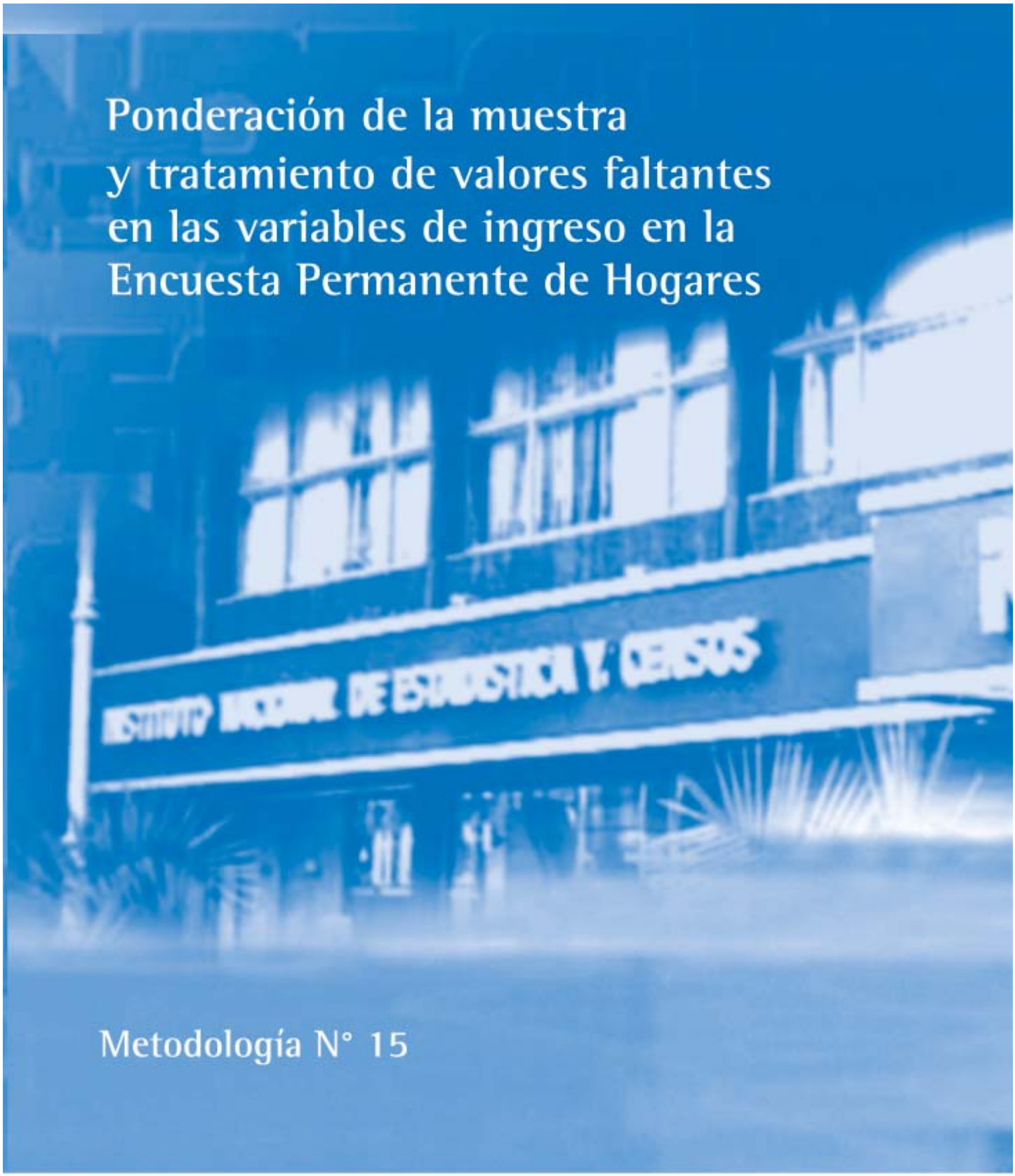
Ponderación de la muestra y tratamiento de valores faltantes en las variables de ingreso en la Encuesta Permanente de Hogares



Metodología N° 15

REPÚBLICA ARGENTINA
MINISTERIO DE ECONOMÍA Y FINANZAS PÚBLICAS
INSTITUTO NACIONAL DE ESTADÍSTICA Y CENSOS

INDEC

Ponderación de la muestra y tratamiento de valores faltantes en las variables de ingreso en la Encuesta Permanente de Hogares

Dirección de Metodología Estadística
Dirección de Encuesta Permanente de Hogares

Claudio Comari	Director de Encuesta Permanente de Hogares
Augusto Hoszowski	Dirección de Metodología Estadística

Índice

Agradecimientos

En el marco de la elaboración de los avances metodológicos expuestos en este documento, se realizaron numerosas reuniones con especialistas, cuyas críticas, observaciones y sugerencias fueron por demás valiosas y a quienes queremos agradecer la generosidad con que aportaron su conocimiento y experiencia para el mejoramiento de nuestra producción estadística.

