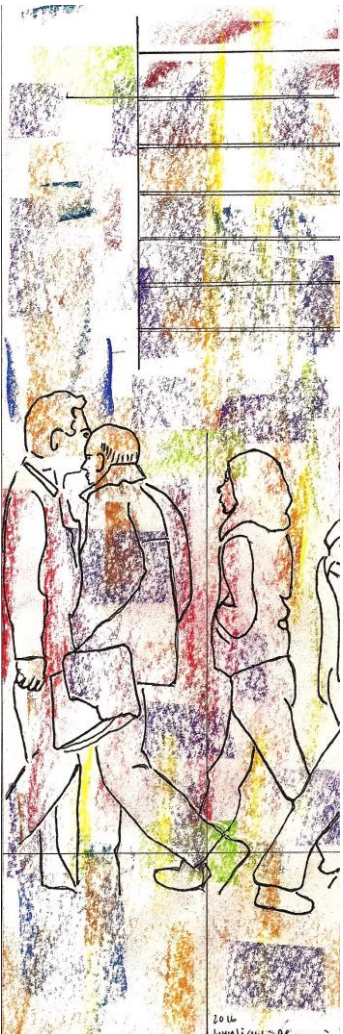
INCASI
Working Paper Series

2019, No. 6

INCASI *International Network for Comparative Analysis of Social Inequalities*



Imputaciones de la no respuesta en las variables de ingreso. Encuesta Permanente de Hogares del Gran Buenos Aires, 1990-2010

Eduardo Donza

European Commission | Horizon 2020 European Union funding for Research & Innovation

Marie Skłodowska-Curie Actions (MSCA)
Research and Innovation Staff Exchange (RISE)
H2020-MSCA-RISE-2015
GA-691004

Imputaciones de la no respuesta en las variables de ingreso. Encuesta Permanente de Hogares del Gran Buenos Aires, 1990-2010

Eduardo Donza[1]

[1] Universidad Católica Argentina
Universidad de Buenos Aires, Argentina
email@email.com

Marie Skłodowska-Curie Actions (MSCA)
Research and Innovation Staff Exchange (RISE)
H2020-MSCA-RISE-2015
GA-691004

This working paper was elaborated in the context of INCASI Network, a European project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie GA, No. 691004, and coordinated by Dr. Pedro López-Roldán. This article reflects only the author's view and the Agency is not responsible for any use that may be made of the information it contains.

Digital edition: https://ddd.uab.cat/record/204786

Dipòsit Digital de Documents
Bellaterra, Cerdantola del Vallès (Barcelona)
Universitat Autònoma de Barcelona

INCASI Working Paper Series
2019, No. 6
Published online: https://ddd.uab.cat/record/204786

INCASI *International Network for Comparative Analysis of Social Inequalities*

Imputaciones de la no respuesta en las variables de ingreso. Encuesta Permanente de Hogares del Gran Buenos Aires, 1990-2010[1]

Eduardo Donza

Resumen

La no respuesta a las preguntas de ingreso constituye uno de los principales problemas de los estudios referidos a la temática y de los que utilizan sus datos para estratificar o clasificar a la población. La ausencia de respuestas depende de factores que exceden la instancia de relevamiento, tales como procedimientos y estrategias en el proceso de investigación, cambios de cuestionarios y factores socioeconómicos. Utilizando los datos relevados por la Encuesta Permanente de Hogares del Instituto Nacional de Estadísticas y Censo para el aglomerado Gran Buenos Aires de la Argentina, retomando y ampliando estudios anteriores, en esta ponencia se determina la evolución y el impacto de la no respuesta a las preguntas de ingresos entre 1990 y 2010. Se demuestra que la no respuesta es elevada en parte del período, fluctuó por cuestiones técnicas y efectos procedentes de la población, y su distribución no es completamente aleatoria. Debido a esto se plantea la necesidad de realizar una imputación de los ingresos no declarados y se aplica el procedimiento de máxima verosimilitud para enmendar la no respuesta. Se realiza la estimación de los ingresos no declarados, a partir de los cuales se recalculan los principales estadísticos sociales. Se verifican importantes diferencias entre los estadísticos calculados al considerar solo los valores de ingresos declarados y los obtenidos al considerar también los ingresos imputados. Se comparan las imputaciones con las realizadas por el organismo oficial de estadísticas en parte del período. Se concluye la necesidad de que los organismos productores de información refuercen sus actividades para generar información de mayor calidad y se recomienda que para el uso de la información sobre ingresos se imputen los valores a las preguntas no respondidas.

Palabras clave

Ingresos, no respuesta, imputación, aglomerado Gran Buenos Aires

Índice

[1] El autor agradece los comentarios recibidos a este documento presentado en la sesión del RC55 "*Imputation and Social Indicators: The Use of Factor Analysis for Imputing Missing Data*" del III ISA Forum, realizado en Viena, Austria el 14 de julio de 2016

1. Planteo del problema

Uno de los problemas que enfrentan los estudios empíricos de ingresos es la no respuesta ante la requisitoria de datos. Este hecho, que es independiente de la fuente de datos, no depende sólo de la instancia del relevamiento sino, también y mayoritariamente de otras instancias del proceso de investigación y de factores contextuales.

Específicamente, los estudios, o puntualmente, las preguntas que relevan los ingresos de las personas suelen presentar una gran incidencia de no respuestas o respuestas parciales. Ante este hecho los diversos tipos de estudios que se basan en estos datos (distribución del ingreso, pobreza e indigencia, evolución de ingresos de la población, evolución de ingresos sectoriales, estrategias familiares, etc.) se ven complejizados, limitados, sesgados y/o expresan solamente en forma parcial los eventos a los que pretenden referirse.

Basado en estos antecedentes, retomando y ampliando trabajos anteriores[2], en este informe se determina la incidencia de la no respuesta a las preguntas referidas a los ingresos, el efecto que esto genera, el carácter o la gravedad del hecho y se recomiendan estrategias supletorias de estos inconvenientes. Esto se realiza, a partir de datos de la Encuesta Permanente de Hogares (EPH) del Instituto Nacional de Estadísticas y Censos (INDEC) para el aglomerado Gran Buenos Aires (GBA) entre los años 1990 y 2010.[3]

2. Consideraciones teóricas sobre la no respuesta

En lo que respecta a las respuestas sobre ingresos, se consideran, en general, dos posibles orígenes de error de medición: subdeclaración y la no declaración.

La primera, se basa en la sospecha de la existencia de una conducta sistemática de subdeclaración de ingresos monetarios por parte de los perceptores, especialmente en referencia a los ingresos provenientes de ganancias, utilidades de capital y transferencias. Estudios como el de Camelo (1998), Llach y Montoya (1999) y Roca y Pena (2001), entre otros, basados en la EPH y otras fuentes (Sistema Integrado de Jubilaciones y Pensiones, Cuentas Nacionales, etc.), ponderan el posible sesgo generado por la subdeclaración y la especifican. Independientemente de la disidencia que plantean Roca y Pena, sobre los análisis de Camelo y de Llach y Motoya, podemos considerar que: "En conclusión, de una primera lectura de los datos de las fuentes analizadas no aparecen indicios claros sobre subdeclaración de ingresos por parte de la EPH en cuanto a los perceptores de ingresos fijos como los jubilados y los asalariados. Las diferencias mayores en los niveles de subdeclaración se concentrarían en los ingresos de perceptores de rentas, ganancias empresariales, e incluso de los trabajadores por cuenta propia que seguramente no declaran correctamente sus ingresos, a veces por la propia dificultad en diferenciar claramente ingresos netos de actividad." (Roca y Pena, 2001).

Teniendo en cuenta estas afirmaciones (coincidentes con apreciaciones de Lindenboim, Kennedy y Graña (2006)), es decir, la no certeza de subdeclaración en perceptores de ingresos fijos y la limitada incidencia del volumen[4] de

[2] La temática es tratada por Camelo (1998), Becaria (1998), Berumen y Muñoz (1996), Feres (1998), Keifman, S. y otros (1998), Muñoz (1996), Medina y Galván (2007), Salvia y Donza (1999), entre otros.

[3] Se parte del año 1990 para considerar el efecto de los cambios económicos generados por el del Plan de Convertibilidad (1991-2001) y los cambios metodológicos aplicados a la EPH (primera parte de la década de 1990 y 2003). El estudio contempla el efecto del cambio de la modalidad de relevamiento puntual a continuo (siendo el último relevamiento contínuo utilizado en este trabajo el de octubre de 2002). Dentro de estos cambios, además del tipo de relevamiento, se realizó una modificación del cuestionario con importantes cambios en las preguntas referidas a ingresos. Por otra parte, para una mejor comparabilidad se utilizaron las versiones de las bases de EPH de modalidad continua que posee el período 2003-2010 completo (esto se debe a que se dispone de otras versiones de las bases EPH continuas para los años 2003-2006).

[4] Se debe tener en cuenta que la distribución según fuente del total de ingresos relevados por la EPH – GBA, para el año 1992, es: asalariada 53,6%, cuenta propia 21,6%, jubilaciones y pensiones 10,7%, utilidades y rentas 1,1% y

ingresos de los posibles subdeclarantes (rentas, ganancias empresariales y trabajadores por su cuenta), se puede considerar como el más importante al segundo posible error de medición: la no declaración de ingresos.

