aumento en la concentración cercana al promedio.

Gráfico 19: Comparación en la distribución de las jubilaciones observadas e imputadas por el método EM



Promedio observado=241.736 - Promedio imputado=238.928
Fuente: VII EPF-INE.

Tabla 51: Resultados muestrales de la imputación de ingresos por jubilación por método EM

Julaciones	Obser.	Promedio	Mínimo	Máximo	SD	CV
Observado	3.187	241.736	57.780	4.280.000	268.378	1,110
- EM	3.327	238.928	57.780	4.280.000	263.011	1,101

Fuente: VII EPF

C. FASE DE ANÁLISIS: EVALUACIÓN DE LAS METODOLOGÍAS DE IMPUTACIÓN, REGLAS DECISIÓN

En esta sección se realiza un análisis descriptivo de las variables post imputación, centrándose en la conservación de la distribución de los datos observados. Además, se calculan dos estadísticos, raíz del error cuadrático medio y error absoluto medio, para apoyo en la elección sobre el método de imputación a utilizar en cada caso.

Tabla 52: Estadísticos descriptivos de la distribución por fuente laboral y método de imputación

Categoría laboral	Promedio	SD	CV	p25	p50	p75
Asalariados	**626.447**	**817.078**	**1,304**	**220.000**	**350.000**	**700.000**
- HD	618.535	808.740	1,308	220.000	350.000	691.282
- Heckman	613.682	792.014	1,291	220.000	350.000	682.232
- Imputación Múltiple	611.691	791.971	1,295	220.000	350.000	680.000
- EM restingido	611.285	788.555	1,290	220.000	350.441	683.900
- EM	607.322	784.241	1,291	230.000	396.763	650.000
Honorarios	**454.240**	**465.242**	**1,024**	**177.777**	**300.000**	**600.000**
- HD	437.502	442.692	1,012	178.889	296.809	550.000
- Heckman	428.462	439.224	1,025	170.000	290.773	546.014
- Imputación Múltiple	422.928	438.753	1,037	163.995	281.571	538.944
- EM restingido	421.527	436.030	1,034	176.800	281.607	509.500
- EM	430.014	430.748	1,002	192.500	300.028	500.000
Negocios por cuenta Propia	**410.537**	**1.011.086**	**2,463**	**75.000**	**200.000**	**400.000**
- HD	407.188	989.770	2,431	76.000	200.000	399.845
- Heckman	399.425	976.519	2,445	80.000	200.000	373.997
- Imputación Múltiple	397.440	976.032	2,456	80.000	200.000	375.000
- EM restingido	398.439	976.713	2,451	80.000	197.440	380.000
- EM	392.660	974.248	2,481	80.000	180.000	350.000
Profesionales Independientes	**738.937**	**1.416.280**	**1,917**	**200.000**	**350.000**	**720.000**
- HD	740.499	1.436.307	1,940	200.000	350.000	720.000
- Heckman	731.310	1.377.041	1,883	200.000	350.000	732.356
- Imputación Múltiple	725.666	1.375.266	1,895	200.000	350.000	710.959
- EM restingido	722.070	1.371.329	1,899	200.000	351.740	700.000
- EM	711.662	1.367.380	1,921	200.000	362.830	700.000

Fuente: VII EPF

Para asalariados, honorarios y profesionales independientes, los métodos IM y EM Restringido, son los métodos cuyo promedio final más se diferencia al calculado con los datos observados. El método Heckman, en tanto, es el que más cambia el promedio para los trabajadores que tienen negocios por cuenta propia. Por otro lado, se puede ver que los tres percentiles, P25, P50 y P75, presentados casi no varían bajo ningún método, con excepción del grupo de honorarios.

Tabla 53: Estadísticos descriptivos de la distribución por método de imputación de las jubilaciones

Categoría laboral	Promedio	SD	CV	p25	p50	p75
Jubilaciones	241.736	268.378	1,110	103.790	147.660	255.944
- HD	240.872	265.181	1,101	104.860	148.535	255.105
- Heckman	238.623	263.258	1,103	105.503	149.800	246.100
- Imputación Múltiple	239.276	263.335	1,101	105.659	149.612	249.310
- EM restingido	239.229	263.316	1,101	105.187	149.800	249.310
- EM	238.928	263.011	1,101	106.198	152.464	245.030

Para las jubilaciones se evidencia una menor diferencia entre los promedios y la distribución si se compara con los métodos implementados. Sobre el coeficiente de variación no se ven diferencias reveladoras, siendo el único diferente el modelo Heckman que denota un CV diferente. El método Hot-Deck es el con mayor promedio posterior a la imputación, pero es el más cercano al set de observaciones con datos completos, así como también lo es el percentil 75.

Tabla 54: Ingreso promedio por hogar

	HOT DECK	HECKMAN	IMPUTACIÓN MÚLTIPLE	EM RESTRINGIDO	EM NO RESTINGIDO
De la Ocupación	**732.711**	**729.454**	**726.301**	**723.601**	**718.771**
Asalariados	514.653	514.295	512.343	510.190	507.281
Honorarios	20.393	19.987	19.809	19.774	20.194
Negocios por cuenta propia	104.193	102.292	101.724	101.480	99.764
Profesionales Independientes	51.969	51.376	50.922	50.654	50.029
Ingresos de Otros Trabajos	41.503	41.503	41.503	41.503	41.503
De Otras Fuentes	152.031	151.595	151.824	151.554	151.551
Rentas de la Propiedad	**32.008**	**32.008**	**32.008**	**32.008**	**32.008**
Jubilaciones	71.689	71.253	71.482	71.212	71.209
Transferencias (sin Gasto)	48.334	48.334	48.334	48.334	48.334
TOTAL sin arriendo imputado	**884.743**	**881.049**	**878.124**	**875.155**	**870.322**

Fuente: VII EPF

Por otro lado, la conclusión debe incluir un análisis de la estructura del ingreso promedio por hogar. En la Tabla 54. Se puede observar que la estructura varía entre cada método, siendo la diferencia alrededor de 48 mil pesos. El método Hot-Deck presenta ingresos por la ocupación para los trabajadores dependientes como para los independientes, quienes en promedio están por encima.

Tabla 55: Estructura del ingreso promedio por hogar

	HOT DECK	HECKMAN	IMPUTACIÓN MÚLTIPLE	EM RESTRINGIDO	EM NO RESTINGIDO
De la Ocupación	**82,82**	**82,79**	**82,71**	**82,68**	**82,59**
Asalariados	58,17	58,37	58,35	58,30	58,29
Honorarios	2,30	2,27	2,26	2,26	2,32
Negocios por cuenta propia	11,78	11,61	11,58	11,60	11,46
Profesionales Independientes	5,87	5,83	5,80	5,79	5,75
Ingresos de Otros Trabajos	4,69	4,71	4,73	4,74	4,77
De Otras Fuentes	**17,18**	**17,21**	**17,29**	**17,32**	**17,41**
Rentas de la Propiedad	3,62	3,63	3,65	3,66	3,68
Jubilaciones	8,10	8,09	8,14	8,14	8,18
Transferencias (sin Gasto)	5,46	5,49	5,50	5,52	5,55
TOTAL sin arriendo imputado	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**

Fuente: VII EPF

La Tabla 55 muestra la estructura en porcentajes que complementa al anterior y permite señalar, por ejemplo, que el método EM no restringido imputa ingresos de la ocupación más bajos en promedio (82,59%), siendo su opuesto el método Hot-Deck (82,82%).

El promedio del porcentaje para las jubilaciones cae con cualquier método de imputación, manteniendo la relación cercana entre Hot-Deck y Heckman. En montos, la estimación de estos ingresos es más semejante entre métodos que los ingresos laborales, ya que la diferencia entre uno y otro es aproximadamente 2.800 pesos.

El error cuadrático promedio (MSE) permite establecer la medida en la que el modelo no se ajusta a la información. Otra conclusión se puede obtener al comparar los diferentes MSE calculados para cada método de imputación, proporcionando una forma para elegir el mejor método: un MSE mínimo en general indicará el mejor método, ya que un MSE=0 significa que el método predice la no respuesta con una precisión perfecta. La principal crítica a este modo de evaluación es que el MSE coloca mayor peso a los errores grandes que en los pequeños (Bosch, 1993). Al calcular la raíz cuadrada del MSE se obtiene una buena medida de precisión.

$$MSE = \frac{\sum \ Errores\ al\ cuadrado}{N\ PERSONAS}$$

Entonces, para proveer otra medida y otro punto de vista se calculó el error absoluto promedio (MAE); el que convierte al error en una sola dirección (positiva). De esta forma, este estadístico del error no tiene el problema de ponderación de errores grandes del MSE. Algunos autores señalan como desventaja que se exprese en la misma unidad que la variable (Aguirre, 1994), aunque esto ayuda a la interpretación. Al igual que el anterior, un menor estadístico señala al mejor método.

$$MAE = \frac{\sum \ Errores\ absolutos}{N\ PERSONAS}$$

Para calcular la raíz cuadrada del MAE y MSE efectivos para cada método se crearon dos tipos de muestras para poder recrear la no respuesta, la primera con un 10% de no respuesta por fuente y otra con el 20% para así poder comparar los estadísticos. Cada columna adyacente al estadístico señala el método con el dato menor y, por lo tanto, el recomendado por las reglas de decisión antes explicadas.

Tabla 56: Raíz de la suma de errores absolutos y al cuadrado por método y fuente

		Muestra del 10%				Muestra del 20%			
		$\sqrt{MAE}$		$\sqrt{MSE}$		$\sqrt{MAE}$		$\sqrt{MSE}$	
A	HECKMAN	0,351	-	16,809	*	0,274	-	12,755	-
	Imputación Múltiple	0,297	*	17,313	-	0,211	*	12,621	*
	HOT DECK	0,426	-	20,659	-	0,306	-	15,948	-
H	HECKMAN	1,677	-	47,662	-	1,178	-	32,146	-
	Imputación Múltiple	1,213	*	39,853	*	0,873	*	28,795	*
	HOT DECK	2,121	-	66,584	-	1,267	-	38,541	-
CP	HECKMAN	0,668	-	32,655	*	0,334	*	44,444	-
	Imputación Múltiple	0,541	*	33,381	-	0,370	-	44,038	*
	HOT DECK	0,985	-	50,102	-	0,609	-	44,679	-
PI	HECKMAN	1,553	-	85,505	*	0,843	*	56,675	-
	Imputación Múltiple	1,220	*	89,127	-	0,893	-	56,664	*
	HOT DECK	2,695	-	165,511	-	1,517	-	76,691	-

NOTA: A: Asalariados, H: Honorarios, CP: Trabajadores por Cuenta Propia y PI: Profesionales Independientes.
(*) Se marcan con un asterisco los valores del estadístico de comparación que son menores por fuente de ingreso
Fuente: VII EPF

Dada la principal crítica a los estadísticos del error[54], se analiza la distribución del error para conocer su comportamiento. Como se observa en la Tabla 57, el promedio menor del error se registra con el método Hot-Deck. Si se comparan los mínimos y máximos de los errores, se evidencia la crítica ya que para Hot-Deck se registraron MSE más altos y en contraposición se observan errores similares comparando los percentiles 10 y 90.