Muchas gracias a la Dirección del INDEC en la persona de Ana María Edwin, a nuestros colegas de la Dirección de Encuesta Permanente de Hogares, de la Dirección de Metodología Estadística, de la Dirección Nacional de Cuentas Nacionales y de la Dirección Nacional de Estadísticas Sociales y de Población, a los titulares y técnicos de las Direcciones Provinciales de Estadística, a los expertos extranjeros: Minezo Fujita, consultor de la Oficina de Estadística de Japón; Jacques Freyssinet, presidente del Comité Científico del Centre d' Etudes de l'Emploi de Francia; Dominique Goux, (chef de mission - animation de la recherche) de la Direction de l'Animation de la Recherche, des Etudes et des Statistiques (DARES, Ministerio de Trabajo de Francia); Giulio Barcaroli, Marco Di Zio, Gianluca Giuliani, Federica Pintaldi, Paolo Consolini y Silvano Vitaletti del ISTAT de Italia y, en particular, a Juan Carlos Feres, Fernando Medina, Marco Galván y Carlos Howes de la División de Estadística y Proyecciones Económicas de la CEPAL.

Subrayando los destacados aportes y esfuerzos de todo el equipo de trabajo y, muy especialmente, de María Alejandra Jorge y María Vargas. A todos los mencionados nuestro mayor reconocimiento, de ellos son los aciertos, de los firmantes los errores.

Introducción

Como en toda encuesta por muestreo, las estimaciones que surgen de la Encuesta Permanente de Hogares están afectadas por errores debidos al muestreo (consecuencia de haber encuestado sólo a una fracción de las viviendas en cada dominio) y errores no debidos al muestreo: no respuesta, respuestas incoherentes, etc.

La incidencia y magnitud de estos errores son, en general, mínimas, y los diferentes análisis que se han hecho a lo largo de la historia de la EPH muestran la alta calidad de la información que surge de la misma.

Las innovaciones metodológicas que se introducen a partir del tercer trimestre de 2009 tienen como objetivo mejorar aún más la calidad de la información que la EPH proporciona, así como la calidad de las bases de microdatos que el INDEC pone a disposición de los usuarios, en dos temáticas:

- Ponderación de la muestra.
- Tratamiento de los datos faltantes en las variables de ingreso.

Los cambios introducidos permitirán a los usuarios que utilizan las bases de microdatos de la EPH realizar análisis más precisos en general y en las temáticas de ingreso en particular.

A continuación se presenta una breve reseña de estos dos avances, sus efectos e implicancias.

I. CALIBRACIÓN DE LOS FACTORES DE EXPANSIÓN

Las estimaciones de totales en una encuesta por muestreo fluctúan debido al error muestral. En una encuesta a hogares, según la muestra de viviendas seleccionadas, se pueden estimar más o menos personas, más o menos ocupados, etc., dependiendo de la composición de los hogares residentes en las viviendas. Según el tamaño de la muestra y el dominio bajo estudio, estas fluctuaciones tienen mayor o menor impacto en las estimaciones.

Desde su inicio en 2003, para la EPH Continua se utiliza un *estimador de razón* para atenuar estas fluctuaciones:

> *Cada trimestre, se ajustan los factores de expansión de diseño (inversa de la probabilidad de selección) para que las estimaciones de población de cada aglomerado coincidan con las proyecciones de población realizadas por la Dirección de Estadísticas Poblacionales del INDEC.*

Esta corrección, si bien determina que no haya variaciones muestrales del total estimado de población, no corrige las variaciones muestrales al *interior de la población*, e.g. en grupos de sexo y tramos de edad. Estas variaciones pueden incidir artificialmente sobre las estimaciones: el aumento o disminución del total de desocupados o activos en cierto tramo de edad, de un trimestre a otro, puede deberse a que en dicho tramo etario hay más o menos población en la muestra de viviendas encuestadas. Una segunda implicación del error muestral en este tipo de encuestas es que dos encuestas a hogares, en igual dominios y períodos, darán estimaciones diferentes en general.

Una técnica que permite atenuar parcialmente estos efectos es la *calibración*.

Dados ciertos totales poblacionales de subpoblaciones (e.g. total de población según sexo y tramos de edad, total de hogares unipersonales, etc.) conocidos por *fuentes externas* (registros administrativos, proyecciones demográficas), y denominados en general valores *marginales poblacionales*, se ajustan las ponderaciones para que los totales estimados coincidan con estos valores externos.

Formalmente, sean $x_1,...x_k$ k variables auxiliares cuyos totales poblacionales $X_1, ...X_k$ conocemos. En una encuesta a hogares los X_j son típicamente totales de población por sexo o tramos de edad, cuyos valores 'verdaderos' se conocen por proyecciones demográficas, registros, etc..

Y sean $d_1,...,d_n$ los factores de expansión originales (inversa de las probabilidades de selección, corregidos por no respuesta) de los elementos de la muestra $i=1,..., n$.

Se busca entonces ajustar estas ponderaciones mediante ciertos factores $c_1,...c_n$ de forma que las estimaciones de los totales $X_1,...X_k$ no tengan error, i.e:

$$\Sigma\, c_i * d_i * x_{ji} = X_j, \text{ para } j=1,...,k$$

Este ajuste se realiza de forma que los factores de expansión *finales* estén lo más próximo posible de los *iniciales* y respeten ciertas condiciones, como ser no negativos, o no duplicar al factor original, etc.

Formalmente, se elige una función $G(w, d)$, que cuantifique la distancia entre los factores originales y los nuevos, que cumpla las condiciones básicas:

- $G(z,z) = 0$
- $G \geqslant 0$
- G estrictamente convexa
- G derivable en w

Si w_i es el nuevo factor de expansión y d_i es el factor de expansión original, se trata entonces de minimizar la función

$$\Sigma\, G(w_i,d_i)$$

sujeta a las condiciones

$$c_i * d_i * x_{ji} = X_j, \text{ para } j=1,...,k$$

Si $w_i = d_i$ entonces $G(w_i,d_i)=0$, por la primer condición .