En este aspecto, la no respuesta o respuesta parcial, puede generar serios impedimentos al realizar los análisis: "Debido a estos "casos perdidos" los estudios sobre remuneraciones o ingresos familiares están impedidos de hacer inferencias al total de la población por el recorte que sufre la muestra. Asimismo, los análisis de asociación también se ven afectados, a no ser que se asuma a ciegas el supuesto –por demás riesgoso- que los casos perdidos presenten distribuciones multidimensionales semejantes a los registros con ingresos informados." (Salvia y Donza, 1999).

Esta afirmación se refuerza por el hecho que en los estudios de distribución del ingreso basados en hogares y los estudios de pobreza, generalmente, la no declaración o declaración parcial de ingresos de un perceptor del hogar impide la consideración de la totalidad de los componentes del hogar (es decir, del hogar) en el estudio. Induciendo además, en segunda instancia, posibles alteraciones en la aplicación de series temporales en las cuales no se podrán diferenciar el efecto generado por el cambio del perfil de perceptores, factores contextuales o cambios metodológicos en el proceso de medición.

3. Incidencia de la no respuesta de ingresos entre 1990 y 2010 en GBA

La tabla 1 y la figura 1 presentan el porcentaje de no respuestas y su efecto en el porcentaje de perceptores, de hogares y de la población total. Como puede observarse, en el período 1991-2002 (último dato disponible por limitaciones en la forma de presentación de los microdatos originales), el porcentaje de preguntas de ingresos no respondidas se redujo de 18,1 % a 11,9 %, presentándose el menor de los valores en 1998 (6,6 %).

Tabla 1. No respuestas de ingresos monetarios, perceptores, hogares y población afectada. Gran Buenos Aires: 1990-2010. Octubre 1990-2002 y 2.º semestre 2003-2010

Porcentaje con respecto al total de referencia

Año	No respuestas de ingresos	Perceptores con ingresos no declarados	Hogares con ingresos no declarados	Población de hogares con ingresos no declarados
Modalidad puntual – 1.er cuestionario				
1990	(1)	**20,0 %**	**26,2 %**	**29,3 %**
1991	**18,1 %**	18,6 %	25,5 %	29,1 %
1992	13,7 %	14,4 %	19,2 %	21,7 %
1993	10,3 %	9,8 %	12,9 %	14,2 %
1994	7,4 %	8,1 %	10,6 %	11,2 %
Modalidad puntual – 2.º cuestionario				
1995	10,5 %	7,9 %	10,7 %	11,1 %
1996	8,6 %	9,2 %	12,3 %	12,8 %
1997	6,9 %	8,3 %	9,7 %	9,8 %
1998	**6,6 %**	**7,3 %**	**9,1 %**	9,6 %
1999	8,6 %	9,2 %	11,7 %	12,2 %
2000	8,4 %	8,9 %	11,3 %	11,6 %
2001	12,3 %	10,6 %	12,8 %	13,8 %
2002	**11,9 %**	11,7 %	14,9 %	14,5 %
Modalidad continua				
2003	(2)	19,4 %	24,8 %	26,1 %
2004	(2)	20,7 %	26,9 %	28,1 %
2005	(2)	18,0 %	23,7 %	24,3 %
2006	(2)	15,7 %	21,0 %	21,7 %
2007	(2)	22,8 %	29,5 %	31,3 %
2008	(2)	20,7 %	26,9 %	28,7 %
2009	(2)	21,6 %	29,0 %	30,7 %
2010	(2)	**24,3 %**	**32,7 %**	**34,8 %**

Notas: (1) No es posible definir la cantidad de preguntas no respondidas debido a inconsistencias en la base de datos. (2) No es posible definir la cantidad de preguntas no respondidas debido al esquema de codificación de las no respuestas utilizado en las bases de datos originales.
Fuente: elaboración propia con base en datos de la EPH, INDEC.

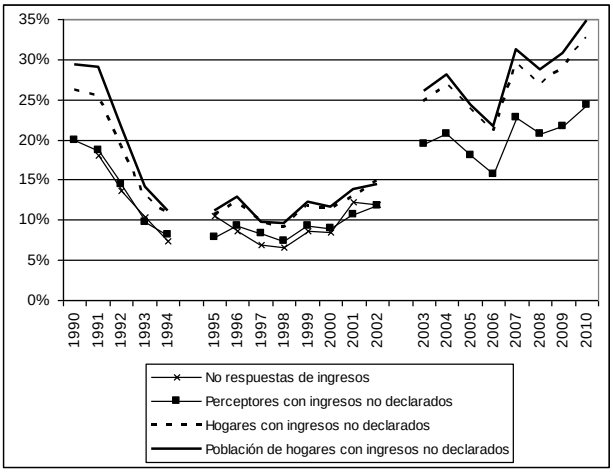
Por su parte, al comienzo de la serie analizada, estas no repuestas involucran a 20,0 % de los perceptores de ingresos, para culminar el período de análisis con 24,3 % de perceptores no declarantes de ingresos o de declaración parcial (que presentan por lo menos una de sus fuentes de ingresos sin declaración). Cabe destacar que la mejor captación se registra en el año 1998, cuando alcanza a 7,3 % de los perceptores. Si para la producción de información y la realización de estudios sobre ingresos se excluyeran estos hogares y su población, tal como se realiza en otras investigaciones, la proporción de hogares

otra fuente de ingresos 3,0%. Lo cual supone, según los autores citados, que un 64% del total de ingresos de la población no posee subdeclaración evidente.

por excluirse variaría entre 26,2 % (1990) y 32,7 % (2010). Asimismo, la población no considerada ascendería a 29,3 % en 1990 y a 34,8 % en 2010. La incidencia de la no respuesta es menor en 1998 cuando no se registran los ingresos completos en apenas 1 de cada 10 hogares.

Figura 1. No respuestas de ingresos monetarios, perceptores, hogares y población afectados por la no declaración Gran Buenos Aires: 1990-2010. Octubre 1990-2002 y 2.º semestre 2003-2010

Porcentaje respecto al total de referencia respectivo



Fuente: elaboración propia con base en datos de la EPH, INDEC.

Del análisis de los datos se deduce una errática evolución en la capacidad de captación de información de ingresos de la EPH. La mejora es muy importante entre los años 1990 y 1994, consigue una relativa estabilización entre 1994 y 2000, y se incrementa levemente hasta 2002. A partir de 2003 se da un importante aumento, luego una leve disminución entre 2004 y 2006, y por último, un marcado incremento entre 2006 y 2010.

Si bien en períodos de inestabilidad monetaria e inflación los perceptores (en especial los de ingresos variables) pueden perder la noción relativa del nivel o de "la cifra" de ingresos que perciben, y/o tener una predisposición negativa a expresarlos, el análisis de los indicadores económicos pone de manifiesto que pasada la hiperinflación de los años 1989-1990, el indicador comúnmente utilizado para expresar el costo de vida, el Índice de Precios al Consumidor (IPC), presenta una marcada disminución y una estabilización con valores muy bajos en toda la etapa de vigencia de la Ley de Convertibilidad (1991-2001); un aumento de casi 40 % a la salida de esta etapa (2002); una relativa estabilización posterior (2003-2006); y finalmente, con controvertidos valores, un incremento anual que oscila entre 13,1 % y 26,9 % en el período 2007-2010.[5]

En este sentido, pese a que la evolución del costo de vida real acompaña en parte la evolución de la no respuesta a preguntas de ingresos, no aparece como un factor de tanta determinación si se consideran los cambios metodológicos aplicados a la EPH y la coyuntura institucional del INDEC. Es posible que la marcada tendencia a la disminución de la no respuesta entre 1991 y 1998 se deba a las actividades de los equipos técnicos de la EPH para mejorar la calidad de la información. La sistematización de los controles de calidad, la intensificación del acompañamiento a los encuestadores, los cambios en las modalidades de "recuperación" de encuestas y la reorganización de la estructura de personal bien pueden ser algunos de los factores que explican la mejora (INDEC, 1998).

Por el contrario, parecería que la implementación de la modalidad continua y la aplicación del nuevo cuestionario trajeron aparejado un incremento acentuado de la no respuesta. Habiendo partido de un piso muy elevado en 2003 (19,4 % de los perceptores que no responden), que se va incrementando hasta 2004, el efecto cuestionario –producto de un cambio en la operacionalización de la condición de actividad económica y del incremento de la cantidad de preguntas referidas a ingresos– se constituye en el principal factor explicativo del aumento de la incidencia.[6] El descenso posterior, entre 2004 y

[5] Para el período 1990-2006 se hace referencia al IPC generado por el INDEC. En el período 2007-2010 se utiliza el IPC-7 Provincias elaborado por el Centro de Estudios para el Desarrollo Argentino (CENDA) a partir de los IPC oficiales correspondientes a los aglomerados de Jujuy, Neuquén, Paraná, Rawson-Trelew, Salta, Santa Rosa y Viedma, generados por las direcciones provinciales de estadística correspondientes. "Se seleccionaron los IPC de estos aglomerados, debido a que los mismos no estaban incluidos en la primera etapa del programa IPC-Nacional y se mantuvieron al margen de las modificaciones del IPC-INDEC" (CENDA, 2011a: 5).