54 Los estadísticos asignan mayor peso a los errores grandes que en los pequeños.

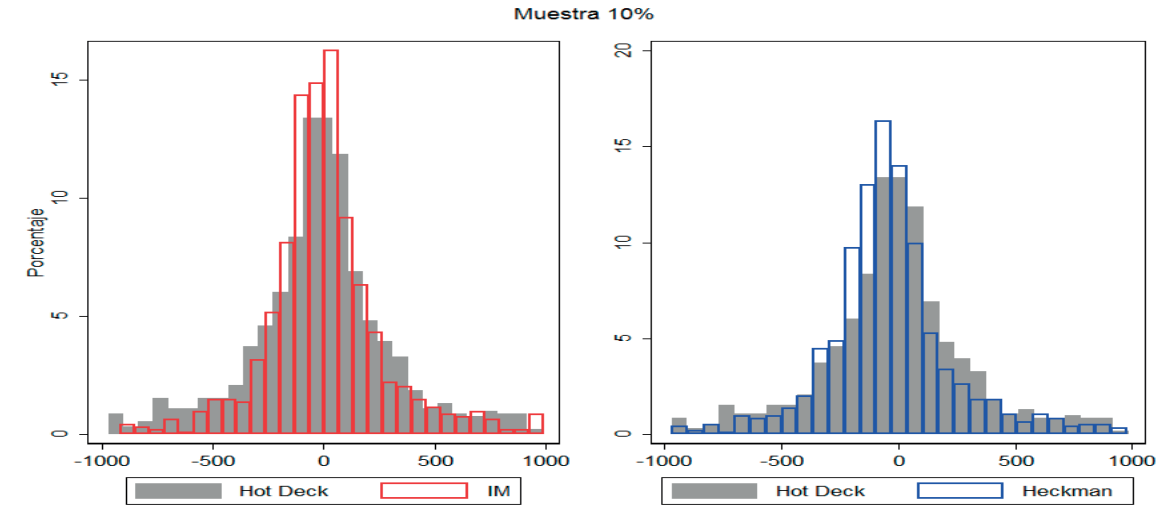
Tabla 57: Distribución del error por método y fuente

		Muestra del 10%					Muestra del 20%				
		PRO-MEDIO	MIN	p25	p75	MAX	PRO-MEDIO	MIN	p10	p75	MAX
A	HECKMAN	34,0	-1.779	-167	92	4.034	-8,4	-2.698	-192	78	11.402
	IM	95,2	-1.668	-109	142	4.123	89,6	-1.876	-112	146	11.564
	HOT DECK	-5,0	-3.760	-191	155	4.601	-24,6	-5.451	-172	138	11.559
H	HECKMAN	-34,3	-943	-191	101	899	-58,7	-969	-204	74	1.404
	IM	54,2	-461	-96	125	863	39,8	-549	-109	156	1.524
	HOT DECK	-109,0	-1.310	-321	50	1.128	-53,3	-1.875	-226	130	1.070
CP	HECKMAN	32,2	-1.756	-111	112	3.890	218,4	-1.675	-50	153	9.217
	IM	104,5	-1.443	-66	153	4.285	197,1	-1.686	-55	146	9.256
	HOT DECK	-99,5	-4.885	-154	131	3.900	22,3	-5.700	-153	118	10.280
PI	HECKMAN	86,6	-1.745	-184	256	2.683	164,3	-1.034	-123	272	4.793
	IM	230,6	-1.608	-73	367	2.769	137,6	-1.602	-145	265	4.879
	HOT DECK	-330,7	-7.550	-450	220	2.700	-115,5	-5.182	-370	216	4.500

Fuente: VII EPF

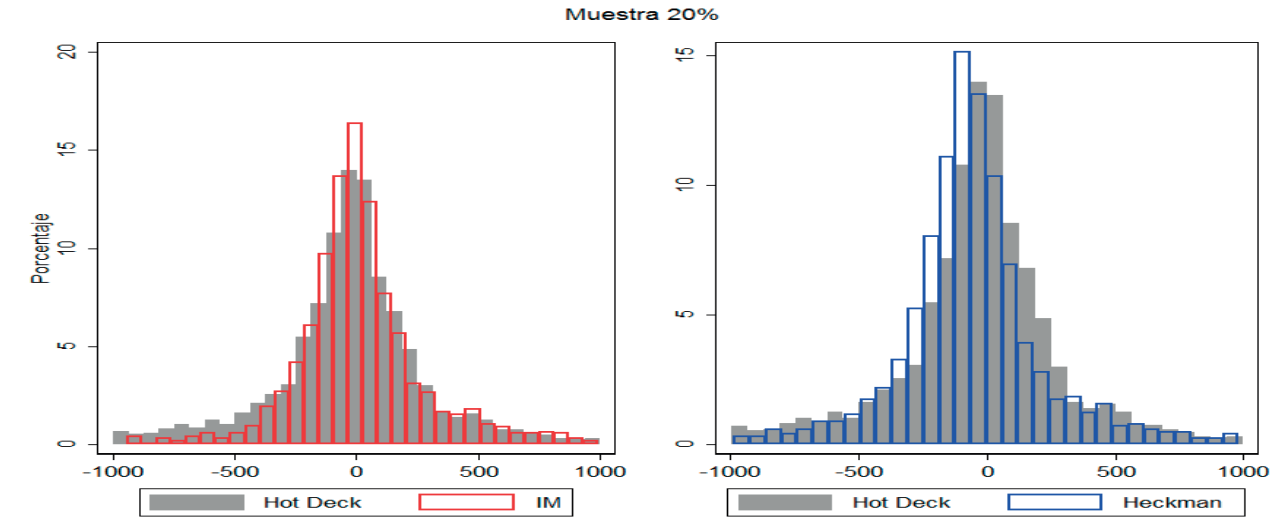
Las siguientes gráficas comparan los histogramas de los errores[55]. La transposición entre dos métodos permite ver lo semejantes que son en este caso con una tasa de no respuesta simulada del 10% y en la siguiente del 20%.

Gráfico 20: Distribución completa del error por método de imputación y tamaño de la muestra



Fuente: VII EPF

55 En el Anexo I se encuentran los histogramas completos y para ambas submuestras.



Fuente: VII EPF

Para finalizar y completar la validación de los métodos, se comparan con dos encuestas que se realizan en Chile en un período cercano a la EPF. Tanto NESI como Casen capturan los ingresos del último trimestre del 2011, que es sólo una parte del tiempo que captura la EPF, pero ambas permiten seleccionar el ámbito geográfico (zonas urbanas de 61 comunas).

Tabla 58: Comparación con otras encuestas y estructura

Fuente	VII EPF		NESI 2011		CASEN 2011	
De la Ocupación	**732.711,4**	**82,82**	**650.075,1**	**82,08**	**762.113,4**	**83,54**
Sueldos y Salarios	535.046,2	60,5	485.072,3	61,2	512.542,2	56,2
Empleadores	51.098,1	5,8	64.342,4	8,1	56.674,9	6,2
Cuenta Propia	105.064,3	11,9	68.788,0	8,7	167.070,7	18,3
Ingresos de Otros Trabajos	41.502,8	4,7	31.872,3	4,0	25.825,6	2,8
De Otras Fuentes	**152.031,4**	**17,18**	**141.907,0**	**17,92**	**150.179,0**	**16,46**
Rentas de la Propiedad	32.008,0	3,6	31.707,4	4,0	47.162,9	5,2
Jubilaciones	71.689,4	8,1	72.947,6	9,2	47.953,0	5,3
Transferencias	48.334,0	5,5	37.252,1	4,7	55.063,2	6,0
TOTAL	**884.742,9**	**100**	**791.982,1**	**100**	**912.292,4**	**100**

Fuente: VII EPF

Los resultados muestran coherencia, tanto en estructura como en niveles entre las tres encuestas. El ingreso promedio alcanzado por EPF, sin incluir el arriendo imputado, es de 884.743 pesos; mientras que NESI capturó 791.982 pesos y la Casen, 912.292 pesos (ajustada a CCNN), considerando que la selección de las comunas y la zona urbana es una aproximación para la comparación. En estructura, se evidencia aún más la similitud de la composición por fuente de ingresos.

VI CONCLUSIONES

En este documento, el equipo técnico de la VII Encuesta de Presupuestos Familiares presenta los métodos de imputación estudiados e implementados en la base de datos acumulada, lo que da cuenta y transparenta los procesos involucrados en la toma de decisiones del proyecto respecto a la imputación de gastos individuales diarios de aquellas libretas contestadas de forma parcial, como a la imputación de los ingresos del trabajo principal e ingresos por jubilaciones que presentaron no respuesta en montos.

Es fundamental presentar las metodologías utilizadas para la imputación de los datos en pro de transparentar todos los procesos contenidos en la elaboración de estadísticas oficiales con el fin de asegurar la calidad de las estadísticas. Además, en un sentido más amplio este tipo de documentos que implementan metodologías contribuyen a la discusión, uso y posicionamiento de la institución en estos temas.

Respecto al gasto, si bien se testearon tres diferentes métodos de imputación de LGI contestadas de forma parcial, los tres métodos se comportaban de forma muy similar, produciendo estructuras similares del gasto a nivel de divisiones, además los distintos deciles de ingreso y de gasto respecto a la participación en el gasto total no variaba significativamente entre los distintos métodos.

En la evaluación de los distintos métodos de imputación de gastos individuales se consideraron además estadísticos básicos como el promedio, el coeficiente de variación, los valores mínimos y máximos además de la distribución del gasto individual a través de los distintos deciles. El método escogido finalmente fue el método de ajuste por Factor No Respuesta que es el método de imputación que se aplicó en la versión pasada de la encuesta, que además es un método fácil de entender y es ampliamente admitido en la experiencia internacional para este tipo de no respuesta parcial en los gastos registrados diariamente.

Para el concepto de ingresos del trabajo y de jubilaciones se probaron cuatro métodos de imputación. Conjuntamente, sobre la base de datos completa se seleccionó dos muestras, una del 10% y otra del 20% por fuente de los ingresos de la actividad principal. Los datos de las personas seleccionadas fueron convertidos a datos faltantes, se replicaron tres métodos de imputación y así se estimaron los ingresos brutos y a continuación los errores efectivos. Los estadísticos, raíz cuadrada de la suma del error cuadrático promedio y raíz cuadrada de la suma del error absoluto promedio, ayudan a la elección del método que menos se equivoca. El problema con ambos indicadores fueron los datos extremos, ya que ponderan más a los errores más grandes. Este ejercicio nos indicó la cercanía entre el método de regresión Heckman y de imputación múltiple, los que proveen los errores menores y por lo tanto son elegibles.

Sin embargo, para los ingresos se utilizó la imputación mediante el método Hot-Deck. En la elección, se consideró la distribución original de los datos (percentiles 10/90 e intercuartiles), además del coeficiente de variación, comparándolos con los resultados de cada método de imputación, llegando a la conclusión que el método de Hot-Deck era el indicado, pues desde un punto de vista global es un método parsimonioso y altera en menor medida los estadísticos de tendencia central y dispersión.

El método Hot-Deck es un método sencillo que entrega buenos resultados con la actual tasa de no respuesta parcial en ingresos, considerando además que las características de quienes son donantes se acomodan a las características de los que tienen montos faltantes. Sin embargo, el tamaño de la muestra es determinante, ya que para las personas que no reportan los montos de sus ingresos, existe un número finito y acotado de donantes que pueden compartir características.

El impacto en los procesos de validación, depuración y más concretamente los de imputación tiene un efecto en las conclusiones a las que llegan los usuarios de las bases de datos y de los productos publicados por la institución. Por lo que, tal como lo menciona Medina (2012), es necesario reconocer que en la actualidad el interés y uso de las encuestas de ingresos y gastos rebasa el ámbito de las cuentas nacionales y del IPC. Se espera que este documento de estudio satisfaga las necesidades de distintos usuarios y tipos de análisis sobre documentación de procesos y sobre las decisiones elegidas.

Una recomendación del equipo técnico de la EPF para el desarrollo de una nueva versión de esta encuesta es estudiar diferentes alternativas de imputación de libretas de gastos individuales. Si bien es cierto que la imputación de libretas completas no se ha realizado en versiones anteriores de la encuesta y en esta versión decidió no realizarse, existe información suficiente para probar distintas alternativas de

imputación de libretas parciales y el impacto que éstas pueden tener en la estructura del gasto de los hogares.

Tal como apunta la teoría sobre imputación de datos, distintos modelos de imputación aplicados a bajos porcentajes de datos a imputar generan diferencias poco significativas entre los distintos métodos de imputación. Esto es avalado por los resultados obtenidos en la imputación de datos en la VII EPF, tanto para gasto como para ingresos. Los distintos métodos testeados en la encuesta generaron resultados similares.

Al revisar el detalle de los estados de las libretas de gastos individuales, se observa que aquellos informantes que recibieron la LGI de forma directa presentan una tasa de respuesta de LGI notablemente más alta que aquellos informantes que recibieron la LGI de forma indirecta (a través de un miembro del hogar). De esta situación se desprende la recomendación para una próxima EPF de intentar entregar las LGI de forma directa a todos los integrantes del hogar buscando generar el compromiso activo de parte de los informantes de participación en la encuesta a través del llenado de dichas libretas.

En general, los métodos de imputación que utilizan clusters de tamaños pequeños generan estimaciones más precisas, mientras que clusters homogéneos pueden mejorar las estimaciones para pequeños grupos de población al realizar imputaciones entre informantes lo más parecidos posible.

Como conclusión general se puede extraer que ninguno de los métodos de imputación tratados en este documento es preferible a otro per se, ya que la decisión sobre qué método de imputación utilizar depende de la estructura y distribución de los datos con que se esté trabajando. Es importante recalcar que explicitar los métodos probados, la forma en que se tomó la decisión sobre el método a utilizar y la estructura y comportamiento de los datos con y sin imputación generan confianza y entregan herramientas sobre la naturaleza de los datos y como éstos fueron trabajados.

Si bien en este documento se explicitaron los métodos que fueron utilizados en la VII EPF, es relevante tener en cuenta que lo ideal es tener un nivel de imputación lo más bajo posible y en ese sentido la VII EPF concentró muchos esfuerzos para intentar disminuir los errores no muestrales asociados al levantamiento de la encuesta.

Se debe resaltar que los procesos tienden a ser de uso continuo en las versiones de cada encuesta, sin embargo, las revisiones metodológicas de mejora dejan un espacio para una re-evaluación del método elegido tanto para ingresos como para gastos, ya que el ajuste depende de la información y la precisión alcanzada, esto en un contexto de desarrollo institucional donde se procura mantener la comparabilidad.

BIBLIOGRAFÍA

Aguirre A. (1994), *"Introducción al tratamiento de series temporales: Aplicación a las Ciencias de la Salud"*, Ediciones Díaz de Santos.

Alfaro R. y M. Fuenzalida (2009). *"Imputación Múltiple Encuestas Microeconómicas"*. Cuadernos de economía, VOL. 46 (noviembre), PP. 273-288.