El origen de esta técnica es el *raking-ratio*, sugerido por primera vez por los estadísticos W.E. Deming y F.F. Stephan en 1940. Esta técnica esencialmente consiste en ajustar iterativamente los factores de expansión de una muestra según dos variables categóricas, cuyos marginales (totales en cada categoría) se conocen. Esto permite estimar correctamente las categorías de una y otra variable.

Por ejemplo, sean dos variables categóricas sexo y tramos de edad (0-14, 15-29, 30-49, 50-64, 65 o más) y cuyos totales poblacionales, en cada categoría, son conocidos: se ajustan los factores de expansión para estimar correctamente el total de personas según sexo, luego se ajustan las factores de expansión para estimar correctamente las personas según tramo de edad, luego se vuelve a ajustar las ponderaciones para estimar correctamente las personas según sexo, y así sucesivamente.

La calibración es considerada una post-estratificación incompleta porque las estimaciones ajustarán los marginales de una y otra variable en forma independiente, pero no las celdas de cruce. Por ejemplo, un marginal es el total de población entre 0 y 14 años y otro el total de mujeres, pero no se ingresa el total de mujeres entre 0 y 14 años.

Esto tiene dos ventajas:

- No requiere conocer los efectivos poblacionales en las ***n x m*** celdas.
- Evita para celdas con pocos casos que una post-estratificación modifique demasiado los factores de expansión originales.

Posteriormente, la calibración amplió esto al caso de variables explicativas continuas.

La técnica de calibración se popularizó al estar disponible para el software SAS una macro escrita por el INSEE[1] de Francia, CALMAR[2] . Aplicativos similares están disponibles actualmente en diversos 'soft' estadísticos.

Hay un multitud de variantes de esta técnica, permitiendo la mayoría de ellas que el usuario acote el factor de ajuste. Para la calibración de las bases de la EPH se determinó una cota superior de 2.5 y una cota inferior de 0.25. En la práctica el 90% de estos factores se encuentran en el intervalo [0.7;1.3] (ver cuadro "Percentiles del factor de ajuste de las ponderaciones").

Remitimos al lector al artículo de Sing A. y Mohl C. (1996)[3] para una exposición formal de la calibración.

---

[1]Institut National de la Statistique et des Études Économiques

[2]CALage sur MARges

[3]Singh, Avi , Mohl, Chris (1996). *Understanding calibration estimators in survey sampling*. Survey Methodology, vol. 22 n°2.

Objetivos de la calibración

La calibración tiene en general cuatro objetivos:

1. Reducir la varianza de los estimadores de totales *correlacionados* con los marginales poblacionales. Luego de calibrar los ponderadores, la estimación de los marginales tiene varianza cero, por lo que es de esperar que estimadores de totales correlacionados o asociados con estos marginales vean disminuida su varianza.
2. Facilitar los análisis longitudinales, al eliminar de un período a otro las oscilaciones artificialmente introducidas por la composición demográfica de la muestra. Si bien la cantidad de viviendas seleccionadas en la EPH se mantiene constante, varía la cantidad y estructura demográfica de las personas que habitan en ellas al cambiar, por efecto de la rotación, las viviendas en el panel. A ello se suman los efectos producidos por la *no respuesta*.
3. Disminuir posibles sesgos en muestras que presentan *no respuesta* o subpoblaciones difíciles de captar, aumentando la ponderación de estas subpoblaciones.
4. Posibilitar que las estimaciones de ciertos totales surgidas de encuestas oficiales a hogares sean coherentes entre sí, para un determinado domino y período.

Se trabaja con el supuesto implícito de que los marginales que se utilizan son correctos, i.e., coinciden con los valores *poblacionales*. Si esto no sucede, el tercer punto puede no cumplirse, pero seguirán cumpliéndose en general los puntos 2 y 4. Aunque esta técnica se origina para reducir la variancia de los estimadores de totales, son los objetivos segundo y cuarto los más relevantes en las encuestas oficiales, al dar coherencia a las estimaciones difundidas.

Conjunto de marginales a utilizar en la calibración

Un aspecto relevante en el proceso de calibración es determinar los marginales a utilizar. Si bien aparentemente muchos marginales 'mejorarán' la calidad, el utilizar mayor información externa obliga al método a modificar en mayor medida los pesos originales, pudiendo así aumentar sensiblemente la varianza de los estimadores de variables que no estén muy correlacionadas con los marginales. La selección de marginales requiere entonces soluciones equilibradas adecuadas al estudio en particular.

En el caso de la calibración de la EPH se optó por una postura conservadora, tomando como marginales las siguientes proyecciones:

- Total de varones.
- Total de mujeres.
- Total de población entre 0 y 14 años.
- Total de población entre 15 y 29 años.
- Total de población entre 30 y 49 años.
- Total de población entre 50 y 64 años.
- Total de población de 65 años o más.

Los tramos de edad son seleccionados en virtud de la asociación de la condición de actividad con rangos etarios.