[6] Otro importante cambio introducido a partir de 2003 fue el tipo de relevamiento: pasó de puntual a continuo; pero el efecto aislado de esta modificación en el nivel de no

2006, puede atribuirse a los esfuerzos introducidos por los equipos de la EPH (cuestiones operativas, capacitación, supervisión, etc.). Por último, es probable que el incremento correspondiente a los cuatro años finales de la serie resulte de la combinación de un factor económico (aumento del costo de precios), los consabidos problemas institucionales que presenta el INDEC desde 2007 (desplazamiento de los equipos técnicos con experiencia en los procesos de relevamiento y análisis de calidad de los datos) y un presumible descontento generalizado entre la población (tanto por la falta de credibilidad de la institución como por problemas coyunturales), que aumentó la propensión a no colaborar en las encuestas.

Conforme a estas afirmaciones y a las evidencias empíricas presentadas, se concluye que para el desarrollo de estudios referidos a remuneraciones, pobreza, distribución del ingreso y estratificación social basada en los ingresos resulta imprescindible aplicar algún tipo de estimación de ingresos no declarados para considerar en el análisis la mayor cantidad posible de casos, evitando así la exclusión de registros que sesguen los resultados.

4. Alternativas ante la no respuesta

Conforme lo expresado por varios autores[7] una de las cuestiones a tener en cuenta para la elección de un procedimiento adecuado de tratamiento de la no respuesta es la distribución de los datos faltantes. Esto se debe a que algunos métodos poseen como supuesto que la distribución de datos no conocidos posee un patrón determinado.

Una de las posibilidades es que estos datos presenten un patrón completamente aleatorio (*Missing Completely At Random*, MCAR). La probabilidad de que una respuesta a una variable sea dato faltante es independiente tanto del valor de esta variable como del valor de otras variables del conjunto de datos. Es decir, la ausencia de la información no está asociada con ninguna variable presente en la matriz de datos. Es frecuente que en la práctica esto no ocurra, ya que la falta de respuesta suele estar asociada a características definidas.

Otra opción es que la respuesta sea dato faltante independientemente de los valores de la misma variable, pero que dependa de los valores de otras variables del conjunto de datos. Es decir, en este caso la ausencia de datos está asociada a variables presentes en la matriz de datos. Se identifica como un proceso de datos omitidos que se genera en forma aleatoria (*Missing At Random*, MAR).[8]

Estas dos posibilidades de datos faltantes suelen denominarse "ignorables" por producir efectos que se pueden ignorar si se controlan adecuadamente las variables que determinan la no respuesta. Por fin, en la mayoría de los patrones observados, la probabilidad de la ausencia de datos depende tanto de los valores que asume la variable analizada como de los valores de otras variables y no sigue un proceso aleatorio (*Not Missing At Random*, NMAR). En este caso los datos faltantes se denominan "no ignorables".

4.1. Tratamientos de la no respuesta que no constituyen imputación

Una de las posibles decisiones ante la no respuesta a las preguntas de ingresos es limitar el análisis a los registros que poseen datos declarados. Tal como lo presenta Otero García (2011), son varios los procedimientos que solo se limitan a la utilización de datos conocidos y no avanzan sobre la constitución de imputaciones, entre ellos los siguientes:

4.1.a Análisis con datos completos (*listwise o case deletion LD*)

Procede a eliminar los registros que no poseen datos y el análisis se limita a las unidades que poseen información completa en todas las

respuesta a las preguntas de ingresos es de difícil determinación.

[7] Puede contarse entre ellos a BID (2004), Feres (2004), Medina (2004), Medina y Galvan (2007), Otero García (2011), etc.

[8] Según Medina y Galván (2007) en las distribuciones de ingresos la omisión presenta un patrón MCAR, si se cumple que, en promedio, los respondentes tienen ingresos similares a los no respondentes. Por otra parte, siguen un patrón MAR si la no respuesta depende, por ejemplo, del nivel educativo pero en cada uno de los niveles educativos la no respuesta no está relacionada con el ingreso.

variables. La ventaja es la simplicidad en la aplicación, pero con ella se excluye del análisis buena parte de los registros originales y, en consecuencia, puede ocasionar sesgos en las estimaciones de los parámetros. Únicamente se podría utilizar bajo la comprobación de que el patrón de no respuesta presenta un comportamiento aleatorio (MCAR o MAR). En caso de que los datos provengan de una muestra probabilística, se debe tener un particular cuidado, ya que las unidades fueron elegidas por procedimientos aleatorios y con probabilidades de selección, y su no consideración puede generar sesgos en las estimaciones de los parámetros. Una forma de evitar este inconveniente es generar nuevos factores de expansión y/o corrección.

4.1.b Análisis con los datos disponibles (*pairwise deletion*)

Como alternativa al análisis de datos completos, este procedimiento consiste en considerar en el análisis todos los datos de que se dispone para cada variable analizada. Al igual que el método anterior, tiene como ventaja la simplicidad. El inconveniente de trabajar con la muestra truncada por este método se agrava al truncarse de forma diferente según las variables utilizadas. Su uso debería limitarse solamente a un proceso de no respuesta de tipo MCAR y, de utilizarse datos provenientes de muestras, se deberían generar nuevos factores de expansión y/o corrección.

4.1.c Ajuste de ponderadores

Consiste en generar nuevos factores de ponderación que suplan la no respuesta. Su mayor utilización se da cuando se tiene falta de información para todos los datos de un caso (ejemplo típico es la encuesta no realizada). Por medio del conocimiento de información auxiliar, se incrementa el peso de algunos casos que sí respondieron para que representen a los que no lo hicieron.[9]

4.2. Procedimientos de imputación

Los métodos de imputación para generar una base de datos completa pueden ser clasificados en *simples* (se asigna un valor para cada valor faltante en función de la propia variable o de otras) o *múltiples* (para cada valor faltante se asignan varios valores; luego, utilizando estos conjuntos de datos completos, se estiman los parámetros necesarios y posteriormente se combinan los resultados). Otra posible clasificación de los procedimientos de imputación es en *determinísticos* (para cada unidad producen la misma respuesta cuando se reproduce el procedimiento bajo las mismas condiciones) y *estocásticos o aleatorios* (para cada unidad pueden producir resultados diferentes cuando se reproduce el procedimiento bajo las mismas condiciones).

A continuación, resumiendo lo expresado por MECOVI (2004), Medina (2004), Medina y Galván (2007) y Otero García (2011), se presenta una breve descripción de los principales métodos.[10]

4.2.a Métodos de imputación simple

a.1 Imputación por la media

Es uno de los procedimientos más antiguos y sencillos. Los valores faltantes de una variable se sustituyen por la media de las unidades observadas en esa variable. Tiene una versión determinística y una aleatoria (que incluye un residuo aleatorio). Presenta dos variantes:

a.1.1 Imputación por media no condicional: consiste en estimar la media de los valores observados. Se asume que los datos faltantes siguen un patrón MCAR. Preserva el valor medio de la variable, pero los estadísticos que definen la forma de la distribución (varianza, percentiles, sesgo, etc.) pueden verse afectados, y lo mismo puede ocurrir con las relaciones entre las variables.

a.1.2 Imputación por media condicional: imputa medias condicionadas a valores observados. Un

[9] El INDEC utilizó el procedimiento de ajuste de ponderadores ante la no respuesta de ingreso en la EPH entre 2003 y 2006. Con posterioridad, para el mismo período, se aplicó la técnica de *hot deck*.

[10] Determinados autores, con el fin de anular un posible sesgo de selección muestral, combinan alguno de estos métodos de imputación con procedimientos como el de Heckman. Ejercicios de este tipo fueron realizados por Zamudio Carrillo (1995), Crosta (2000), Arrazola y Hevia (2005), Tenjo Galarza, Ribero Medina y Bernat Diaz (2006), González Espiti (s/f), Perlbach de Maradona y Calderón (s/f), entre otros. Esta investigación parte del supuesto de que la incidencia del sesgo de este tipo es despreciable.

método común consiste en agrupar los valores observados y no observados en clases e imputar los valores faltantes por la media de los valores observados en la misma clase.

a.2 Imputación deductiva

Es un método determinístico que se aplica en situaciones donde las respuestas que faltan se pueden deducir del resto de la información proveniente del conjunto de datos, es decir, los valores se asignan mediante relaciones lógicas entre las variables.

a.3 Imputación *cold deck*

Este procedimiento utiliza información externa a la base de datos considerada. Consiste en asignar valor a los datos faltantes en función de información conocida de otras encuestas, datos de registros, datos históricos, etc. Se suele usar en encuestas de tipo panel, en las cuales ya se cuenta con información de algunas de las unidades de registro. La desventaja principal de este método es que la calidad de los resultados dependerá de la calidad de la información externa disponible.