American Association for Public Opinion Research [AAPOR] (2011), *"Standard Definitions, Final Dispositions of Case Codes and Outcome Rates for Surveys"*.

Australian Bureau of Statistics [ABS] Australia (2006). *"Household Expenditure Survey and Survey of Income and Housing. User Guide 2003-2004"*. Australia, Junio.

[ABS] Australia (2012). *"Household Expenditure Survey and Survey of Income and Housing. User Guide 2009-2010"*.

Banco Central de Chile [BCCH] (2011). *"Encuesta Financiera de hogares: metodología y principales resultados EFH 2007"*.

Belsley, D. A., E. Kuh, y R. E. Welsh (1980). *"Regression Diagnostics"*. New York, NY: John Wiley & Sons, Inc.

Bisquerra, R. (1987). *"Introducción a la estadística aplicada a la investigación educativa"*. Barcelona. Editorial PPU (Promociones y Publicaciones Universitarias, S.A.).

Bosch A. (editor) (1993). *"Estadística"*, 2da edición.

Bureau of Labor Statistics [BLS] Estados Unidos (2008). Consumer Expenditure Survey, Anthology 2008. Reporte anual de la encuesta. Estados Unidos.

Dayal N., J. Gomulka, L. Mitton, H. Sutherland y R. Taylor [Dayal et al.] (2000), Enhancing Family Resources Survey Income Data with Expenditure Data from the Family Expenditure Survey: Data Comparisons. Department of Applied Economics, University of Cambridge.

Departamento Administrativo Nacional de Estadística [DANE] Colombia (2009). Metodología Encuesta Nacional Ingresos y Gastos, 2006-2007.

Departamento Nacional de Planeación, Colombia (2003). *"Evaluación de dos métodos de imputación de ingresos totales en la población en edad de trabajar con encuestas de hogares"* http://www.eclac.cl/deype/mecovi/docs/TALLER14/12.pdf

Eurostat (2003), Household Budget Surveys in the EU, Methodology and recommendations for harmonization – 2003.

Felcman, Kidyba y Ruffo (2003) *"Medición del ingreso laboral: ajustes a los datos de la encuesta permanente de hogares para el análisis de la distribución del Ingreso (1993–2002)"*.

Heckman, J. (1979) *"Sample selection bias as a especification error"*, Econométrica 47.

Huisman, M. (2010). Missing Data, Mechanisms and Determinants. Presentación en Simposio de la Asociación de Educación (Education Partnership EPP), disponible en: http://www.gmw.rug.nl/~huisman/md/EPP2_2010.pdf

IBM-SPSS (2012), *"Missing values 22"*, Manual de usuario del módulo análisis de datos faltantes del programa SPSS.

Institut national de la statistique et des études économiques [Insee] France (2007) Sources et méthodes. Enquête Budget de famille 2006.

Instituto Brasileiro de Geografia e Estatística [IBGE] Brasil (2010) Pesquisa de Orçamentos Familiares 2008-2009. Despesas, Rendimentos e Condições de Vida.

Instituto Nacional de Estadística [INE] España (2008).Principales características. Encuesta de Presupuestos Familiares. Base 2006. Diciembre 2008

[INE] España (2010). Módulo 4. Encuesta de Presupuestos Familiares, Fundamentos y Prácticas de las Encuestas a los Hogares, 3ª Edición. Instituto Nacional de Estadística España.

Instituto Nacional de Estadística y Censos [INEC] Costa Rica (2006). Metodología Encuesta Nacional de Ingresos y Gastos de los Hogares 2004.

Instituto Nacional de Estadística, Geografía e Informática [INEGI] México (2009). Encuesta nacional de ingresos y gastos de los hogares 2008. Ingresos y Gastos de los Hogares.

Instituto Nacional de Estadísticas [INE] Chile (2003), Método PRINCALS para la Clasificación Socioeconómica del Censo 2002. Miguel Guerrero.

[INE] Chile (2009), METODOLOGÍA, VI Encuesta de Presupuestos Familiares 2006-2007.

[INE] Chile (2013), METODOLOGÍA, VII Encuesta de Presupuestos Familiares. Instituto Nacional de Estadísticas de Chile.

Instituto Nacional de Estadística [INE] Portugal (2008). Inquérito às Despesas das Famílias 2005-2006.

Instituto Nacional de Estadísticas [INE] Uruguay (2006). Encuesta Nacional de Gastos e Ingresos de los Hogares 2005-2006. Metodología y Resultados.

Instituto Nacional de Estadísticas y Censos [INDEC] Argentina (2007). Encuesta Nacional de Gastos de los Hogares 2004/2005. Resumen Metodológico.

Istituto Nazionale Di Statistica [ISTAT] Italia (2011) I consumi delle famiglie. Anno 2009., Marzo.

Little, R. (1988). *"A test of Messing completely at random for multivariate data with Missing values"*. Revista American Statistical Association. Diciembre, vol. 83 No. 404, Teoría y métodos.

Little, R. J. A. y Rubin, D. B. (2002).Statistical Analysis with Missing Data, 2nd ed. Wiley, New York.

Medina F. (2012) *"Los desafíos de medición en las encuestas de ingresos y gastos en América Latina"* Seminario de Cuentas Nacionales de América Latina CEPAL. http://www.eclac.cl/deype/noticias/noticias/8/48378/2012-SemCN_CEPAL-FMedina.pdf

Medina, F. y M. Galván (2007). *"Imputación de datos: teoría y práctica"*. Serie estudios estadísticos y prospectivos N° 54. División de estadísticas y proyecciones económicas d la Comisión económica para América Latina y el Caribe.

Michael N. Mitchell (2012), A Visual Guide to Stata Graphics, Third Edition. Stata Press

Naciones Unidas, Banco Mundial, Fondo Monetario Internacional, Comisión de las Comunidades Europeas, Organización para la Cooperación y el Desarrollo Económico [UN et al.] (1993), Sistema de Cuentas Nacionales.

Naciones Unidas [UN] (2001), Clasificaciones de Gastos por Finalidades, Nueva York.

Office for National Statistics [ONS] (2010) Family Spending. A report on the Living Costs and Food Survey 2009. Reino Unido.

Oficina Internacional del trabajo [OIT] (1991), *"Clasificación internacional uniforme de ocupaciones: CIOU-88"*, Ginebra.

[OIT] (2003), Organización Internacional del Trabajo, Informe II, Estadísticas de Ingresos y Gastos de los Hogares, Decimoséptima Conferencia Internacional de Estadísticos del Trabajo.

Paulin G. y D. Ferraro, (1994). *"Imputing income in the consumer expenditure survey"*. Monthly Labor Review. Diciembre.

Perlbach de Maradona I. y M.I. Calderón (1998). Estimación del sesgo de selección para el mercado laboral de Mendoza. Documento de investigación de la Asociación Argentina de Economía política y Universidad Nacional de Cuyo. Disponible en: http://www.aaep.org.ar/anales/works/works1998/perlbach-de-maradona_calderon.pdf

Rubin D., *"Multiple imputation for nonresponse in surveys"*, Wiley Classics Library, 2004.

Schafer, J. (1997). "*Analysis of Incomplete Multivariate Data*". Chapman & Hall/CRC.

Schafer, J. y Graham, J. (2002). *"Missing Data: Our View of the State of the Art"*. Psychological Methods

Stata Corporation (2009). *"Stata multiple-imputation reference manual"*. Edición 11. Manual de imputación múltiple con Stata.

Statistics Canada (2013). *"User Guide for the Survey of Household Spending, 2011"*

Statistics South Africa (2012). *"Income and Expenditure of Households 2010/2011"*. Statistical release.

Taylor R., H. Sutherland y J. Gomulka (2001), "*Using POLIMOD to Evaluate Alternative Methods of Expenditure Imputation*", Department of applied Economics, University of Cambridge.

Von Hippel P. (2004). *"Biases in SPSS 12.0 Missing Value Analysis"*, The American Statistician, Vol. 58, No. 2.

Wang, H. (2007). *"Missing data analysis in structural equation modeling: Expectation maximization and multiple imputation methods"* The University of Alabama, ProQuest, UMI Dissertations Publishing. Número de publicación: 3313750.

Anexo A: Módulo de control de Libretas de Gastos Individuales de la Hoja de Ruta

CONTROL DE LIBRETAS DE GASTOS INDIVIDUALES (CL)

Registre el folio y anote el número de línea de todos los integrantes del hogar de acuerdo al RPH.
Marque con una "X" la persona que cumple la función de administrador de gastos del hogar en la columna correspondiente (de acuerdo a IP03 del RPH). Respecto a la entrega, anote el código correspondiente (enunciados a continuación de la tabla). Registre las fechas de entrega y retiro de las libretas. Además señale si el estado de las libretas es completo, parcial o sin registro. Independiente del estado de las libretas, en caso de rechazo, marque con una "X" en la celda correspondiente.

RECUERDE QUE DÍAS CON GASTO IGUAL A CERO SON DÍAS CON REGISTRO.

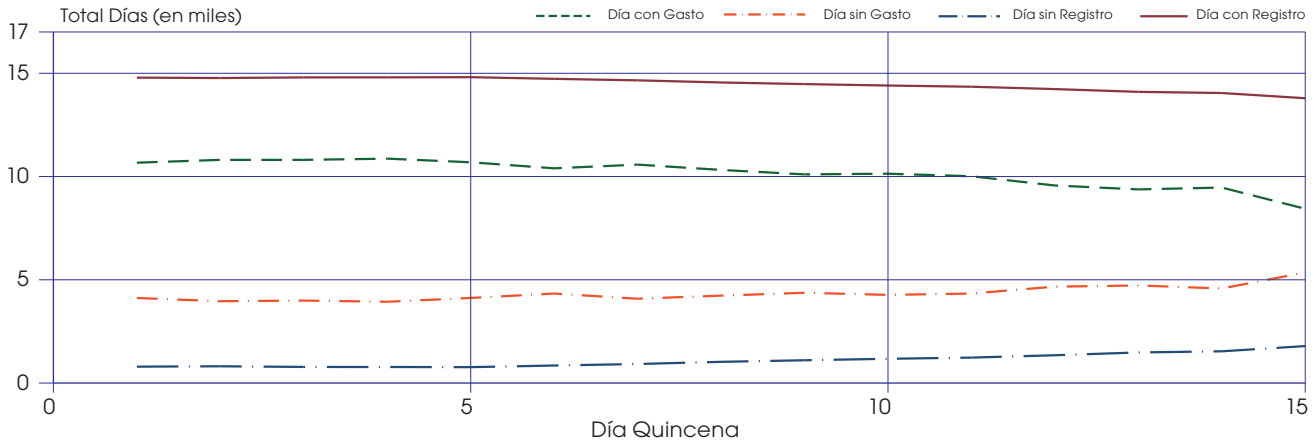
FOLIO	N° LÍNEA (RPH)	ADM. HOGAR	CÓDIGO ASIGNACIÓN	FECHA						ESTADO			RECHAZO	TOTAL DÍAS DE REGISTRO	DÍAS DE REGISTRO FIN DE SEMANA
				ENTREGA			RETIRO			COMPLETA	PARCIAL	SIN REGISTRO			
				DÍA	MES	AÑO	DÍA	MES	AÑO						
(01)	(02)	(03)	(04)	(05)	(06)	(07)	(08)	(09)	(10)	(11)	(12)	(13)	(14)	(15)	(16)

CÓDIGOS DE ASIGNACIÓN DE LIBRETA DE GASTOS INDIVIDUALES

81	Entrega de forma personal.
82	Entrega de forma indirecta (a través de un miembro del hogar).
83	No entrega, menor de 15 años (TODO MENOR DE 15 AÑOS DEBE LLEVAR ESTE CÓDIGO).
84	No entrega, persona con capacidades mentales no aptas para contestar el estudio, Ej.: Síndrome de Down, Alzheimer.
85	No entrega, persona con capacidades físicas no aptas para contestar el estudio, Ej.: postrado.
86	No entrega, persona con síndrome de adicción.
87	No entrega, miembro del hogar no ubicable en el periodo de referencia, Ej.: motivos vacacionales, trabajos fuera del hogar entre otros.