Etapas del proceso general de construcción de los ponderadores

Para obtener los factores finales de expansión se parte de:

- Probabilidades de selección corregidas por no respuesta.
- Proyecciones de población para el aglomerado y el trimestre.
- Estructura de población según sexo y tramos de edad, para los departamentos que incluyen al aglomerado, por año (una estructura proyectada por año).

Mediante las estructuras de población según sexo y tramo de edad (una estructura por año) y el total proyectado de población para el aglomerado y el trimestre, se determinan los totales de población por sexo y tramo de edad, que serán los marginales a utilizar en el proceso de calibración.

El método que se utilizó es una modificación del método de Huang-Fuller[4], implementado en Stata 9.0.

[4] Singh, Avi , Mohl, Chris (1996). *Understanding calibration estimators in survey sampling*. Survey Methodology, vol. 22 no. 2.

Dispersión de los factores de ajuste

A continuación, a modo de ejemplo, se presenta un cuadro que muestra la dispersión de los factores de ajuste de la calibración para los trimestres correspondientes a los años 2004 a 2006 inclusive.

Percentiles del factor de ajuste de las ponderaciones
***Ajuste=Peso/Pondera* siendo Peso= Pondera calibrado**

Periodo	Percentil 5	Percentil 25	Percentil 50	Percentil 75	Percentil 95
1° T 2004	0.77	0.92	1.00	1.08	1.23
2° T 2004	0.78	0.91	0.99	1.08	1.23
3° T 2004	0.79	0.92	1.00	1.07	1.21
4° T 2004	0.78	0.92	1.00	1.07	1.21
1° T 2005	0.78	0.93	1.00	1.08	1.23
2° T 2005	0.77	0.92	0.99	1.08	1.25
3° T 2005	0.77	0.91	1.00	1.09	1.26
4° T 2005	0.78	0.93	1.01	1.08	1.22
1° T 2006	0.80	0.93	1.01	1.08	1.20
2° T 2006	0.78	0.93	1.00	1.09	1.22
3° T 2006	0.78	0.93	1.01	1.09	1.22
4° T 2006	0.79	0.93	1.00	1.08	1.22

Cambios en las estimaciones

La calibración introduce modificaciones mínimas en las estimaciones de las principales tasas del mercado laboral. Se presenta a continuación las tasas de actividad, empleo y desocupación antes y luego de la calibración para los tres años considerados.

Total de aglomerados

Tasas de actividad, empleo y desocupación según ponderadores sin calibrar y calibrados

Periodo	Tasa actividad sin calib.	Tasa actividad con calib.	Tasa empleo sin calib.	Tasa empleo con calib.	Tasa desocup. sin calib.	Tasa desocup. con calib.
1° T 2004	45.4	46.1	38.9	39.5	14.4	14.3
2° T 2004	46.2	46.7	39.4	39.8	14.8	14.7
3° T 2004	46.2	46.6	40.1	40.5	13.2	13.1
4° T 2004	45.9	46.4	40.4	40.8	12.1	12.0
1° T 2005	45.2	46.0	39.4	40.0	13.0	12.9
2° T 2005	45.6	46.0	40.1	40.5	12.1	12.0
3° T 2005	46.2	46.7	41.1	41.5	11.1	11.1
4° T 2005	45.9	46.4	41.3	41.7	10.1	10.0
1° T 2006	46.0	46.4	40.7	41.2	11.4	11.3
2° T 2006	46.7	46.9	41.8	42.1	10.4	10.3
3° T 2006	46.3	46.7	41.6	42.0	10.2	10.1
4° T 2006	46.1	46.4	42.1	42.4	8.7	8.6

Estructura según sexo

Periodo	Sexo	% de población con calibración	% de población sin calibración
1° T 2004	Varón	47.9	47.4
2° T 2004	Varón	47.9	47.2
3° T 2004	Varón	47.9	47.5
4° T 2004	Varón	47.9	47.8
1° T 2005	Varón	47.9	47.7
2° T 2005	Varón	48.0	47.8
3° T 2005	Varón	48.0	47.3
4° T 2005	Varón	48.0	47.5
1° T 2006	Varón	48.0	47.8
2° T 2006	Varón	48.0	47.7
3° T 2006	Varón	48.0	47.5
4° T 2006	Varón	48.0	47.4

Estructura según sexo

Periodo	Sexo	% de población con calibración	% de población sin calibración
1° T 2004	Mujer	52.1	52.6
2° T 2004	Mujer	52.1	52.8
3° T 2004	Mujer	52.1	52.5
4° T 2004	Mujer	52.1	52.2
1° T 2005	Mujer	52.1	52.3
2° T 2005	Mujer	52.0	52.2
3° T 2005	Mujer	52.0	52.7
4° T 2005	Mujer	52.0	52.5
1° T 2006	Mujer	52.0	52.2
2° T 2006	Mujer	52.0	52.3
3° T 2006	Mujer	52.0	52.5
4° T 2006	Mujer	52.0	52.6

Estructura estimada según tramo de edad

Periodo	Tramo de edad	% de población con calibración	% de población sin calibración
1° T 2004	0-14	25.3	26.0
2° T 2004	0-14	25.3	25.5
3° T 2004	0-14	25.3	25.5
4° T 2004	0-14	25.3	25.7
1° T 2005	0-14	25.1	25.9
2° T 2005	0-14	25.1	25.3
3° T 2005	0-14	25.1	25.5
4° T 2005	0-14	25.1	25.6
1° T 2006	0-14	24.8	25.3
2° T 2006	0-14	24.8	25.0
3° T 2006	0-14	24.8	25.1
4° T 2006	0-14	24.8	25.2