a.4 Imputación *hot deck*

Es un proceso de duplicación de valores conocidos. Consiste en asignar al registro que no posee valor el dato correspondiente a un registro con valor conocido. La técnica utilizada para la elección del registro "donante" de información genera diferentes variantes del método *hot deck*:

a.4.1 Imputación aleatoria *hot deck* (Imputación *hot deck* por muestreo aleatorio simple): se asigna aleatoriamente un valor recogido en la muestra de la variable a imputar. Conserva la distribución de los respondentes pero no considera si es factible la imputación ni la correlación con otras variables. Es un método estocástico.

a.4.2 Imputación aleatoria *hot deck* por grupos: es similar a la técnica anterior, pero se aplica con un procedimiento de clasificación asociado. Todas las unidades de la muestra están clasificadas en grupos lo más homogéneos posible. A cada valor que falte se le asigna un valor del mismo grupo. Así, el supuesto que se está utilizando es que dentro de cada grupo de clasificación la no respuesta sigue la misma distribución que los que responden. Desde ya, las variables de clasificación deben estar relacionadas con los valores que falten y con los valores de los que contestan. Si esto no se mantiene, el procedimiento *hot deck* puede llevar a resultados erróneos. Es un método estocástico.

a.4.3 Imputación *hot deck* secuencial: se usa cuando la muestra tiene algún tipo de orden dentro de cada grupo de clasificación. Cada valor faltante se reemplaza por el del registro perteneciente al mismo grupo e inmediatamente anterior a él; si el primer registro tiene un dato faltante, se lo reemplaza por un valor inicial que puede obtenerse de información externa. En caso de que se impute un gran número de registros en el mismo grupo, se presenta la dificultad de asignarle el mismo valor a todos los casos no conocidos, lo que lleva a una pérdida de precisión de las estimaciones.

a.4.4 Imputación *hot deck* - vecino más cercano: procedimiento no paramétrico basado en la suposición de que los individuos cercanos en un mismo espacio tienen características similares. Es un método de imputación determinístico. Para aplicarlo se requiere definir una medida de distancia.

En líneas generales, un inconveniente que puede generar el método *hot deck* es la duplicación del mismo valor muchas veces. Esto ocurre cuando en los grupos de clasificación hay muchos valores faltantes y pocos valores registrados. Resulta más confiable cuando se trabaja con tamaños de muestra grandes o censos, para así poder seleccionar valores que reemplacen las unidades faltantes.

a.5 Imputación por regresión.

Consiste en aplicar modelos de regresión para imputar información en la variable Y, a partir de variables $(X_1, \ldots, X_k)$ correlacionadas con Y. Este procedimiento consiste en eliminar las observaciones con datos incompletos y ajustar la ecuación de la regresión para predecir los valores faltantes.

Sea n el tamaño muestral, consideremos la variable Y que presenta los primeros r valores observados y los n-r valores faltantes. Supongamos que las k variables $-X = (X_1, \ldots, X_k)-$ no presentan valores perdidos. Si para el caso i tenemos que el valor Yi no se observa, este valor faltante es imputado mediante el modelo de regresión construido.

Según el tipo de variable Y y de la función de distribución se pueden identificar los siguientes modelos: Y es una variable continua (ya sea determinística o estocástica); Y es una variable binaria; Y es una variable de tipo cómputo; Y es una variable categórica (con más de dos categorías); e Y es una variable mixta.

a.6 Imputación mediante el método de regresión secuencial multivariante (*sequential regression multiple imputation*)

Procedimiento estocástico que considera elementos aleatorios. La estrategia básica se basa en crear imputaciones por medio de una secuencia de regresiones. El tipo de regresión depende de la variable que será imputada y se pretende recoger la correlación de todas las variables.

Se parte de una iteración inicial en la cual se imputa, mediante un modelo de regresión, el valor de la variable con menos falta de respuesta, Y_1, sobre las variables explicativas X. Una vez obtenida una predicción de Y_1 se incorpora esta variable a la matriz X de las variables completas, se obtiene una nueva matriz y se realiza la regresión de Y_2 sobre esta última matriz, y así sucesivamente.

Una vez realizada esta iteración de regresiones según el modelo correspondiente en función del tipo de variable, se tiene una primera imputación de todos los valores faltantes. En las iteraciones siguientes lo que se hace es repetir esta iteración inicial pero incluyendo como variables explicativas todas las variables, ya que ahora no hay valores faltantes en ninguna de ellas. Este paso da lugar a actualizaciones de las imputaciones hechas en el paso inicial, que incorporan la información de las variables que se imputan después. El proceso se detiene cuando se alcanza el número de iteraciones predeterminado por el usuario.

a.7 Estimación por máxima verosimilitud (MV)

Su objetivo es realizar estimaciones verosímiles de los parámetros de una distribución cuando existen datos faltantes, suponiendo que los datos completos siguen un determinado modelo multivariante. Es importante elegir un modelo lo suficientemente flexible para reflejar las características de los datos estudiados.

Parte de considerar los valores observados (*Yobs*), los valores faltantes (*Ymis*) y un parámetro o parámetros (θ) que definen la distribución poblacional por medio de una función de densidad que incluye conjuntamente a *Yobs* e *Ymis*. La función de densidad marginal de *Yobs* es obtenida integrando sobre los valores faltantes *Ymis*. La función de verosimilitud $l(\theta|Yobs)$ es proporcional a $f(Yobs|\theta)$ que determina la verosimilitud de los posibles valores de θ.

Los estimadores máximos verosímiles se suelen obtener maximizando la función de verosimilitud respecto de θ. Para simplificar, los cálculos se realizan maximizando el logaritmo de dicha función.

Un procedimiento eficaz para maximizar la verosimilitud cuando existen datos faltantes es el algoritmo EM (*Expectation-Maximization* / Maximización Esperada): un algoritmo iterativo general basado en factorizar la función de verosimilitud que permite obtener estimaciones máximo verosímiles cuando hay datos no completos con unas estructuras determinadas. Puesto que este algoritmo se basa en la idea de imputar los valores faltantes e iterar, ha sido utilizado a lo largo de los años en diferentes contextos. Cada iteración del algoritmo EM consiste en un paso E (*expectation*) y un paso M (*maximization*). Ambos pasos son conceptualmente sencillos y fáciles de implementar en programas informáticos.

Una ventaja adicional de este algoritmo es que puede converger de forma fiable, en el sentido de que en condiciones generales cada iteración incrementa el logaritmo de la función de verosimilitud. Por el contrario, una desventaja del algoritmo EM es que la convergencia se hace más lenta proporcionalmente a la cantidad de datos faltantes.

En el paso E se calculan los valores esperados en la información ausente a partir de los valores observados y de las estimaciones actuales de θ, para posteriormente reemplazar la información ausente con los valores esperados obtenidos. Se debe tener en cuenta en este caso que por información ausente no se entiende cada uno de los valores faltantes *Ymis*, sino las funciones de *Ymis* que intervienen en la función de log-verosimilitud para datos completos $l(\theta|Y)$.

El paso M determina $\theta(t+1)$ maximizando la función soporte obtenida en el paso E.

Las estimaciones iniciales de θ pueden ser realizadas mediante diferentes procedimientos alternativos: (1) análisis de datos completos, (2) análisis de datos disponibles, (3) imputación de los valores faltantes y (4) cálculo de las medias y varianza con los valores observados fijando las covarianzas a cero.

4.2.b Imputación múltiple (*Multiple Imputation*, MI)

Se trata de un método desarrollado a comienzos de la década de los 80. A diferencia de los métodos anteriores, que imputan un valor único a cada dato desconocido, la imputación múltiple se basa en la imputación de más de un valor para cada valor ausente. MI consiste en generar $m>1$ valores aleatorios para cada valor perdido por no respuesta, de manera que se disponga de m conjuntos de datos completos. Luego, se realizan los análisis estadísticos usuales a partir de cada uno de los m conjuntos de datos, generando m estimaciones. Finalmente, las distintas estimaciones son combinadas para producir una estimación con buenas propiedades estadísticas y con la posibilidad de estimar la varianza de las estimaciones.

El método MI consta de tres etapas: (1) cada valor perdido se reemplaza por un conjunto de $m>1$ valores generados por simulación, con lo que se crean m conjuntos de datos completos; (2) se aplica a cada uno de ellos el método de análisis deseado; y (3) los resultados obtenidos se combinan mediante reglas simples para producir una estimación global.

El objetivo de la imputación múltiple es hacer un uso eficiente de los datos que se han recogido, obtener estimadores no sesgados y reflejar adecuadamente la incertidumbre que la no respuesta parcial introduce en la estimación de los parámetros.

El número óptimo de bases de datos (m) depende del porcentaje de información faltante y se ubica entre tres y diez grupos.

Cada una de las m estimaciones anteriores se puede crear con una gran variedad de métodos, desde los más simples, como la imputación por media, hasta los más complejos, como los modelos de Monte Carlo con cadenas de Markov (MCMC-*Markov Chain Monte Carlo*). Inicialmente se desarrolló con técnicas de imputación simple para generar los valores a imputar; sin embargo, los métodos más utilizados en la actualidad son: aproximación bayesiana "*bootstrap*" y Monte Carlo con cadenas de Markov.

b.1 Imputación Múltiple Markov Chain Monte Carlo (MCMC)

Es uno de los procedimientos que se consideran más adecuados para generar imputaciones. MCMC es una colección de procesos de simulación generados por métodos de selección aleatoria mediante cadenas de Markov.