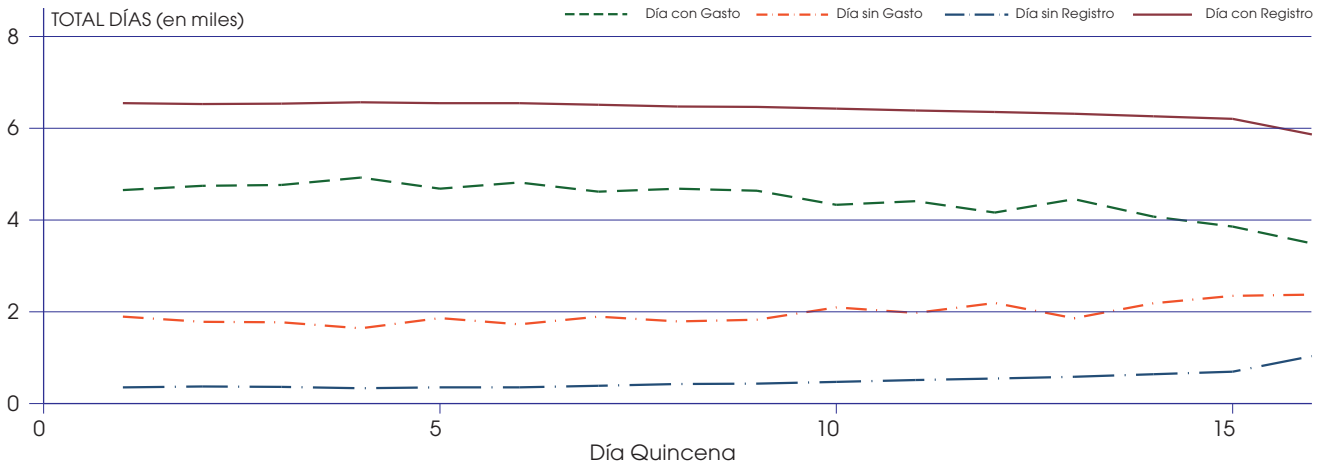
Anexo B. Tipos de registros de gastos individuales según días de la quincena (diferenciados por tipos de quincenas)

Quincenas de 15 días



Nota: sólo se consideran las 16 quincenas de 15 días
Fuente: VII EPF-INE.

Quincenas de 16 días



Nota: sólo se consideran las 7 quincenas de 16 días
Fuente: VII EPF-INE.

Anexo C. Correlaciones por tramos etarios con la variable gasto diario individual

Correlación de punto biserial Tramos Etarios	ln Gasto Individual de la Persona		
	TOTAL	Hombres	Mujeres
15-20	-0,3852	-0,369	-0,3974
20-25	-0,1519	-0,1405	-0,1578
25-30	0,0157	0,0065	0,0235
30-35	0,0967	0,0885	0,1056
35-40	0,1118	0,1098	0,113
40-45	0,1013	0,086	0,1108
45-50	0,0946	0,0689	0,11
50-55	0,0793	0,0916	0,0694
55-60	0,0467	0,0588	0,0376
60-65	0,0333	0,0388	0,0271
65-70	0,0166	0,0403	0,0005
70-75	-0,0026	0,005	-0,0074
75-80	-0,0199	-0,0143	-0,024
80-85	-0,0086	-0,0023	-0,0161
85 o más	-0,0254	-0,008	-0,0394

Fuente: VII EPF

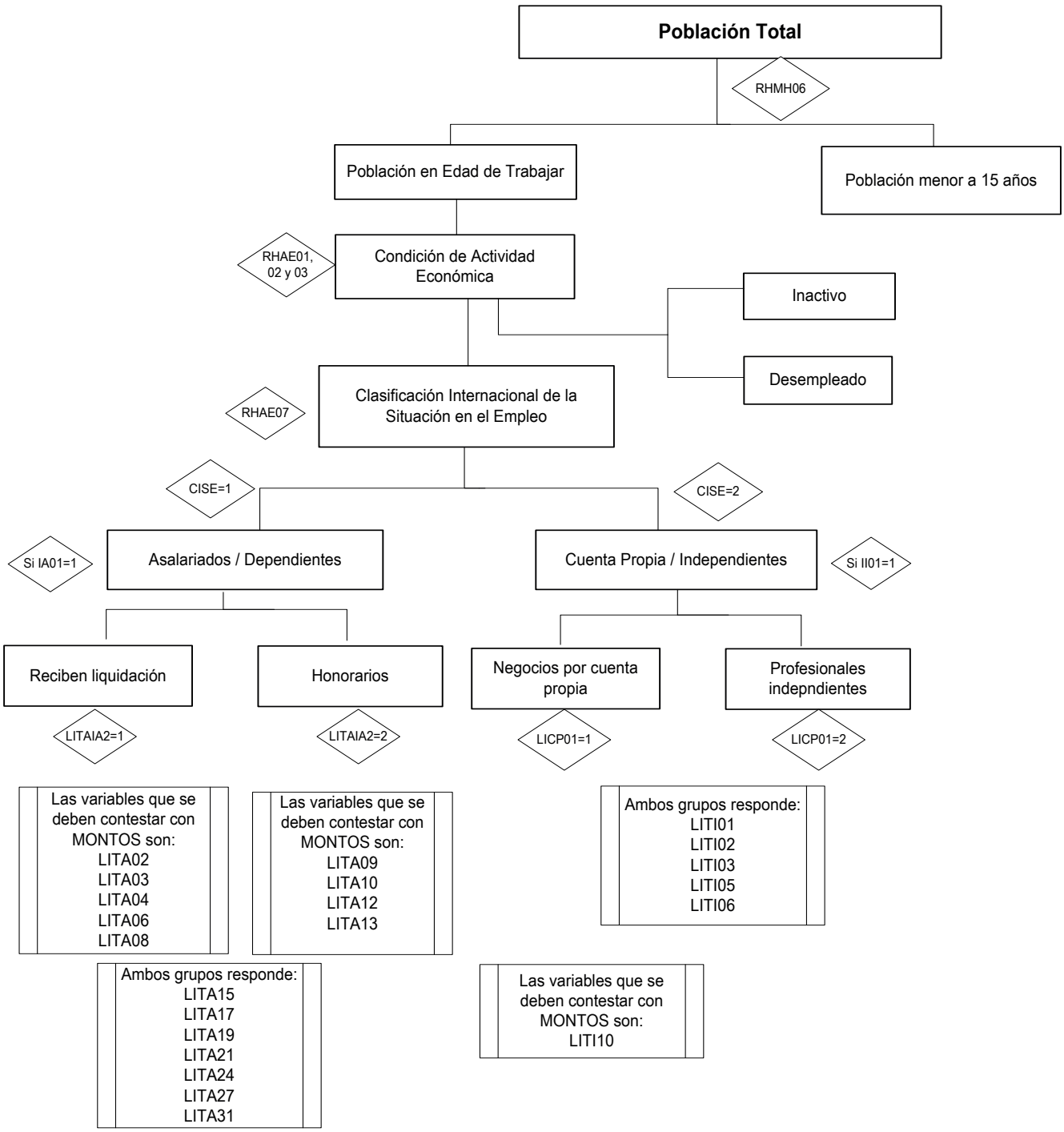
Anexo D. Categorías de respuesta de la LGI por condición de actividad económica, tramos etarios (cada 5 años) y sexo

EDAD TRAMOS	Situación de LGI. Hombres CAE=Trabaja durante la semana de referencia					Situación de LGI. Hombres CAE=Desocupado o Inactivo durante la semana de referencia				
	LGI Completa	%	LGI Parcial	%	Total	LGI Completa	%	LGI Parcial	%	Total
15-19	113	47,9%	68	28,8%	181	703	58,2%	243	20,1%	946
20-24	402	50,4%	216	27,1%	618	408	61,4%	126	18,9%	534
25-29	501	56,0%	222	24,8%	723	133	55,9%	45	18,9%	178
30-34	577	64,5%	169	18,9%	746	50	51,5%	17	17,5%	67
35-39	568	63,5%	189	21,1%	757	42	56,0%	13	17,3%	55
40-44	587	65,4%	178	19,8%	765	40	55,6%	15	20,8%	55
45-49	584	62,9%	200	21,5%	784	49	51,6%	28	29,5%	77
50-54	598	65,1%	200	21,8%	798	63	63,0%	17	17,0%	80
55-59	462	66,8%	150	21,7%	612	68	71,6%	16	16,8%	84
60-64	306	67,0%	98	21,4%	404	105	69,5%	32	21,2%	137
65-69	222	71,4%	56	18,0%	278	191	70,7%	48	17,8%	239
70-74	93	70,5%	18	13,6%	111	203	76,0%	36	13,5%	239
75-79	39	69,6%	13	23,2%	52	157	71,4%	44	20,0%	201
80-84	7	63,6%	3	27,3%	10	100	73,5%	27	19,9%	127
85 o más	5	71,4%	2	28,6%	7	59	80,8%	8	11,0%	67
No Responde		-		-	-	-	-	-	-	-
Total	**5.064**	**62,3%**	**1.782**	**21,9%**	**6.846**	**2.371**	**63,0%**	**715**	**19,0%**	**3.086**

EDAD TRAMOS	Situación de LGI. Mujeres CAE=Trabaja durante la semana de referencia					Situación de LGI. Mujeres CAE=Desocupado o Inactivo durante la semana de referencia				
	LGI Completa	%	LGI Parcial	%	Total	LGI Completa	%	LGI Parcial	%	Total
15-19	96	61,9%	49	31,6%	145	886	63,3%	295	21,1%	1.181
20-24	371	59,7%	153	24,6%	524	627	63,7%	213	21,6%	840
25-29	539	67,0%	184	22,9%	723	347	73,1%	75	15,8%	422
30-34	589	72,8%	158	19,5%	747	255	75,2%	61	18,0%	316
35-39	592	74,6%	140	17,6%	732	318	81,1%	59	15,1%	377
40-44	639	74,7%	170	19,9%	809	352	79,6%	76	17,2%	428
45-49	682	76,4%	163	18,3%	845	372	79,0%	86	18,3%	458
50-54	617	75,2%	162	19,8%	779	397	79,9%	74	14,9%	471
55-59	414	76,4%	104	19,2%	518	351	77,3%	75	16,5%	426
60-64	229	77,6%	55	18,6%	284	422	80,4%	87	16,6%	509
65-69	116	77,3%	23	15,3%	139	398	77,9%	93	18,2%	491
70-74	56	72,7%	19	24,7%	75	348	78,6%	69	15,6%	417
75-79	14	63,6%	7	31,8%	21	261	75,7%	58	16,8%	319
80-84	6	75,0%	2	25,0%	8	208	76,8%	40	14,8%	248
85 o más	4	66,7%	2	33,3%	6	104	72,7%	20	14,0%	124
No Responde		-		-	-	1	100%	-	-	1
Total	**4.964**	**72,5%**	**1.391**	**20,3%**	**6.355**	**5.647**	**73,4%**	**1.381**	**18,0%**	**7.028**

*Se consideran parciales aquellas LGI que tengan al menos un día de registro.
Nota: Se excluye del cuadro 1 informante que no declaró condición de actividad económica.
Fuente: VII EPF

Anexo E. Flujo para el cálculo de la tasa de no respuesta por categoría de ingresos del trabajo



Anexo F. Correlaciones de algunas variables con el ingreso (en logaritmos naturales).

Macrozona \ Ingreso Laboral	Correlación	%
Norte	0,0716	16,33
Centro	-0,0581	28,71
Sur	0,0348	12,74
Metropolitana	-0,0238	42,22

Fuente: VII EPF

Estrato Económico \ Ingreso Laboral	Correlación	%
Bajo	-0,188	27,09
Medio	-0,1081	57,37
Alto	0,3782	15,54

Fuente: VII EPF

Rango de Edad \ Escolaridad (1)	Correlación	%
15 a 24 años	-0,0011	10,89
25 a 34 años	0,1838	22,47
35 a 44 años	0,0697	23,45
45 a 54 años	-0,0883	24,57
Más de 55 años	-0,1744	18,63

(1) Se toma la escolaridad como variable continua
Fuente: VII EPF

Rango de Edad \ Escolaridad (1)	Asalariados	Honorarios	Negocios Cuenta Propia	Profesionales Independientes
15 a 24 años	-0,0377	-0,0499	0,0928	-0,0506
25 a 34 años	0,1795	0,2085	0,1121	0,2102
35 a 44 años	0,061	-0,0526	0,0946	0,1126
45 a 54 años	-0,0918	-0,0817	-0,0293	-0,0769
Más de 55 años	-0,1492	-0,0674	-0,1766	-0,1741