Estructura estimada según tramo de edad

Periodo	Tramo de edad	% de población con calibración	% de población sin calibración
1° T 2004	15-29	25.8	26.2
2° T 2004	15-29	25.8	26.1
3° T 2004	15-29	25.8	25.9
4° T 2004	15-29	25.8	26.0
1° T 2005	15-29	25.8	26.0
2° T 2005	15-29	25.8	26.2
3° T 2005	15-29	25.8	25.6
4° T 2005	15-29	25.8	25.6
1° T 2006	15-29	25.6	26.2
2° T 2006	15-29	25.6	26.5
3° T 2006	15-29	25.6	26.5
4° T 2006	15-29	25.6	25.9
1° T 2004	30-49	25.1	24.1
2° T 2004	30-49	25.1	23.9
3° T 2004	30-49	25.1	24.3
4° T 2004	30-49	25.1	24.3
1° T 2005	30-49	25.3	23.8
2° T 2005	30-49	25.3	23.7
3° T 2005	30-49	25.3	24.5
4° T 2005	30-49	25.3	24.5
1° T 2006	30-49	25.5	24.2
2° T 2006	30-49	25.5	24.2
3° T 2006	30-49	25.5	24.3
4° T 2006	30-49	25.5	24.7
1° T 2004	50-64	13.5	13.3
2° T 2004	50-64	13.5	13.8
3° T 2004	50-64	13.5	13.8
4° T 2004	50-64	13.5	13.5
1° T 2005	50-64	13.6	13.8
2° T 2005	50-64	13.6	14.2
3° T 2005	50-64	13.6	13.9
4° T 2005	50-64	13.6	13.9
1° T 2006	50-64	13.7	14.0
2° T 2006	50-64	13.7	13.9
3° T 2006	50-64	13.7	13.7
4° T 2006	50-64	13.7	14.1

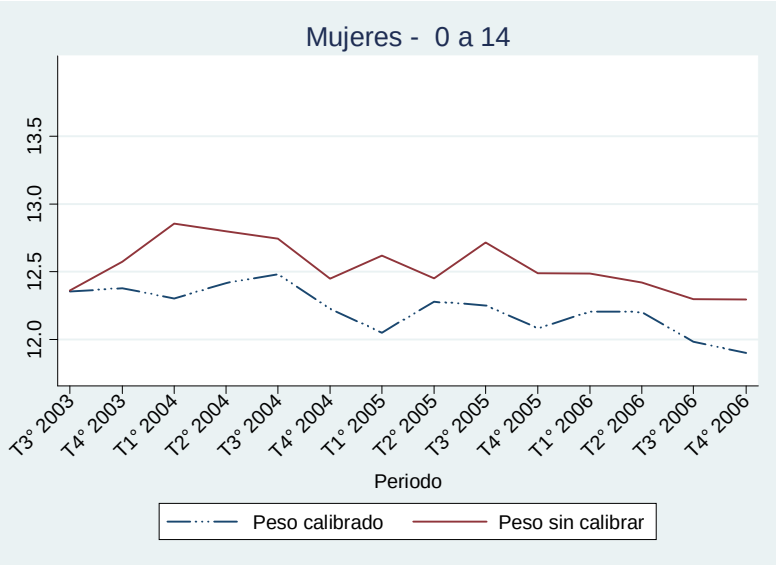
Estructura estimada según tramo de edad

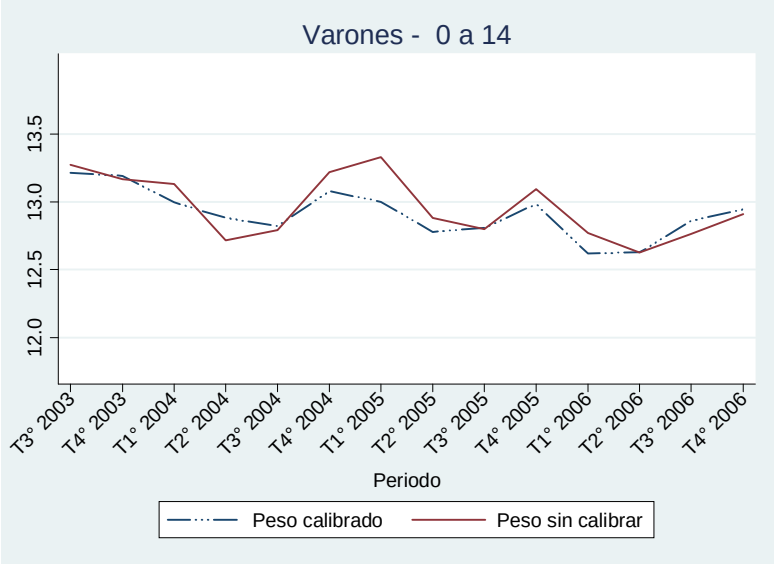
Periodo	Tramo de edad	% de población con calibración	% de población sin calibración
1° T 2004	65 y más	10.3	10.4
2° T 2004	65 y más	10.3	10.6
3° T 2004	65 y más	10.3	10.5
4° T 2004	65 y más	10.3	10.5
1° T 2005	65 y más	10.4	10.5
2° T 2005	65 y más	10.3	10.5
3° T 2005	65 y más	10.3	10.5
4° T 2005	65 y más	10.3	10.4
1° T 2006	65 y más	10.4	10.4
2° T 2006	65 y más	10.4	10.3
3° T 2006	65 y más	10.4	10.5
4° T 2006	65 y más	10.4	10.1