MCMC utiliza simulación paramétrica generando muestras aleatorias a partir de métodos bayesianos y, en el método MI, se aplica para generar las m selecciones independientes de valores faltantes, las cuales se utilizan en la etapa de inferencia.

Asumiendo que los datos provienen de una distribución normal multivariable, la agregación de los datos es aplicada desde la inferencia bayesiana a datos faltantes a través de la repetición de los siguientes pasos: (1) Imputación: con la estimación del vector de la media y la matriz de covarianzas, el primer paso consiste en simular los valores faltantes para cada una de las observaciones de forma independiente, (2) Distribución posterior: concluida la simulación del primer paso, se obtiene el vector de media de la población y de la matriz de covarianza de la muestra completa. Luego, estas nuevas estimaciones son usadas en el primer paso. Finalmente se realizan varias iteraciones. El objetivo es que estas iteraciones converjan a la distribución estacionaria y entonces se obtiene una estimación aproximada de los valores faltantes. El resultado de la estimación por algoritmo EM puede ser un buen valor inicial para comenzar el proceso MCMC.

5. Implementación del modelo de máxima verosimilitud (MV)

A partir de lo expuesto, se presume la existencia de ventajas comparativas que llevan a seleccionar un método de imputación múltiple. Ellas son: el incremento de la eficacia de los estimadores por la minimización de errores estándares, las inferencias válidas debidas a la combinación de

las inferencias obtenidas y la posibilidad de determinar la sensibilidad de las inferencias generadas a través de los diferentes modelos de no respuestas.

Pese a ello, los modelos de imputación múltiple poseen una importante desventaja: no producen una única respuesta. Debido a esto, Medina y Galván (2007) y Otero García (2011) recomiendan no hacer uso de procedimientos de estimación múltiples cuando se realizarán operaciones y clasificaciones de las unidades de análisis a partir de datos estimados. Asimismo (Medina y Galván, 2007: 36) expresan que

"un procedimiento que compite con los métodos de IM es el procedimiento de máxima verosimilitud (MV) que utiliza el algoritmo EM. Ambas propuestas aplican métodos numéricos y […] se demuestra que para tamaños de muestra grandes generan resultados similares."

La evidencia empírica y la descripción de cada uno de los modelos, con sus ventajas y desventajas, lleva a pensar que para el objetivo de esta investigación el mejor método de imputación de la no respuesta a las preguntas de ingresos es el modelo de máxima verosimilitud (MV).

En función de esto, con el objetivo de dar solución a la problemática originada por la no respuesta a las preguntas de ingresos, en esta investigación se realiza una imputación del valor de los ingresos no declarados por medio del procedimiento de máxima verosimilitud disponible en el Paquete Estadístico para Ciencias Sociales (SPSS 18.0). Dentro de las opciones de esta técnica, se seleccionó el método que considera el algoritmo EM puesto que la distribución de la no respuesta no es aleatoria.

En primer lugar, se imputan los ingresos de los individuos por perceptor y tipo de fuente que no fueron declarados, mediante el procedimiento de MV para cada una de las variables de ingresos recabadas por la EPH. Se parte del convencimiento de que, al realizar las estimaciones con la mayor desagregación posible de las preguntas de ingresos, disminuye el error en el que se puede incurrir. Posteriormente se agregan los ingresos de cada una de las fuentes para generar el ingreso total personal.

Además, para todos los años se estiman los ingresos de la ocupación principal y los ingresos horarios de la ocupación principal en forma independiente.

El modelo teórico propuesto considera como variables sociodemográficas, ocupacionales y económicas predictivas: el sexo, la edad, el máximo nivel de instrucción alcanzado, la relación con el jefe del hogar, la condición de actividad, la categoría ocupacional, la calificación laboral, el carácter de la tarea y la cantidad de ocupaciones (tabla 2).

Tabla 2. Variables predictivas y categorías utilizadas en procedimiento de imputación de valores

Variables	Categorías
Sexo	Varón Mujer
Edad	Hasta 24 años Entre 25 y 44 años Entre 45 y 64 años 65 años y más
Nivel de instrucción	Primario incompleto Secundario incompleto Secundario completo Superior o universitario completo
Relación con el jefe del Hogar	Jefe No jefe
Condición de actividad	Ocupado Desocupado Inactivo
Categoría ocupacional	Patrón o empleador Cuentapropista Obrero o empleado
Calificación laboral	Profesional Calificado No calificado
Carácter de la tarea	Producción Administrativo–Contable Comercialización Transporte, seguridad, y servicios
Cantidad de ocupaciones	Solo una ocupación Dos o más ocupaciones

Fuente: elaboración propia.

A partir de estas variables, por medio del procedimiento de imputación se procede a asignar valores en las fuentes de ingreso de los registros que no realizan la declaración de alguna de las preguntas.[11] Hipotéticamente, un perceptor

[11] Para cada tipo de ingreso, el procesamiento parte de la base de datos de la EPH, a la cual se le seleccionan los registros que declararon o tendrían que haber declarado el tipo de ingreso considerado. Posteriormente, el procedimiento de imputación aplicado con el SPSS genera una segunda base de datos que incluye solo el conjunto de

puede tener un ingreso declarado y uno o más ingresos no declarados; por medio de la imputación se reconstruye el total de los ingresos percibidos. La ecuación que lo interpreta es:

$$yTi = yD + yJ_1 + \ldots + yJ_n$$

donde yT_i es el ingreso total del perceptor i, yD es el ingreso declarado, yJ_1 es el ingreso imputado en el tipo de ingreso J_1, e yJ_n es el ingreso imputado en el tipo de ingreso J_n. De este modo se obtiene la estimación completa de los ingresos totales personales.

Por medio de otro procedimiento de MV, con las mismas variables predictoras, se realiza la imputación en forma independiente del ingreso de la ocupación principal y del ingreso horario de la ocupación principal.

6. La distribución de los ingresos personales pre y pos imputación

Considerando que la no respuesta de ingresos no se distribuye de forma aleatoria entre el total de perceptores, es posible que la consideración de los valores imputados ocasione variaciones en las distribuciones de ingresos.

Con el objetivo de evidenciar estos cambios se analizan tres distribuciones:

- Distribución de valores declarados: se genera al considerar solo los datos relevados (coincidiría con el procedimiento de *listwise*, que no considera los registros no respondentes para los análisis).
- Distribución de valores imputados: son los valores que surgen por medio de la aplicación de la técnica de imputación para los casos no declarantes.
- Distribución de variable completa: se origina al considerar en la misma distribución los datos relevados y los datos estimados por el procedimiento de imputación (valores declarados + valores imputados).

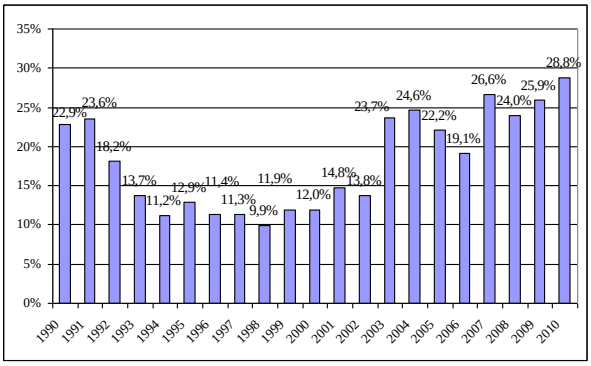
6.1. Efectos de la imputación sobre los ingresos relevados

La primera consecuencia de incluir en el análisis casi la totalidad de los perceptores monetarios es la posibilidad de integrar en los estudios de ingresos un volumen monetario que, de no realizarse la imputación, no sería considerado.

En la figura 2 se observa cómo la evolución del porcentaje de ingresos imputados respecto del total de ingresos personales (variable completa) fluctúa acompañando la incidencia de la no respuesta.[12] Este porcentaje es mínimo en 1998 y máximo en 2010. En estos años, de no haberse realizado el procedimiento de imputación, no se habrían considerado en los análisis 9,9 % y 28,8 %, respectivamente, del total de ingresos de los perceptores.

Figura 2. Efecto de la imputación en el ingreso total personal. Incidencia de los ingresos imputados en el total de ingresos personales Gran Buenos Aires: 1990-2010. Octubre 1990-2002 y 2.º semestre 2003-2010

En porcentaje de ingresos imputados respecto al total de ingresos de cada año



Fuente: elaboración propia con base en datos de la EPH, INDEC.

6.2. Efectos de la imputación sobre las medias de ingresos totales personales

Otra consecuencia de la integración casi total de los registros en los estudios de ingresos es la variación de la media. Tal como se expresó anteriormente, las características de los perceptores no respondentes se asocian, en su mayoría, a un perfil de ingresos más elevado que el promedio de perceptores. Debido a esto, la media de ingresos por perceptor se incrementa al realizar la imputación de datos.