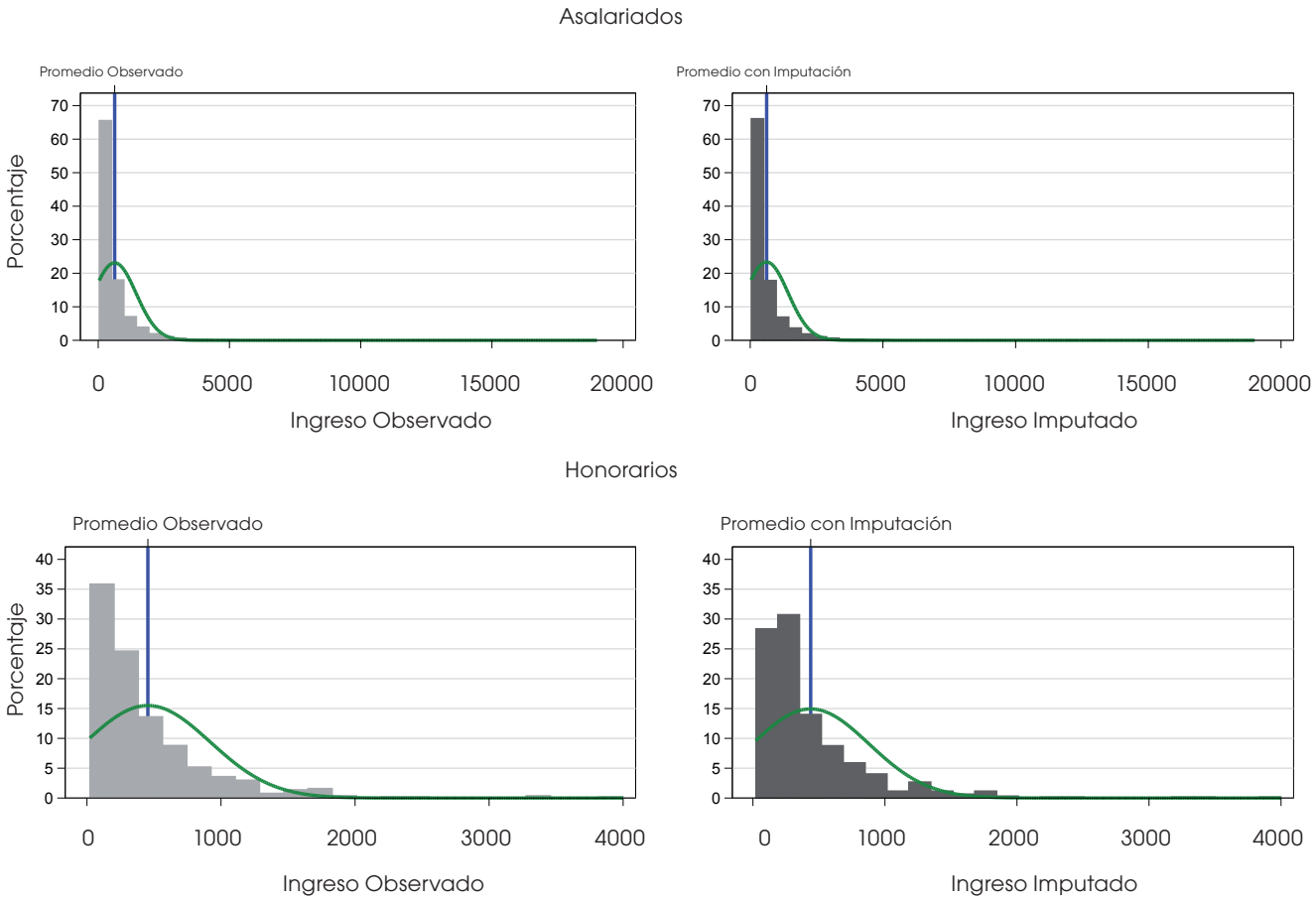
(1) Se toma la escolaridad como variable continua
Fuente: VII EPF

Anexo G. Ámbito geográfico de la VII Encuesta de Presupuestos Familiares

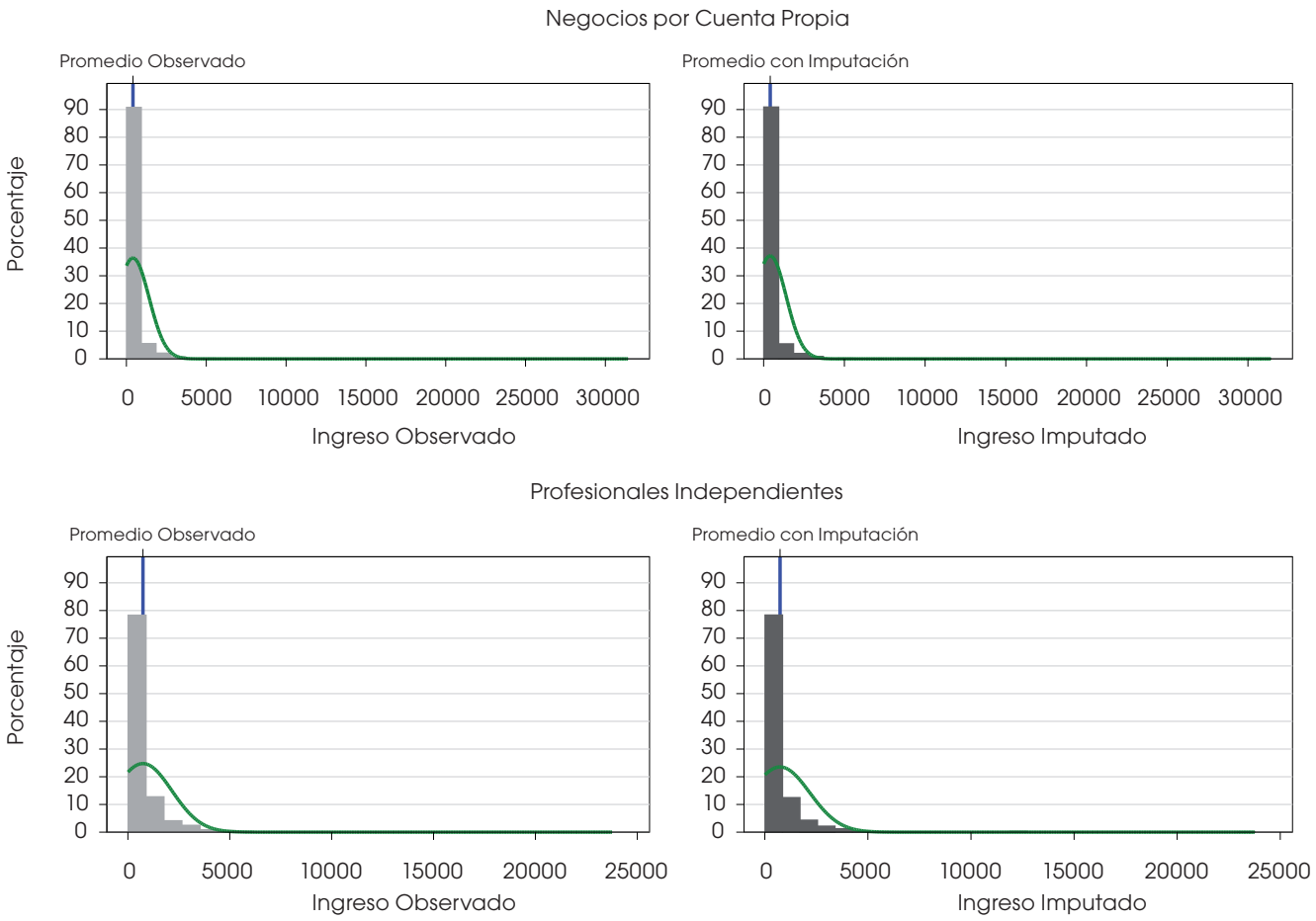
ÁMBITO GEOGRÁFICO VII ENCUESTA DE PRESUPUESTOS FAMILIARES			
Nº	REGIÓN	CAPITAL	COMUNAS INCLUIDAS
1	Tarapacá	Iquique	Iquique
			Alto Hospicio
2	Antofagasta	Antofagasta	Antofagasta
3	Atacama	Copiapó	Copiapó
4	Coquimbo	La Serena	La Serena
			Coquimbo
5	Valparaíso	Valparaíso	Valparaíso
			Viña del Mar
			Concón
			Quilpué
			Villa Alemana
6	O'Higgins	Rancagua	Rancagua
7	Maule	Talca	Talca
8	Bio-Bío	Concepción	Concepción
			Chiquayante
			Penco
			San Pedro de la Paz
			Talcahuano
			Hualpén
9	La Araucanía	Temuco	Temuco
			Padre Las Casas
10	Los Lagos	Puerto Montt	Puerto Montt
11	Aysén	Coyhaique	Coyhaique
12	Magallanes	Punta Arenas	Punta Arenas
13	Metropolitana	Santiago	Comunas Provincia de Santiago
			Puente Alto
			San Bernardo
			Padre Hurtado
14	Los Ríos	Valdivia	Valdivia
15	Arica y Parinacota	Arica	Arica

Anexo H. Comparación de las distribuciones antes y después de cada método de imputación por fuente laboral desagregada

INGRESOS IMPUTADOS POR HOT DECK

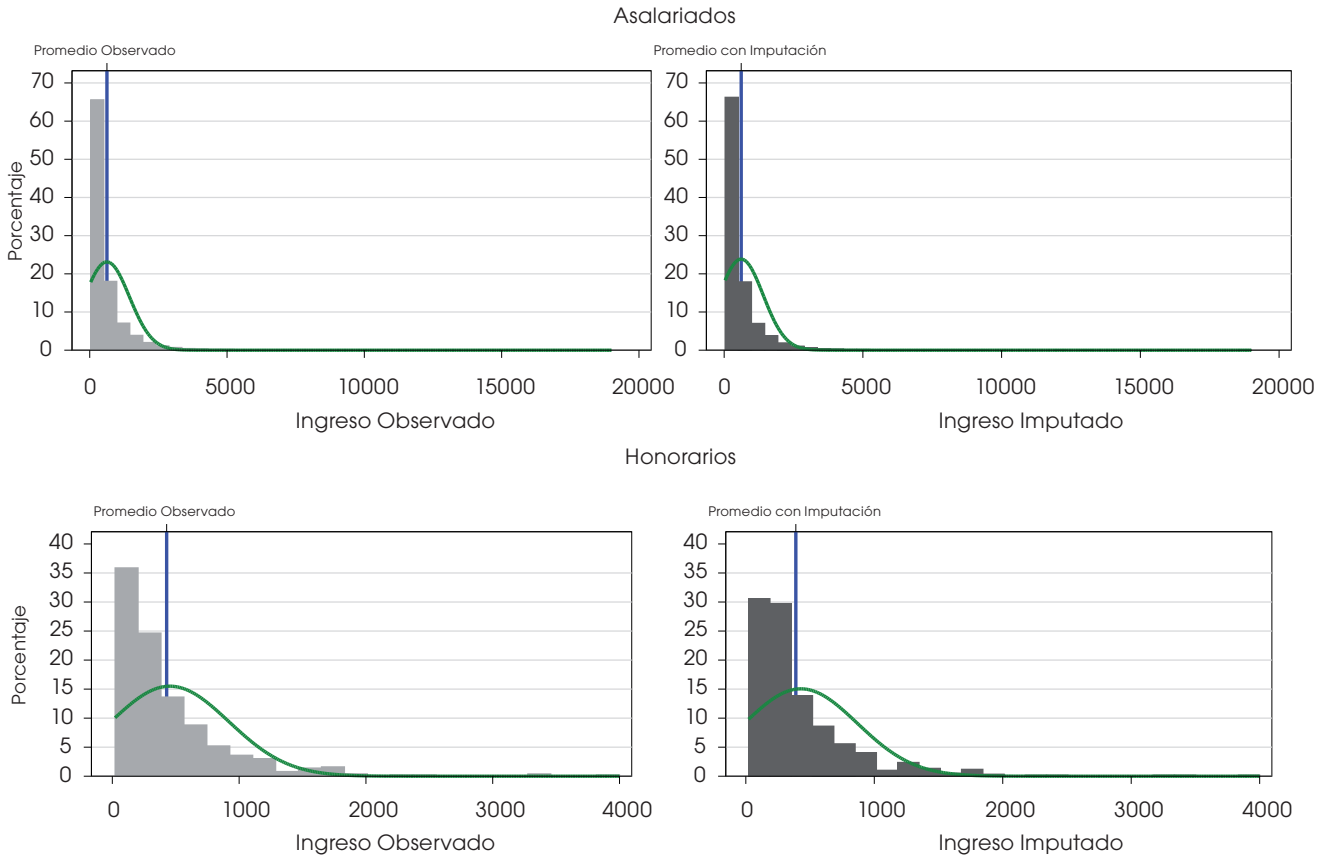


Fuente: VII EPF-INE.



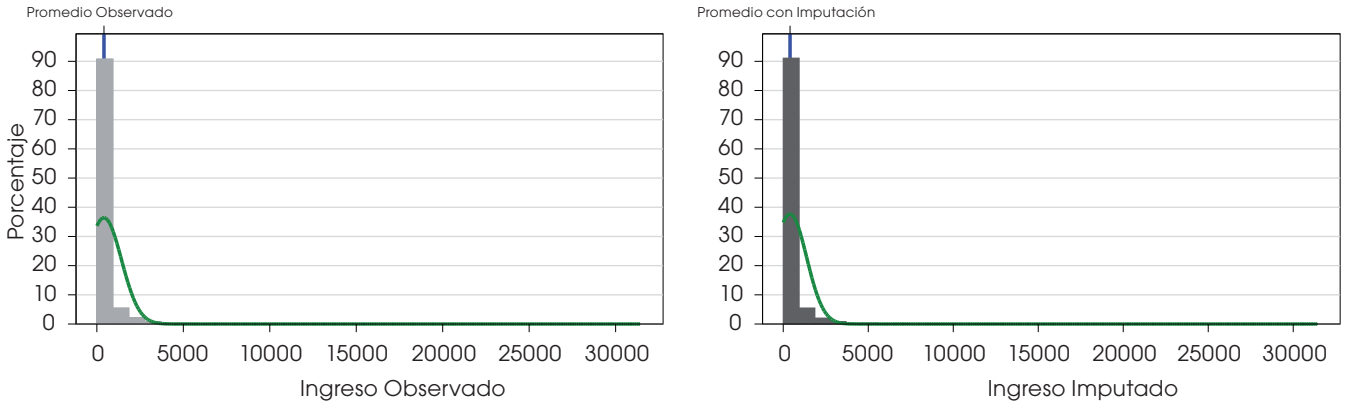
Fuente: VII EPF-INE.