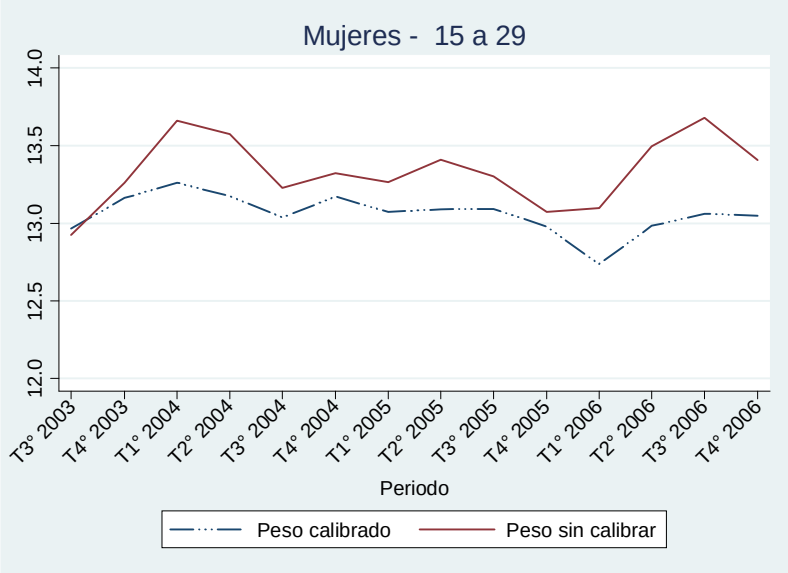
Luego de la calibración, la estructura de edad y de sexo prácticamente no presenta variaciones a lo largo del tiempo.

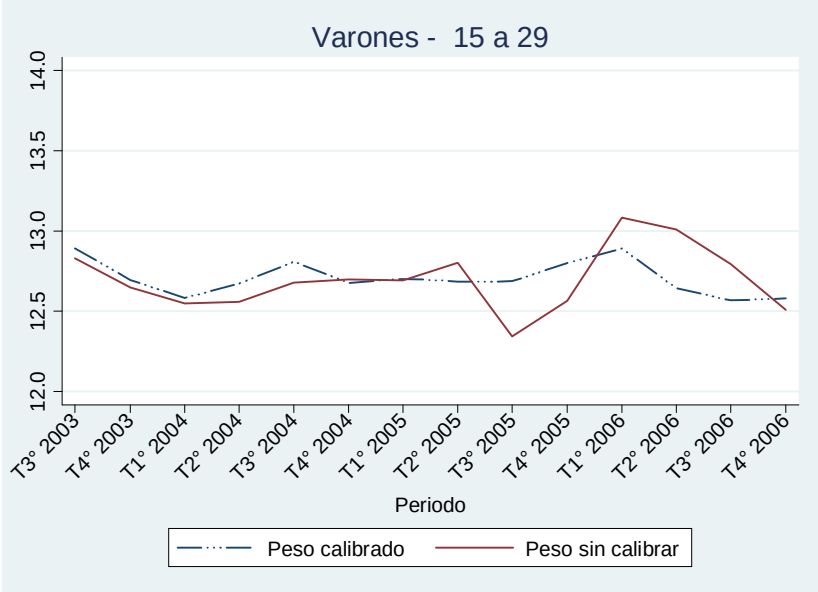
A continuación se procede a comparar la estructura de la población cruzando ambas variables. Se obtienen, entonces, 10 subpoblaciones sobre las que se grafica el porcentaje respecto de la población total para los trimestres estudiados.

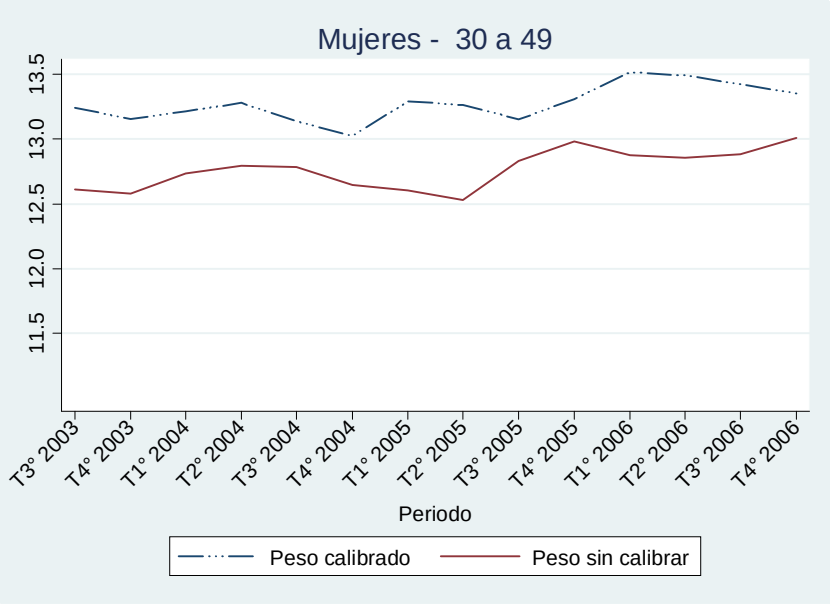
Estructura de la población sin calibración y con calibración
Período 2003-2006