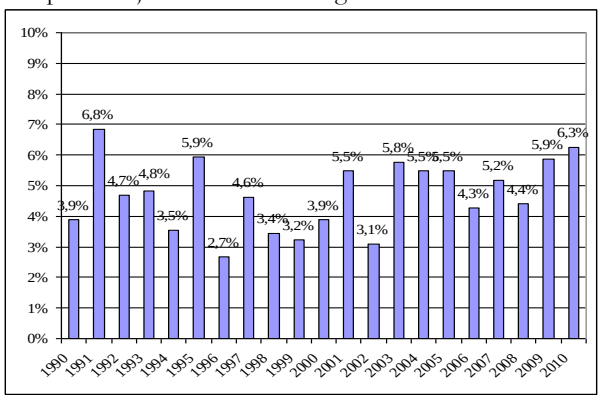
variables predictivas y la variable de ingreso que se desea imputar. Esta variable, en la nueva base, posee tanto los valores declarados como los valores imputados. Luego se procede a reemplazar, en la base de datos original y para la fuente de ingreso considerada, los valores no declarados por los imputados.

[12] El coeficiente de correlación de Pearson entre el porcentaje de perceptores que no declaran ingresos y el porcentaje del volumen de ingresos imputados es de 0,99.

Al considerarse la distribución completa, el promedio total de ingreso por perceptor aumenta con respecto a la media que surge de considerar solamente los valores declarados. El menor incremento observado es de 2,7 % en 1996, y el mayor, de 6,8 % en 1991 (figura 3). La limitada amplitud entre estos valores entre mínimo y máximo expresa el homogéneo comportamiento del procedimiento de imputación en los diversos escenarios socioeconómicos, los diferentes cuestionarios utilizados por la EPH y los cambios en el proceso de generación de información. Se puede considerar que solo una parte de la disparidad entre las variaciones se debe al porcentaje de no respuesta de cada año, pudiéndose determinar el resto por las particularidades de cada relevamiento, la composición de la no respuesta, la situación socioeconómica de contexto u otros factores.[13]

Figura 3. Efecto de la imputación de ingresos. Variación del ingreso medio por perceptor entre los valores declarados y los declarados más los imputados. Gran Buenos Aires: 1990-2010. Octubre 1990-2002 y 2.º semestre 2003-2010

En porcentaje de la media de ingreso declarado cada año



Fuente: elaboración propia con base en datos de la EPH, INDEC.

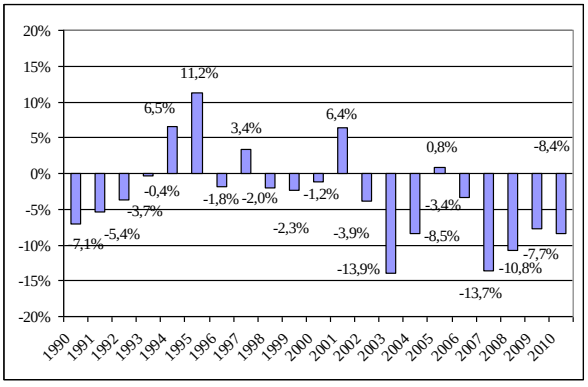
6.3. Efectos de la imputación en la dispersión de los ingresos

Con el fin de analizar el efecto generado por la imputación en la dispersión del ingreso personal, se compara el coeficiente de variación de la distribución completa con el coeficiente de la distribución declarada. La figura 4 presenta las variaciones porcentuales de los coeficientes de variación de ambas distribuciones; en la mayoría de las mediciones (76 % de los años considerados) se reduce la dispersión debido a la integración en el análisis de los valores imputados.[14]

La máxima reducción de dispersión se observa en los años 2003 y 2007 (13,9 % y 13,7 %, respectivamente) y el máximo incremento en 1995 (11,2 %). Considerando la totalidad de los años analizados, se observa que los datos imputados generan una marcada disminución de la dispersión al aumentar el porcentaje de no respuesta de los perceptores de ingresos.[15]

Figura 4. Efecto de la imputación de ingresos. Variación del coeficiente de variación del ingreso por perceptor entre la distribución completa y la distribución declarada. Gran Buenos Aires: 1990-2010. Octubre 1990-2002 y 2.º semestre 2003-2010

En porcentaje del coeficiente de variación de cada año respecto del declarado



Fuente: elaboración propia con base en datos de la EPH, INDEC.

6.4. Efectos de la imputación en los ingresos laborales y no laborales[16]

En la figura 5 se expone la evolución de la media y la mediana de los ingresos laborales del período 1990-2010, en pesos de octubre de 2010,

[13] Esta afirmación responde a que el coeficiente de correlación de Pearson entre las variaciones de las medias de ingreso total por perceptor y el porcentaje de no respuesta de los perceptores es de solo 0,56.

[14] Es importante considerar que la disminución de la varianza robustece los resultados obtenidos. Algunos autores (entre ellos INDEC, 2010; Otero García, 2011) presentan como una desventaja del procedimiento *hot deck* aleatorio el aumento de la varianza que produce.

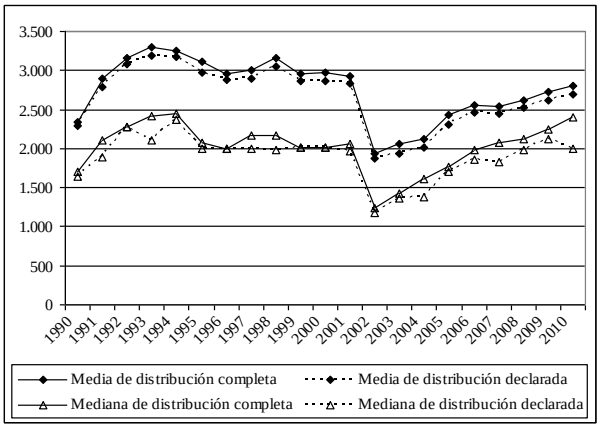
[15] El coeficiente de correlación de Pearson entre ambas distribuciones es de -0,78.

[16] Se considera ingresos laborales a las retribuciones al trabajo asalariado como obrero o empleado, utilidades derivadas de actividades por cuenta propia y ganancias empresariales. Por su parte, los ingresos no laborales están formados por jubilaciones o pensiones, rentas y utilidades financieras y transferencias públicas o privadas.

correspondiente a las distribuciones de ingresos declarados y completos. En ambas medidas, mayormente, los estadísticos que contienen los ingresos imputados poseen valores superiores a los que solo tienen los ingresos declarados. La gráfica de la evolución de ambas medias es similar: se incrementan los ingresos en períodos de expansión económica y disminuyen en los de contracción.

Figura 5. Efecto de la imputación de los valores no declarados en la media y la mediana de los ingresos laborales Gran Buenos Aires: 1990-2010. Octubre 1990-2002 y 2.º semestre 2003-2010

En pesos de octubre de 2010



Fuente: elaboración propia con base en datos de la EPH, IPC-INDEC y CENDA (2011b).

Entre el promedio de ingresos declarados y el de la distribución completa, la mínima variación es de 2,2 % y la máxima de 5,6 %, en 1990 y 2003 respectivamente. El limitado rango de las variaciones puede interpretarse como un indicador de la solidez y previsibilidad de las imputaciones realizadas mediante el procedimiento de máxima verosimilitud.

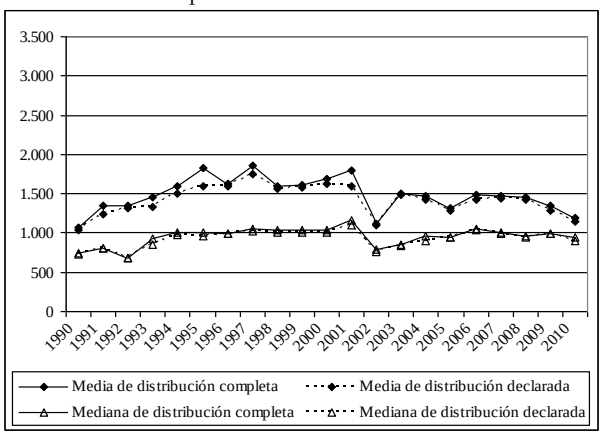
Al examinar la evolución de las medianas de ingresos declarados y las de la distribución completa (figura 5) se observa que ambas gráficas se acompañan, pero poseen variaciones pronunciadas; en algunos años no existe diferencia entre los valores (1992, 1996, 1999 y 2000), mientras que en otros la diferencia es muy importante (20 % en 2010).

En los primeros casos, la imputación de ingresos no alteró el valor del primer 50 %, ya sea porque no se estimaron casos en ese tramo de los datos o por la elevada concentración de datos en el valor de la mediana de datos declarados. En las variaciones extremas (1991, 1993, 2004, 2007 y 2010) todo indica que el valor que surge de los valores imputados completa los datos declarados de una forma eficiente, ya que amortigua abruptas variaciones de la mediana de ingresos de los datos declarados (figura 5).

La figura 6 presenta las evoluciones de las medias y medianas de los ingresos no laborales declarados y de la distribución con datos completos. Se observa en este tipo de ingresos un grado levemente mayor de independencia de las situaciones económicas debido al importante peso relativo que poseen en ellos las jubilaciones y pensiones.