INGRESOS IMPUTADOS POR HECKMAN

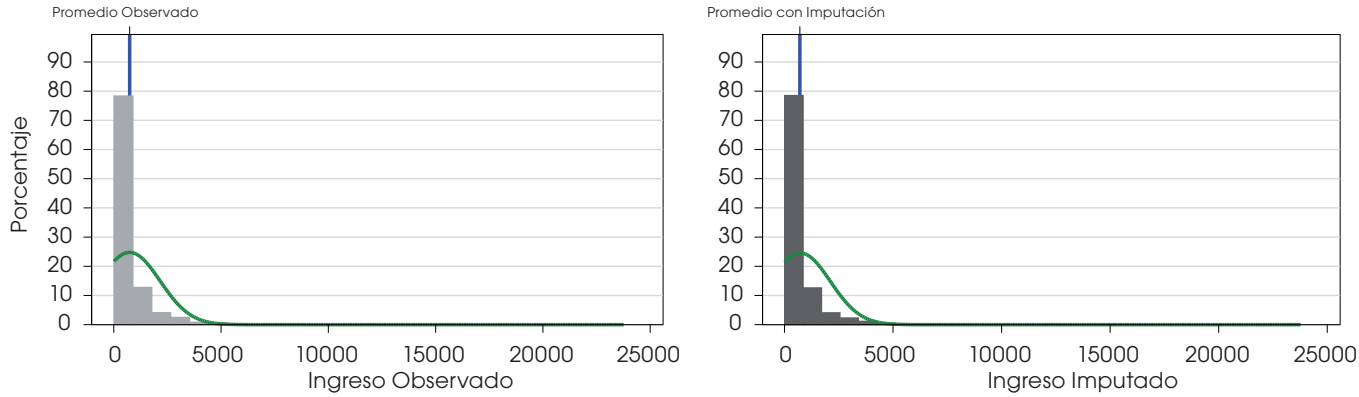


Fuente: VII EPF-INE.

Negocios por Cuenta Propia

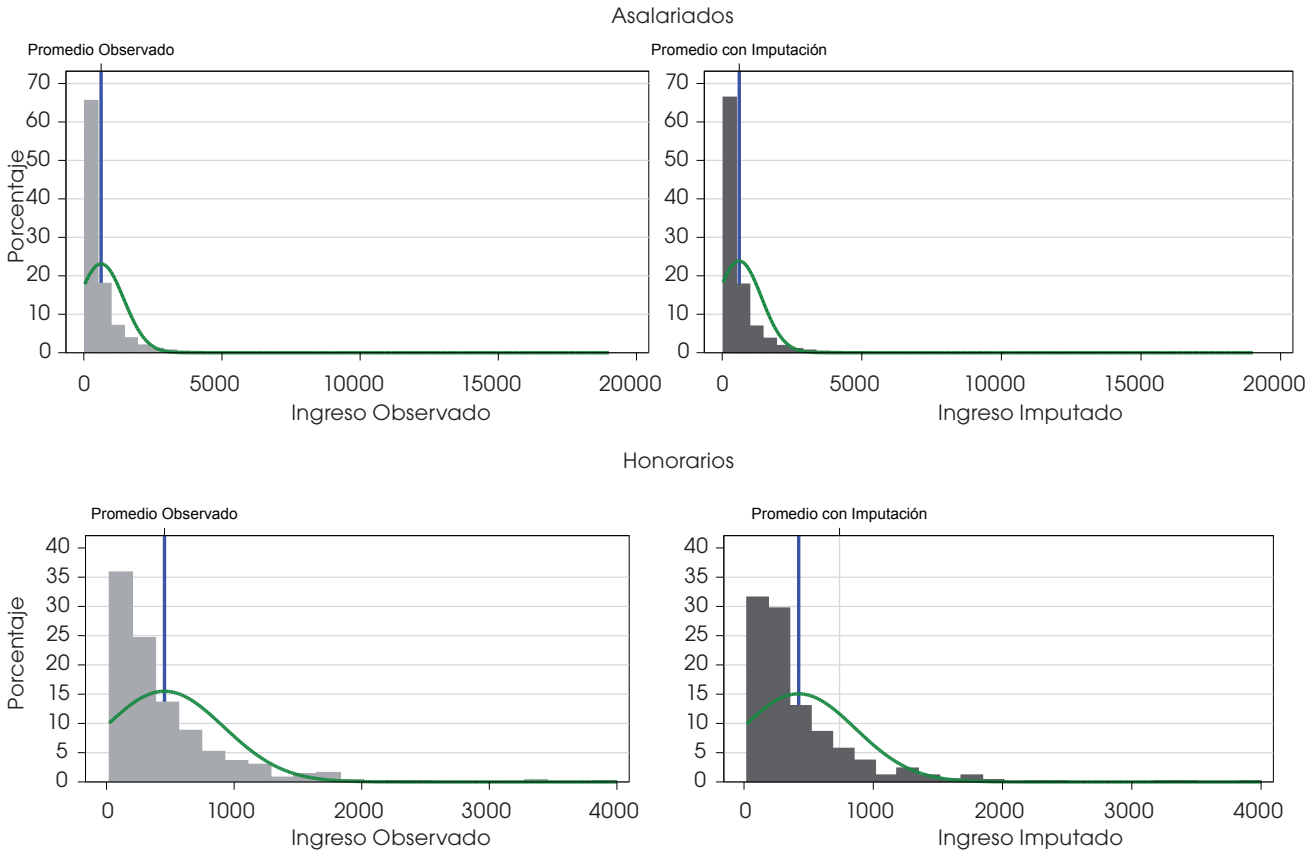


Profesionales Independientes



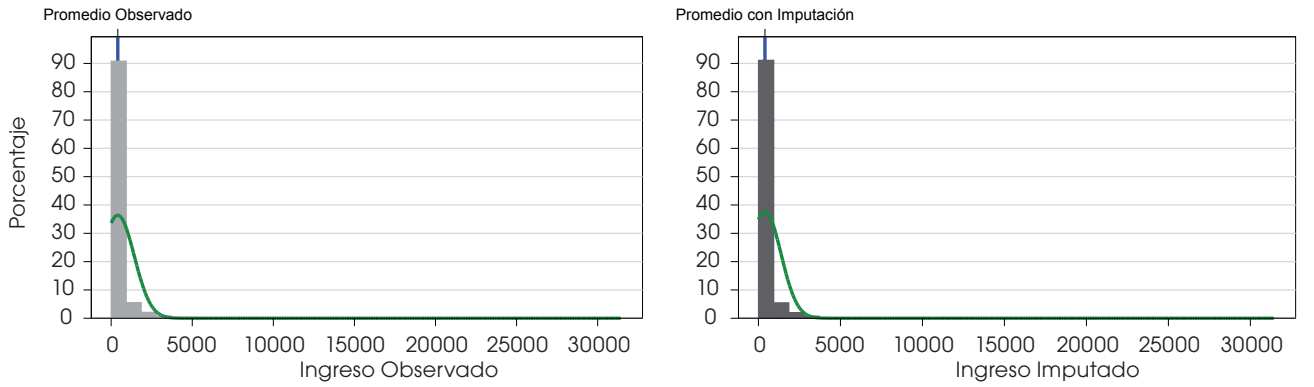
Fuente: VII EPF-INE.

INGRESOS IMPUTADOS POR IMPUTACIÓN MÚLTIPLE

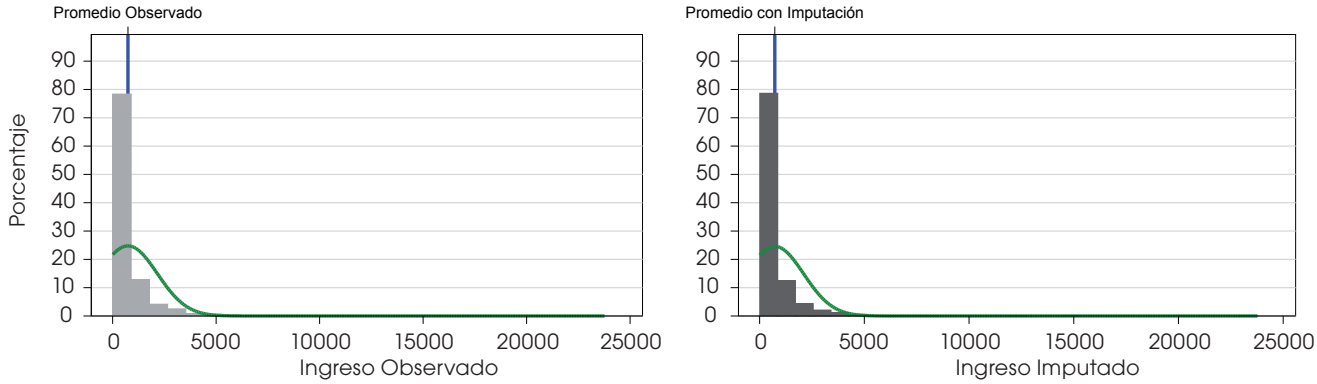


Fuente: VII EPF-INE.