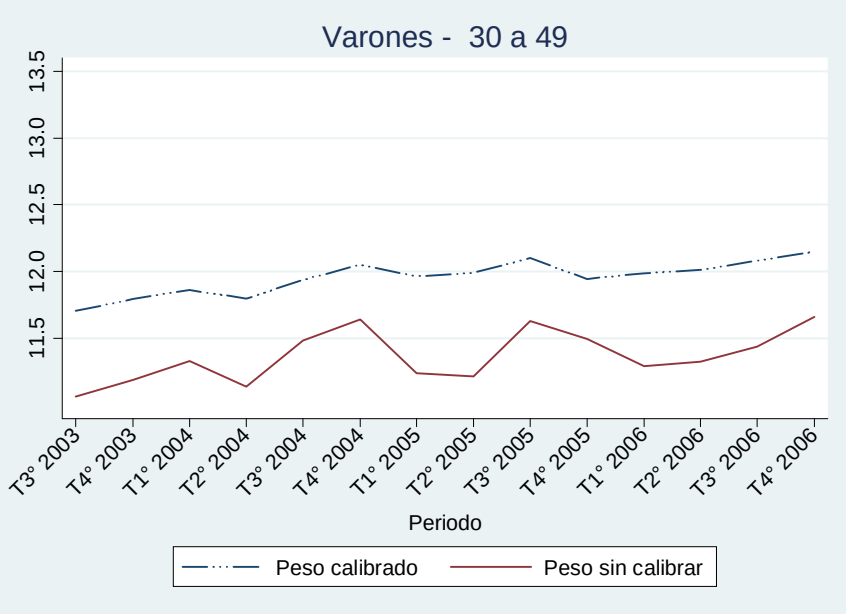


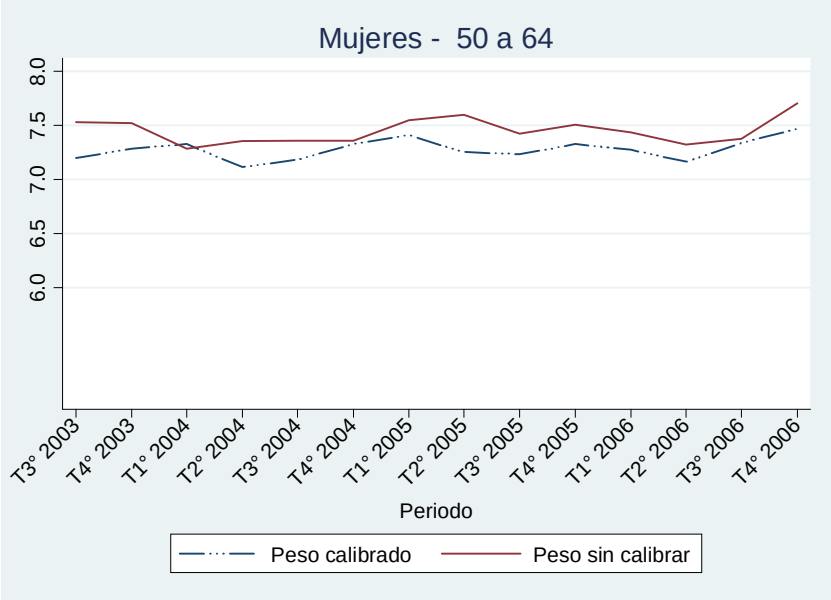


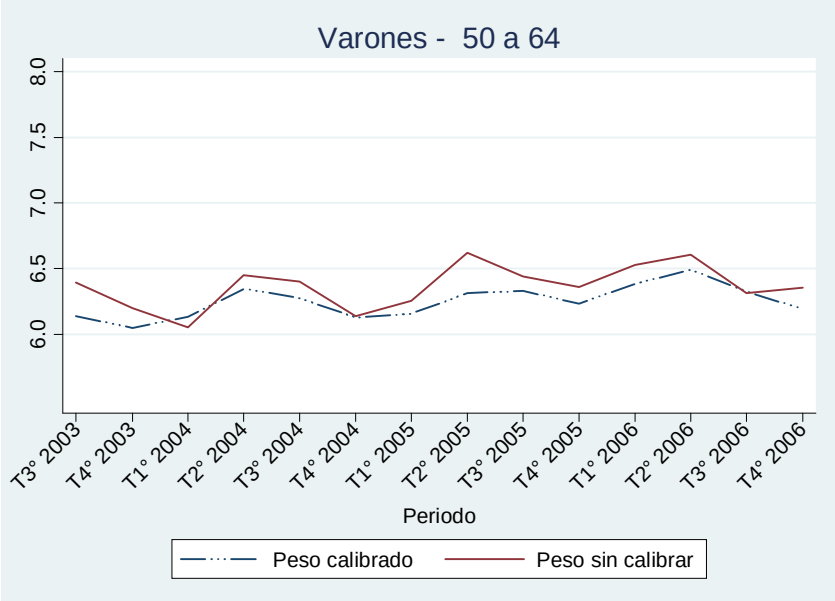

Mujeres - 15 a 29
14.0
13.5
13.0
12.5
12.0
T3° 2003
T4° 2003
T1° 2004
T2° 2004
T3° 2004
T4° 2004
T1° 2005
T2° 2005
T3° 2005
T4° 2005
T1° 2006
T2° 2006
T3° 2006
T4° 2006
Periodo
Peso calibrado
Peso sin calibrar