Figura 6. Efecto de la imputación de los valores no declarados en la media y la mediana de los ingresos no laborales. Gran Buenos Aires: 1990-2010. Octubre 1990-2002 y 2.º semestre 2003-2010

En pesos de octubre de 2010



Fuente: elaboración propia con base en datos de la EPH, IPC-INDEC y CENDA (2011b).

La evolución de la media de ingresos no laborales es similar para la distribución que integran los valores imputados y la que solo posee los declarados; exceptuando los años 1991, 1993, 1995 y 2001, cuando las variaciones son de 8,7 %, 9,3 %, 14,6 % y 11,7 %, respectivamente.

La evolución de las medianas conserva la forma para ambas distribuciones y presenta una mínima variación entre los valores de la distribución completa y los valores declarados. Los mayores valores son 9,6 % en 1993 y 8,3 % en 2004.

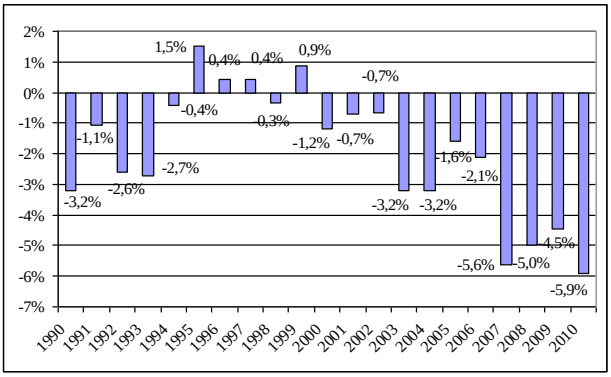
De la consideración de ambos tipos de ingresos y medidas se puede concluir la importancia de realizar la imputación para evaluar los ingresos de la población con una mejor calidad. De esta manera se considera la totalidad de los perceptores y se anulan los factores externos al fenómeno que se desea medir.

6.5. Efectos de la imputación en el coeficiente de desigualdad de Gini

En la figura 7, que presenta las variaciones porcentuales entre el valor del coeficiente de Gini encargado de expresar la desigualdad de la distribución declarada de ingresos laborales y el de Gini correspondiente a la distribución completa del mismo tipo de ingreso, se observa una tendencia según la cual la imputación de ingresos laborales disminuye la desigualdad entre los perceptores: solo en 20 % de los años la variación del Gini fue cercana a cero; en el resto de las mediciones el valor del coeficiente de desigualdad disminuye. Si se comparan los datos presentados con los porcentajes de no respuesta de los perceptores (figura 1), se advierte que ante un mayor porcentaje de imputación disminuye marcadamente el valor del coeficiente de Gini.

Figura 7. Efecto de la imputación de valores de los ingresos laborales en el coeficiente de Gini. Variación porcentual entre los coeficientes de Gini de la distribución declarada y la distribución completa. Gran Buenos Aires: 1990-2010. Octubre 1990-2002 y 2.º semestre 2003-2010

Porcentaje de variación respecto de coeficientes de Gini de la distribución declarada

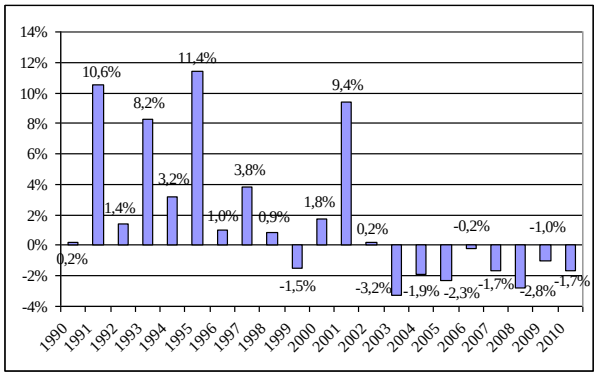


Fuente: elaboración propia con base en datos de la EPH, INDEC.

En la figura 8 se presentan las variaciones porcentuales entre los valores del coeficiente de Gini que expresa la desigualdad de la distribución declarada de ingresos no laborales y el correspondiente a la distribución completa del mismo tipo de ingreso. En este caso no se puede aseverar que la imputación de ingresos altere con un patrón definido la desigualdad de ingresos no laborales. El procedimiento de imputación de valores produce un aumento de la desigualdad en la mitad de los años considerados.

Figura 8. Efecto de la imputación de valores de los ingresos no laborales en el coeficiente de Gini. Variación porcentual entre los coeficientes de Gini de la distribución declarada y la distribución completa. Gran Buenos Aires: 1990-2010. Octubre 1990-2002 y 2.º semestre 2003-2010

Porcentaje de variación respecto de coeficientes de Gini de la distribución declarada



Fuente: elaboración propia con base en datos de la EPH, INDEC.

De lo expuesto se deduce una tendencia a una leve disminución en la desigualdad de la distribución de ingresos laborales como producto de la imputación de los valores no declarados. En lo que respecta a los ingresos no laborales, si bien no se identifica un patrón definido, la imputación de valores genera una leve tendencia a aumentar la desigualdad.

Todas estas observaciones muestran que el procedimiento de imputación de valores permite recuperar casi la totalidad de los registros no declarados. Se presentaron muy pocos casos en los que no se pudo realizar la estimación. La imposibilidad de la imputación se debió a que esos registros no poseen información en las variables de predicción o bien el procedimiento de MV obtiene estimaciones de valor cero que no son aceptadas como válidas. Dichos registros, que en ninguno de los años analizados supera el 0,5 % del total de perceptores, son excluidos de las bases analizadas. Consecuentemente, se puede afirmar que la implementación del procedimiento de imputación MV permite trabajar con casi la totalidad de los hogares y la población representada en la muestra.[17]

[17] De la comparación entre las bases de datos originales de la EPH y las bases completas obtenidas luego de la imputación, para el período 1990-2010 se verificó que el procedimiento aplicado logró obtener en los años más desfavorables definiciones de ingresos del 99,2 % de los hogares y del 99,4 % de la población.

Por otra parte, la integración de casi la totalidad de los casos al análisis genera cambios en los principales estadísticos sociales: se incrementa el promedio de ingresos por perceptor, generalmente disminuye la dispersión de las distribuciones de ingresos y aumenta el valor de corte de los percentiles. Asimismo, a causa de la menor no respuesta de los ingresos de jubilaciones y pensiones, el valor de la media y la mediana aumenta en mayor proporción en los ingresos laborales que en los no laborales.

7. Conclusiones

Se determinó que la incidencia de la no respuesta a las preguntas de ingreso varía considerablemente en el período analizado. El porcentaje de no respuesta disminuyó marcadamente entre 1990 y 1994, se estabilizó relativamente entre 1995 y 2000, se incrementó levemente hasta 2002, y aumentó ligeramente a partir de 2003 y marcadamente a partir de 2007.

Debido a esto, las sub-muestras que pueden surgir de solo considerar los registros con datos completos estarían ampliamente sesgadas y no representarían al universo del que se desea predicar. A partir de estas conclusiones se plantea la necesidad de realizar una imputación a los registros de las preguntas de ingreso no respondidas para poder trabajar con una muestra que conserve la pretensión de ser representativa del universo analizado.

Se verificó que estas fluctuaciones dependen de cuestiones técnico-metodológicas y del contexto socioeconómico. Dentro de las primeras, se identificó el impacto de los cambios de cuestionarios de 1995 y 2003, del incremento de las preguntas referidas al ingreso en los cuestionarios individuales, de las acciones de capacitación de encuestadores, del aumento de la supervisión y del incremento de instancias de recuperación de información. Por otra parte, se verificó que en contextos económicos de acentuado incremento del costo de vida aumentó la no respuesta a las preguntas de ingreso debido a la pérdida de la noción relativa del monto percibido y a ciertos factores sociales que han incidido alterando la propensión de la población a contestar las preguntas de ingresos o, directamente, a colaborar en los operativos oficiales de relevamiento de información por falta de credibilidad en el organismo responsable.

Posteriormente, se reseñaron los principales métodos de imputación de registros faltantes y se identificó el procedimiento de máxima verosimilitud como uno de los más aptos para la problemática analizada. Este procedimiento se aplicó a los registros no respondidos de todas las variables de ingresos de la Encuesta Permanente de Hogares del Gran Buenos Aires de 1990 a 2010. Se reconstruyeron los ingresos por perceptor, generándose una base de datos con información casi completa de los ingresos, gracias a la cual se pudo analizar más de un 99 % de los perceptores de ingresos y de los hogares.

Asimismo, se determinó la importancia de las potencialidades del procedimiento y su impacto en las estadísticas sociales. Sobre este aspecto, se aprecia con claridad que el procedimiento de máxima verosimilitud (MV) permite considerar casi la totalidad de los ingresos monetarios y que la incidencia en los indicadores utilizados para analizar las condiciones de vida de la población son sumamente relevantes: se incrementan los ingresos medios por perceptor. Específicamente, en los ingresos laborales aumenta el valor de la media y la mediana. Por su parte, la imputación de ingresos no laborales genera un menor impacto en los estadísticos a causa de la escasa incidencia de la no respuesta en los ingresos relativamente fijos de las jubilaciones.