Negocios por Cuenta Propia

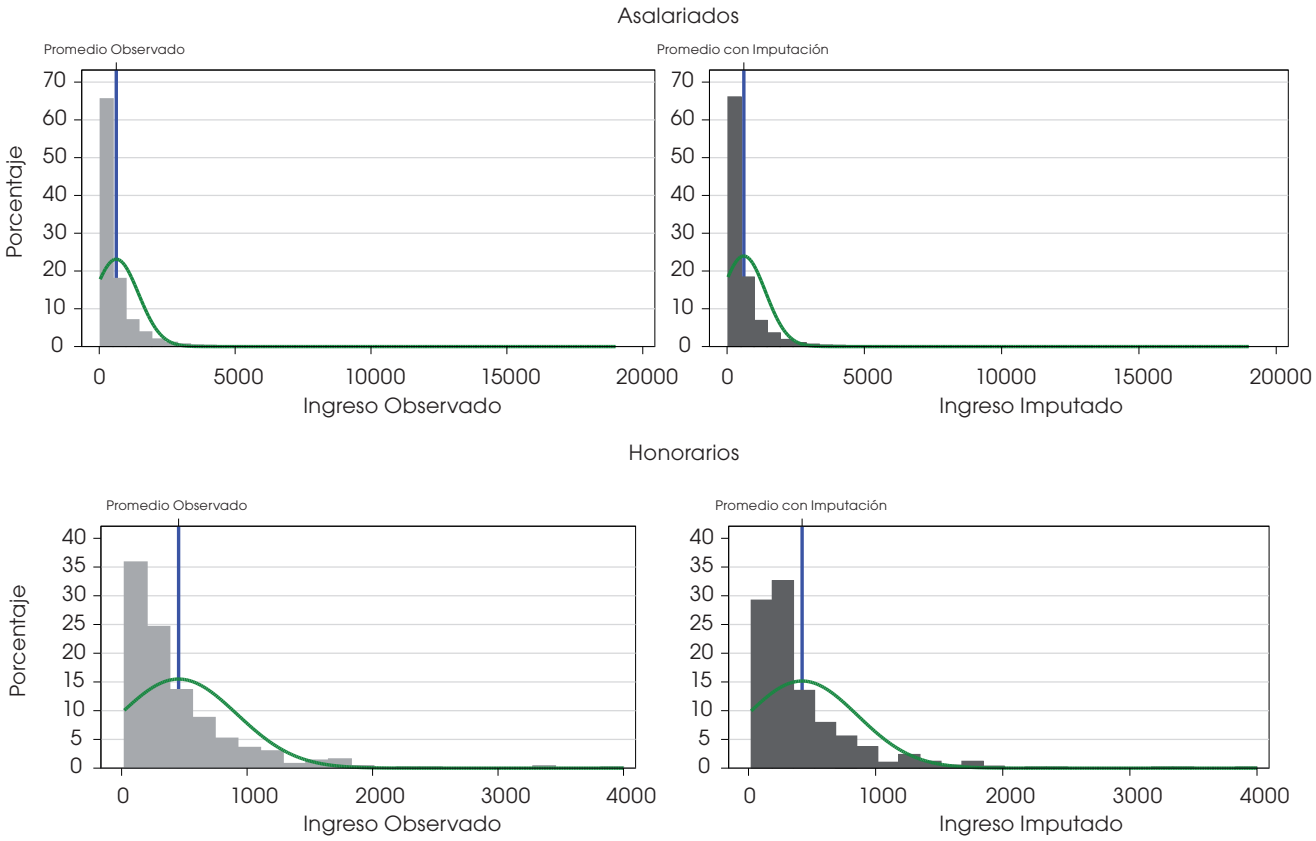


Profesionales Independientes

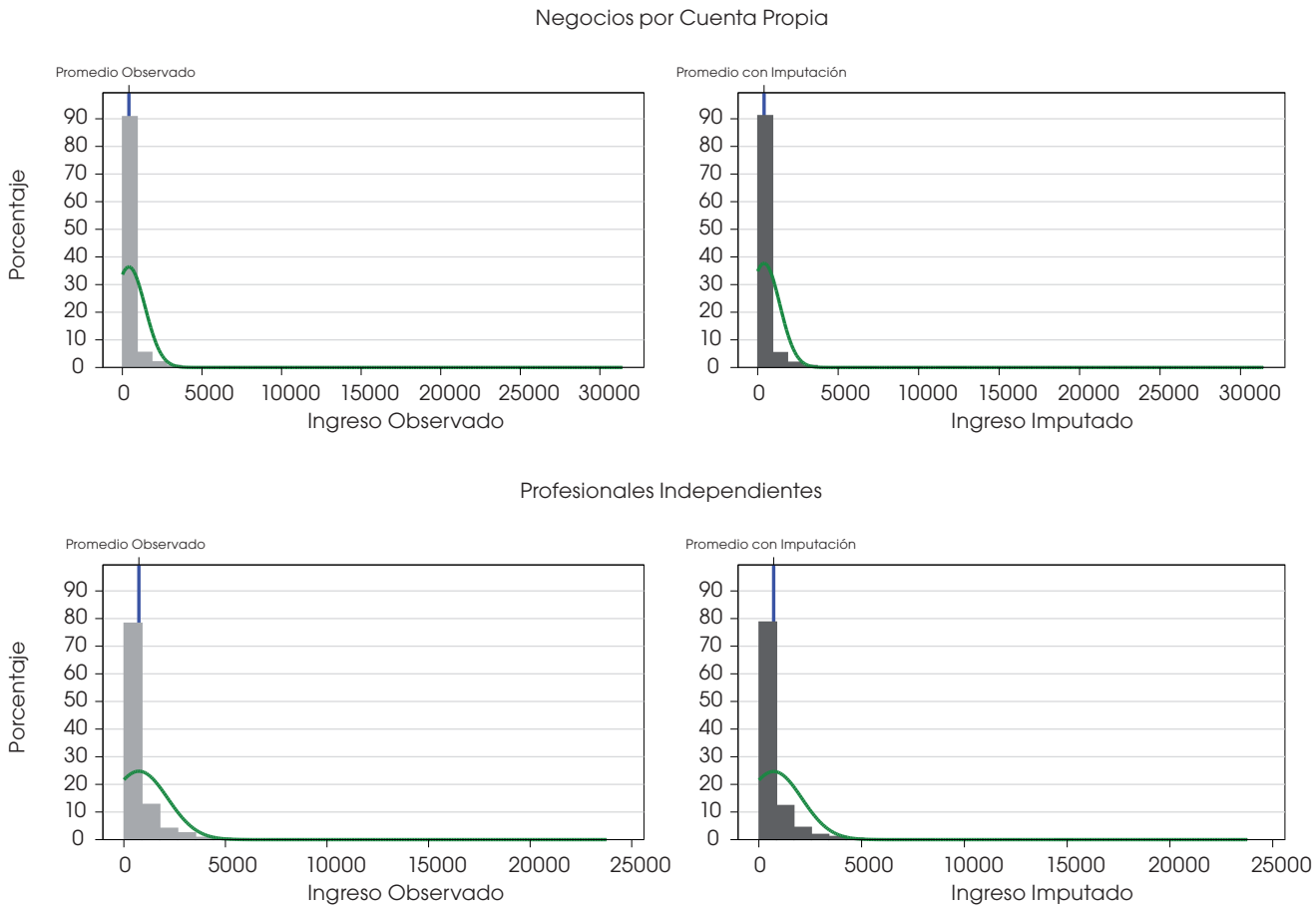


Fuente: VII EPF-INE.

INGRESOS IMPUTADOS POR EM RESTRINGIDO

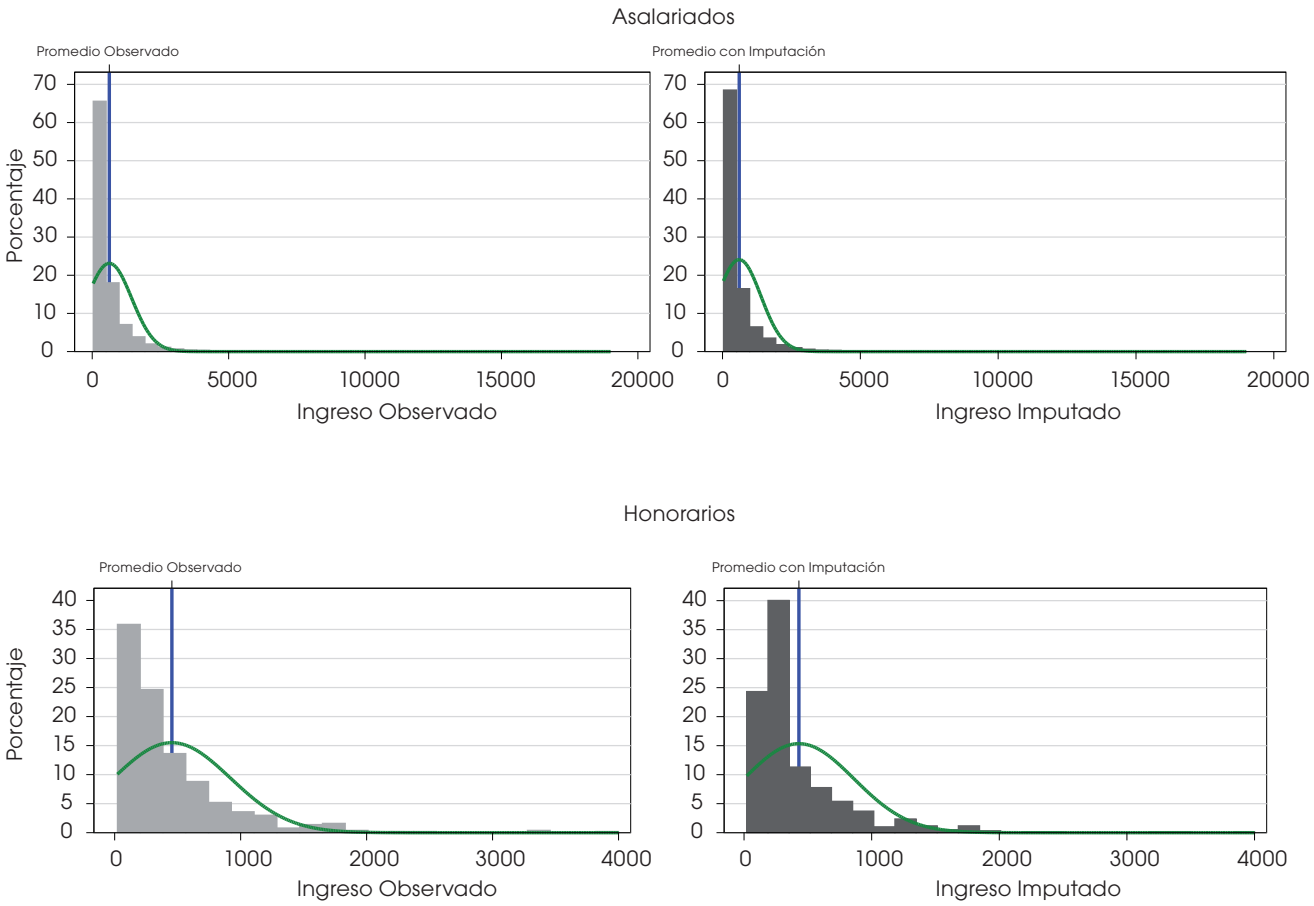


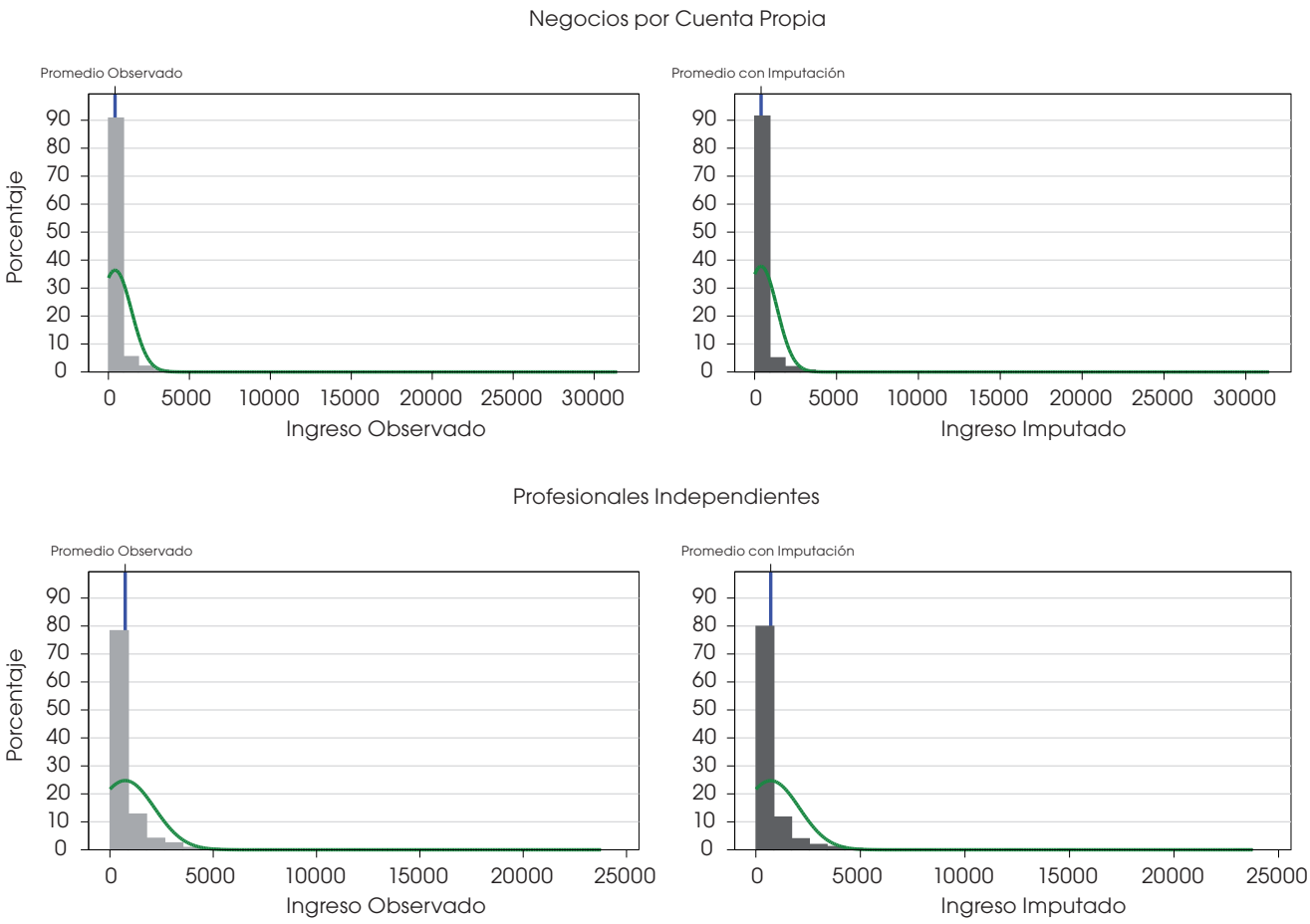
Fuente: VII EPF-INE.



Fuente: VII EPF-INE.

INGRESOS IMPUTADOS POR EM

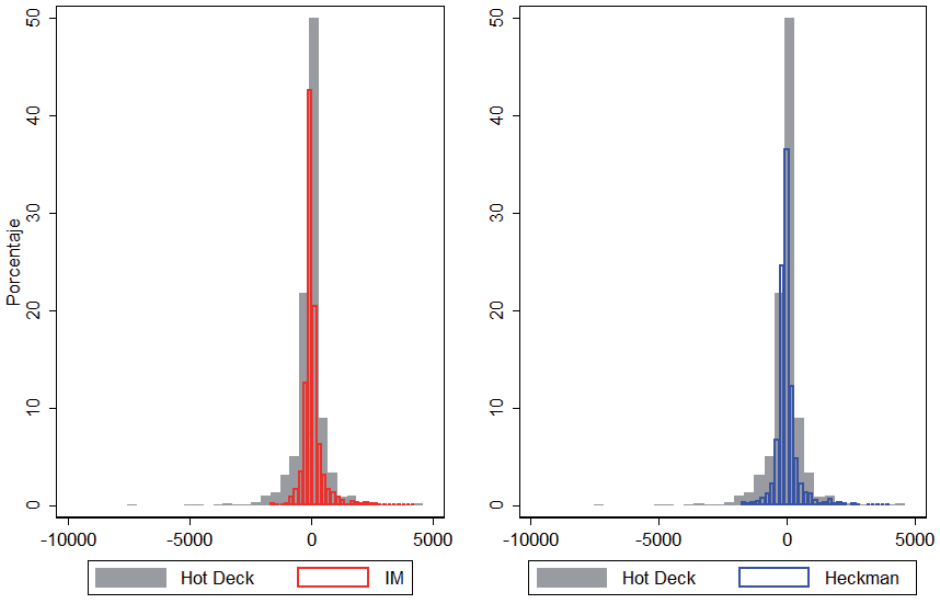




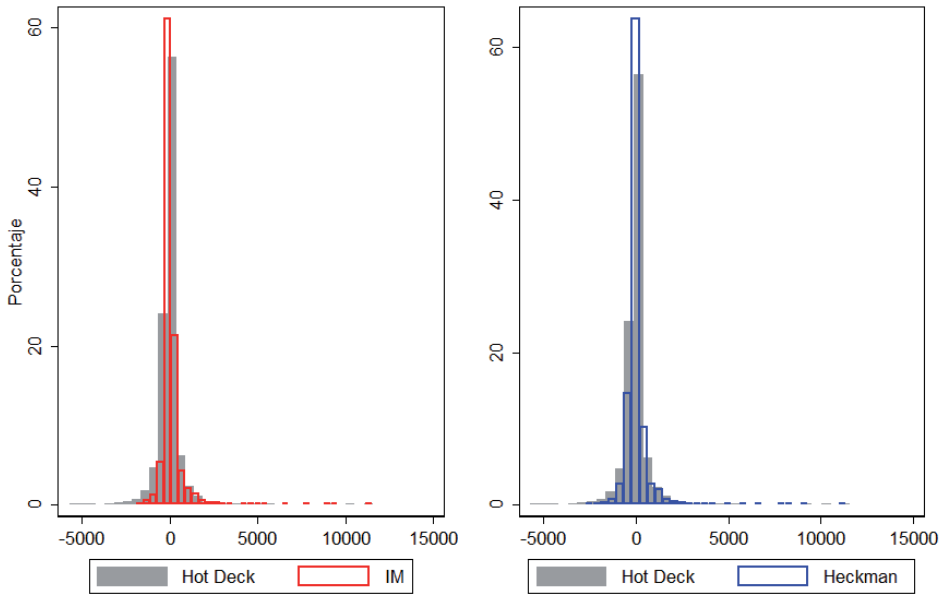
Fuente: VII EPF-INE.