Por todo lo expresado anteriormente, se remarca, como estrategia supletoria ante la no respuesta, la importancia de realizar imputaciones válidas y fiables de los ingresos no declarados.

8. Referencias

Arrazola, M. y Hevia, J. (2005). "Estimación de los efectos de un tratamiento: una aplicación a la educación superior en España". En VV. AA. (eds.), *Actas de las VI Jornadas de Economía Laboral*, Alicante: Universidad de Alicante.

Beccaria, L. (1998). *Criterios operativos de las encuestas de hogares y la medición de los ingresos. Programa para el mejoramiento de las encuestas y la medición de las condiciones de vida en América Latina y el Caribe*, Buenos Aires: MECOVI.

Berumen, E. y Muñoz, J. (1996). *Aspectos del diseño y la puesta en marcha de las encuestas que inciden en la calidad de los datos recogidos. Programa para el mejoramiento de las encuestas y la medición de las*

condiciones de vida en América Latina y el Caribe, Paraguay: MECOVI.

BID (2004). "Imputación múltiple y los missing de ingresos en las encuestas de hogares". 14.º Taller regional MECOVI: *Imputación de Datos en las Encuestas de Hogares: Los Procedimientos Metodológicos y sus Implicaciones*. Buenos Aires, 17 al 19 de noviembre de 2004.

Camelo, H. (1998). *Subdeclaración de ingresos medios en las encuestas de hogares, según quintiles de hogares y fuente de ingresos. Programa para el mejoramiento de las encuestas y la medición de las condiciones de vida en América Latina y el Caribe*, Buenos Aires: MECOVI.

CENDA (2008). *El trabajo en Argentina: Condiciones y perspectivas. Informe trimestral n.° 14*.

— (2011a). *El trabajo en Argentina: Condiciones y perspectivas. Informe trimestral n.° 20*.

— (2011b). *IPC - 7 Provincias*. Centro de Estudios para el Desarrollo Argentino. Consulta: 5 de junio de 2011, www.cenda.org.ar

Crosta, F. (2000). *La medición de la pobreza en la Argentina. Revisión metodológica y estimaciones. Trabajo de tesis de la Maestría de Economía*, La Plata: Facultad de Ciencias Económicas, Universidad Nacional de La Plata.

Donza, E. (2011). "Incidencia de la no respuesta a las preguntas de ingresos en la Encuesta Permanente de Hogares, consideraciones teóricas y efectos. Gran Buenos Aires 1990-2010". En VV. AA. (eds.), ponencia presentada en las *IX Jornadas de Sociología*, Buenos Aires: UBA.

Felcman, D.; Kidyba, S. y Ruffo, H. (2004). *Medición del ingreso laboral: Ajustes a los datos de la Encuesta Permanente de Hogares para el análisis de la distribución del ingreso (1993-2002)*. 14º Taller regional MECOVI, cit.

Feres, J. (1998). *Falta de respuestas a las preguntas sobre el ingreso. Su magnitud y efectos en las encuestas de hogares de América Latina. Programa para el mejoramiento de las encuestas y la medición de las condiciones de vida en América Latina y el Caribe*, Buenos Aires: MECOVI.

— (2004). *Confiabilidad de la medición del ingreso en las encuestas de hogares*, CEPAL, 14º Taller regional MECOVI, cit.

González Espiti, C. (s/f). *Sesgo de selección muestral con STATA*. Cali: Departamento de Economía, Universidad Icesi. Consulta: 7 de febrero de 2011, http://www.icesi.edu.co/e_portafolio/artefact/file/download.php?file=2736&view=286

INDEC (1984). "La pobreza en Argentina". *Series Estudios n.º 1*, Buenos Aires.

— (1989). "Perfil y estrategias de reformulación temática de la EPH". En *Segunda reunión del Comité de Expertos en Estadísticas Sociodemográficas*, Buenos Aires, agosto 22-25.

— (1995). "Encuesta Permanente de Hogares. Desarrollo actual y perspectiva". En *Seminario Internacional sobre medición del empleo*. Buenos Aires, diciembre.

— (1997a). "Evaluación de la calidad de los datos y avances metodológicos. 1.ª parte", Censo Nacional de Población y Viviendas 1991, Serie J, n.º 2, Buenos Aires.

— (1997b). *Encuesta Nacional de Gastos de los Hogares 1996/97*, Buenos Aires.

— (1998). "Encuesta a Hogares: Reformulación de la Encuesta Permanente de Hogares de Argentina". En *Primera reunión sobre estadística pública del Instituto Interamericano de Estadística*. Buenos Aires.

— (2003a). *La nueva Encuesta Permanente de Hogares de Argentina*, Buenos Aires.

— (2003b). *Encuesta Permanente de Hogares (EPH). Cambios metodológicos*, Buenos Aires.

— (2006). *Encuesta Nacional de Gastos de los Hogares 2004/2005*, Buenos Aires.

— (2010). *Ponderación de la muestra y tratamiento de valores faltantes en las variables de ingreso en la Encuesta Permanente de Hogares Metodología n.° 15*, Buenos Aires.

Keifman, S., Manzano, G., Rodríguez, C. y Viler, A. (1998). "Imputación de ingresos de hogares: la experiencia de la Encuesta Nacional de Gastos de la Argentina", *Programa para el mejoramiento de las encuestas y la medición de las condiciones de vida en América Latina y el Caribe*, Buenos Aires: MECOVI.

Lindenboim, J., Graña, J. y Kennedy, D. (2006). "Concepto, medición y utilidad de la distribución funcional del ingreso. Argentina 1993–2005". En AA. VV., *V Jornadas sobre Mercado de Trabajo y Equidad en Argentina, Área Estado, Mercado y Actores Sociales en la Argentina Contemporánea*, Buenos Aires: Instituto de Ciencias, Universidad Nacional de General Sarmiento.

Llach, J. J. y Montoya, S. (1999). *En Pos de la Equidad. La pobreza y la distribución del ingreso en*

el Área Metropolitana de Buenos Aires: diagnóstico y alternativas de políticas, Buenos Aires: IERAL.

MECOVI (2004). "Resumen y Conclusiones", 14.° *Taller Regional Imputación de datos en las encuestas de hogares: los procedimientos metodológicos y sus implicaciones*, cit.

Medina, F. (2004). "Los métodos de imputación de datos en las encuestas de hogares: Teoría y práctica", 14.° *Taller Regional MECOVI*, cit.

Medina, F. y Galván, M. (2007). *Imputación de datos: teoría y práctica. Serie Estudios estadísticos y prospectivos 54*, Santiago de Chile: CEPAL.

Muñoz, J. (1996). "¿Cómo mejorara la calidad de la información? Opciones para mejorar la organización del trabajo de campo, el sistema de entrada de datos, el análisis de consistencia y el manejo de la base de datos". *Programa para el mejoramiento de las encuestas y la medición de las condiciones de vida en América Latina y el Caribe*, Paraguay: MECOVI.

Otero García, D. (2011). "Imputación de datos faltantes en un Sistema de Información sobre Conductas de Riesgo. Tesis de Máster Interuniversitario". En *Técnicas Estadísticas*. Santiago de Compostela: Universidad de Santiago de Compostela, Universidad de La Coruña, Universidad de Vigo.

Perlbach de Maradona, I. y Calderón, M. (s/f). *Estimación del sesgo de selección para el mercado laboral de Mendoza*. Consulta: 7 de febrero de 2011, http://cdi.mecon.gov.ar/biblio/docelec/aaep/98/perlbach-de-maradona_calderon.pdf

Ramírez, G. (1998). "Imputación de datos". Presentación en el *1.er Taller de MECOVI. Planificación y Desarrollo de Encuestas de Hogares para la Medición de las Condiciones de Vida*, Aguascalientes: MECOVI.

Roca, E. y Pena, H. (2001). "La Declaración de Ingresos en las Encuestas de Hogares". En AA. VV. (eds.), *Actas del 5.° Congreso Nacional de Estudios del Trabajo*, 1 al 3 de agosto, Buenos Aires: ASET.

Salvia, A. y Donza, E. (1999). "Problemas de medición y sesgos de estimación derivados de no respuesta a las preguntas de ingresos en la Encuesta Permanente de Hogares (1990-1998)". *Revista de la Asociación Argentina de Especialistas en Estudios del Trabajo*, n.° 18, Buenos Aires.

Tenjo Galarza, J.; Ribero Medina, R. y Bernat Díaz, L. (2006). "Evolución de las diferencias salariales de genero en seis países de América Latina". En Piras C. (ed.), *Mujeres y trabajo en América Latina. Desafíos para las políticas públicas*, BID.

Trabuchi, C. y Pok, C. (1995). "Encuesta Permanente de Hogares: Desarrollo actual y perspectivas", Seminario Internacional sobre medición del empleo, 5 al 7 de diciembre, Buenos Aires.

Zamudio Carrillo, A. (1995). "Rendimientos a la educación superior en México: ajuste por sesgo utilizando máxima verosimilitud". *Economía Mexicana Nueva Época*, vol. IV, n.° 1, 5-67.