Anexo I. Distribución completa del error por método de imputación

Muestra 10%



Muestra 20%



Anexo J. Definición de variables utilizadas en el documento[56]

Administrador de gastos del hogar ECOMPRAS: el administrador del hogar corresponde a la persona que habitualmente realiza las compras para el hogar. Es posible señalar a más de una persona por hogar.

Años de escolaridad de la persona EDUE: corresponde al número de años de escolaridad formal para cada uno de los miembros del hogar (se excluye del cálculo los cursos de párvulo). Los años de escolaridad varían dependiendo de los niveles educacionales alcanzados, por ejemplo, una persona con estudios de enseñanza media completa tiene un total de 12 años de escolaridad.

Clasificación internacional uniforme de ocupaciones (1 dígito) AECIUO: la Clasificación Internacional de Ocupaciones (CIUO-88) es un sistema clasificador que agrupa las ocupaciones tomando como base la similitud de las competencias requeridas para cumplir las tareas y funciones del empleo. Cada empleo es clasificado a partir del nivel de competencias (en función de la variedad y complejidad de las tareas) y la especialización de las competencias (el tipo de conocimientos aplicados, la naturaleza de los bienes y servicios producidos, etc.). La estructura original propuesta por la OIT se compone de cuatro niveles de desagregación, subdivididos en 10 grandes grupos; 28 subgrupos principales; 116 subgrupos y 390 grupos primarios. En la encuesta la variable se encuentra agregada en 9 grandes grupos y un grupo residual.

EDAD: señala la edad de cada uno de los miembros del hogar en años efectivamente cumplidos al momento de aplicar la encuesta.

Espacio geográfico, Manzana, Comuna, Región y Macrozona: corresponde al código numérico asignado a cada uno de los espacios geográficos y sus diferentes niveles. La macrozona es una categorización de las regiones. Norte compuesta por las regiones de: Arica y Parinacota, Tarapacá, Antofagasta, Atacama y Coquimbo. Centro, compuesta por: Valparaíso, O´Higgins, Maule y Biobío. Sur, conformada por: La Araucanía, Los Ríos, Los Lagos, Aysén, Magallanes; y la cuarta categoría, Región Metropolitana. Por temas de multicolinealidad en las filas de la matriz, se debe excluir a una categoría dejándola como base, en este caso será la Región Metropolitana.

ESTRATO SOCIOECONÓMICO: proviene del diseño muestral y su construcción proviene de una clasificación de los deciles. El primer estrato corresponde a los 3 primeros deciles (más pobres), el segundo estrato se compone de los deciles 4 a 9, mientras que el tercer estrato está formado solamente del último decil.

Identificador del hogar FOLIO: número identificador del hogar encuestado. Se encuentra compuesto de cinco dígitos que generan un valor único para la vivienda y un dígito posterior al guión, el cual indica el número del hogar dentro de la misma.

Identificador de la persona PERSONA: es el número correlativo que identifica a los miembros del hogar, definiéndose éste como quién reside habitualmente en el hogar y comparte un presupuesto común para alimentación y servicios básicos.

Ingreso disponible total por hogar (sin arriendo imputado) INGDHOG_HD: corresponde a la sumatoria de todas las fuentes de ingreso del hogar que puede destinarse íntegramente al consumo o bien al ahorro, devengados al mes anterior al levantamiento de la encuesta.

Ingreso disponible autónomo por hogar: son todos los ingresos disponibles del hogar recibidos por conceptos de trabajo dependiente, trabajo por cuenta propia (ya sea negocios por cuenta propia o profesionales independientes), ingresos por jubilaciones y/o pensiones de vejez pagados por el Instituto de Previsión Social (IPS), Administradora de Fondos de Pensiones (AFP), así como de los sistemas de las FFAA y pensiones de vejez.

Ingreso disponible autónomo por persona: son todos los ingresos disponibles de la persona recibidos por conceptos de trabajo dependiente, trabajo por cuenta propia (ya sea negocios por cuenta propia o profesionales independientes), ingresos por jubila-

56 En este anexo se presentan las definiciones utilizadas en el presente documento, para mayor información sobre todas las variables de la encuesta consultar la metadata disponible en www.ine.cl/epf

ciones y/o pensiones de vejez pagados por el Instituto de Previsión Social (IPS), Administradora de Fondos de Pensiones (AFP), así como de los sistemas de las FFAA y pensiones de vejez.

Nivel de análisis. Nivel educativo: se construye la variable NIVEL_ANALISIS la que toma los siguientes valores: 1 (Nivel Desconocido) 2 (Niveles Básicos o Códigos Especiales) 3 (Educación Media Incompleta) 4 (Educación Media Completa) 5 (Educación Superior Incompleta) 6 (Educación Centro de Formación Técnica Completa) 7 (Educación Instituto Profesional Completa) 8 (Educación Universitaria Completa) 9 (Postítulos) 10 (Nivel de Magíster o Doctorado).

SEXO: señala el sexo de cada uno de los miembros del hogar, hombre o mujer. En este caso corresponde a una variable dicotómica que toma el valor 0 para hombre y 1 para mujer.

ZONA: identificador de la zona a la cual pertenece el hogar, Gran Santiago y Resto de Capitales Regionales.

DIRECCIONES REGIONALES Y PROVINCIALES INE

DIRECCIÓN	TELÉFONO	FAX	CASILLA	CORREO ELECTRÓNICO
REGIÓN DE ARICA Y PARINACOTA				
Dirección Regional ARICA Sotomayor N° 216, Piso 5° Edificio Sacor ARICA	58-2232 471 58-2233 403 58-2250 435 58-2250 074	58-2232 471	-	ine.arica@ine.cl
REGIÓN DE TARAPACÁ				
Dirección Regional IQUIQUE Tomás Bonilla N° 1037 IQUIQUE	57-415 683 57-423 119	57-423 119	-	ine.iquique@ine.cl
REGIÓN DE ANTOFAGASTA				
Dirección Regional ANTOFAGASTA Av. José Miguel Carrera 1701, Piso 5° Edificio de Fomento Productivo - Corfo ANTOFAGASTA	55-269 112 55-283 459 55-497 405	55-222 743	1143	ine.antofagasta@ine.cl
REGIÓN DE ATACAMA				
Dirección Regional COPIAPÓ Chacabuco N° 546, Of. 14, Piso 1° Edificio Copayapu COPIAPÓ	52-230 856 52-212 565 52-218 912 52-239 549	52-230 856 52-212 565 52-218 912 52-239 549	405	region.atacama@ine.cl
Oficina Provincial HUASCO Arturo Prat N° 535, Of. 41, Piso 4° Edificio Domeyko VALLENAR	51-614 396	51-614 396	-	provincia.huasco@ine.cl
REGIÓN DE COQUIMBO				
Dirección Regional LA SERENA Matta N° 461, Of. 104 Edificio Servicios Públicos LA SERENA	51-2215 841 51-2224 506	51-2224 506 51-2215 841	23	ine.coquimbo@ine.cl
LIMARÍ Oficina Provincial Avda. Aristía Oriente N°354, Edificio Alameda, Oficina 309, Piso 3°				
CHOAPA Oficina Provincial ILLAPEL Avda. Ignacio Silva N° 98, esquina Buin, Edificio Maray, Of. N° 206, Piso 2°				
REGIÓN DE VALPARAÍSO				
Dirección Regional VALPARAÍSO 7 Norte N° 519 esquina 2 poniente VIÑA DEL MAR	32-2385800 32-2385803	32-2385801 32-2385868	-	ine.valparaiso@ine.cl
Oficina Provincial LOS ANDES Avenida Chacabuco 122-124 Edificio Gobernación Provincial LOS ANDES	34-405 060	34-405 060	-	ine.losandes@ine.cl
Oficina Provincial QUILLOTA Prat N° 20 Piso 3° QUILLOTA	33-317 657	33-317 657	-	ine.quillota@ine.cl
Oficina Provincial SAN ANTONIO Av. Providencia N° 102, oficina 6A, Piso 3° Edificio Gobernación Provincial SAN ANTONIO	35-288422	35-288422	-	ine.sanantonio@ine.cl
REGIÓN DE O'HIGGINS				
Dirección Regional RANCAGUA Ibieta N° 090 RANCAGUA	72-959 594 72-959 595	72-959 596	-	ine.rancagua@ine.cl
Oficina Provincial SAN FERNANDO Carampangue 684, Letra "B" SAN FERNANDO	72-959 619 72-959 620 72-959 621	72-959 596	-	

DIRECCIONES REGIONALES Y PROVINCIALES INE

DIRECCIÓN	TELÉFONO	FAX	CASILLA	CORREO ELECTRÓNICO
REGIÓN DEL MAULE				
Dirección Regional TALCA 3 Norte N° 1139 TALCA	71-231 013 71-238 227 71-224 131 71-215 595	71-231 013	294	ine.talca@ine.cl
Oficina Provincial CURICÓ San Martín Nº 477 Piso 1° CURICÓ	75-327531	75-327531	-	ine.curico@ine.cl
Oficina Provincial LINARES Manuel Rodríguez Nº 580, Piso 3° LINARES	73-2220 004	73-2220 004	433	
REGIÓN DEL BIOBÍO				
Dirección Regional CONCEPCIÓN Caupolicán N° 567, Piso 5° Edificio La Hechicera CONCEPCIÓN	41-2469300	41-3165732	-	ine.concepción@ine.cl
Oficina Provincial ÑUBLE Edificio Gobernación, Piso 3° CHILLÁN	42-2251201	42-2251201	-	mirta.rodriguez@ine.cl
Oficina Provincial BIOBÍO Edificio Gobernación, Piso 3° LOS ÁNGELES	43-2114401	43-2211404	-	ine.losangeles@ine.cl
REGIÓN DE LA ARAUCANÍA				
Dirección Regional TEMUCO Aldunate N° 620, Of. 704, Piso 7° Edificio Inversur TEMUCO	45-591200	45-591201	849	ine.temuco@ine.cl
REGIÓN DE LOS RÍOS				
Dirección Regional de LOS RÍOS Av. Maipú N° 130, Of. 201, Piso 2° Edificio Consorcio VALDIVIA	63-213 457		-	ine.valdivia@ine.cl
REGIÓN DE LOS LAGOS				
Dirección Regional PUERTO MONTT Edificio Torres Plaza Juan Soler Manfredini N° 11, Piso 11. Of. 1102 PUERTO MONTT	65-253 063 65-259 886	65-259 886 65-253 063	493	ine.puertomontt@ine.cl
Oficina Provincial OSORNO O'Higgins N° 645, Piso 3° OSORNO	64-242 850	64-242 850	144	ine.osorno@ine.cl
Oficina Provincial CHILOÉ O'Higgins N° 480, Piso 3° CASTRO	65-635 774	65-635 774	47	ine.castro@ine.cl
REGIÓN DE AYSÉN				
Dirección Regional COYHAIQUE Avenida Baquedano Nº 496 interior COYHAIQUE	67-211 144 67-214 578	67-231 914	-	ine.coyhaique@ine.cl
REGIÓN DE MAGALLANES Y LA ANTÁRTICA				
Dirección Regional PUNTA ARENAS Croacia Nº 722, Piso 9° Edificio Servicios Públicos PUNTA ARENAS	61-714 550 61-714 567	61-714 558	86	ine.puntaarenas@ine.cl