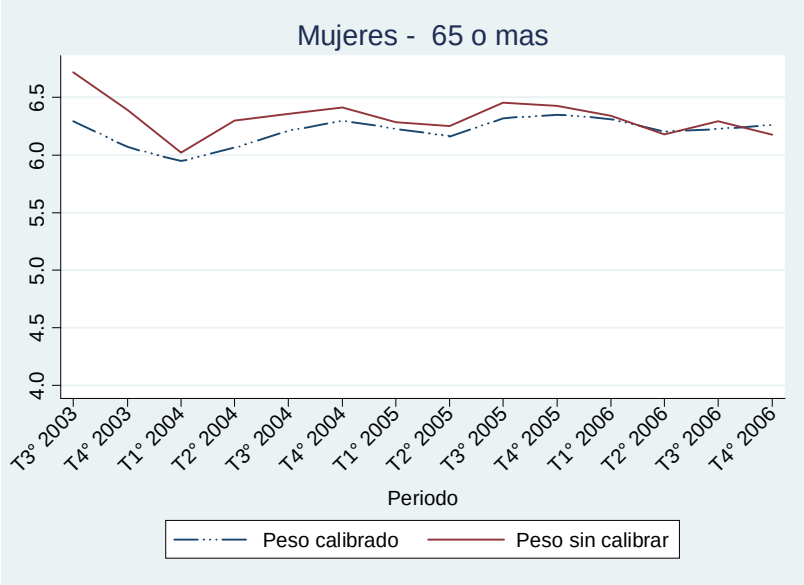

Varones - 15 a 29
14.0
13.5
13.0
12.5
12.0
T3° 2003
T4° 2003
T1° 2004
T2° 2004
T3° 2004
T4° 2004
T1° 2005
T2° 2005
T3° 2005
T4° 2005
T1° 2006
T2° 2006
T3° 2006
T4° 2006
Periodo
Peso calibrado
Peso sin calibrar


Mujeres - 30 a 49
11.5
12.0
12.5
13.0
13.5
T3° 2003
T4° 2003
T1° 2004
T2° 2004
T3° 2004
T4° 2004
T1° 2005
T2° 2005
T3° 2005
T4° 2005
T1° 2006
T2° 2006
T3° 2006
T4° 2006
Periodo
Peso calibrado
Peso sin calibrar


Varones - 30 a 49
11.5
12.0
12.5
13.0
13.5
T3° 2003
T4° 2003
T1° 2004
T2° 2004
T3° 2004
T4° 2004
T1° 2005
T2° 2005
T3° 2005
T4° 2005
T1° 2006
T2° 2006
T3° 2006
T4° 2006
Periodo
Peso calibrado
Peso sin calibrar


Mujeres - 50 a 64
8.0
7.5
7.0
6.5
6.0
T3° 2003
T4° 2003
T1° 2004
T2° 2004
T3° 2004
T4° 2004
T1° 2005
T2° 2005
T3° 2005
T4° 2005
T1° 2006
T2° 2006
T3° 2006
T4° 2006
Periodo
Peso calibrado
Peso sin calibrar


Varones - 50 a 64
8.0
7.5
7.0
6.5
6.0
T3° 2003
T4° 2003
T1° 2004
T2° 2004
T3° 2004
T4° 2004
T1° 2005
T2° 2005
T3° 2005
T4° 2005
T1° 2006
T2° 2006
T3° 2006
T4° 2006
Periodo
Peso calibrado
Peso sin calibrar


Mujeres - 65 o mas
4.0
4.5
5.0
5.5
6.0
6.5
T3° 2003
T4° 2003
T1° 2004
T2° 2004
T3° 2004
T4° 2004
T1° 2005
T2° 2005
T3° 2005
T4° 2005
T1° 2006
T2° 2006
T3° 2006
T4° 2006
Periodo
Peso calibrado
Peso sin calibrar


Varones - 65 o mas
4.0
4.5
5.0
5.5
6.0
6.5
T3° 2003
T4° 2003
T1° 2004
T2° 2004
T3° 2004
T4° 2004
T1° 2005
T2° 2005
T3° 2005
T4° 2005
T1° 2006
T2° 2006
T3° 2006
T4° 2006
Periodo
Peso calibrado
Peso sin calibrar

Se puede observar que las ponderaciones previas a la calibración sobrestiman a las poblaciones de mujeres y varones entre 0 y 14 años, mujeres entre 15 y 29 años y mujeres y varones de 65 o más, en forma sistemática aunque con diferentes niveles de sobrestimación.

Se trata de poblaciones con mayor probabilidad de hallarse en la vivienda al momento de la encuesta, y por ende con menor probabilidad de ser omitida en el proceso de relevamiento.

Simétricamente, el factor de expansión sin calibrar subestima sistemáticamente a las mujeres y varones entre 30 y 49 años, el tramo con mayor tasa de actividad de los grupos considerados.

II. IMPUTACIÓN DE LAS VARIABLES DE INGRESO

No respuesta en las encuestas a hogares

Toda encuesta (no sólo las encuestas probabilísticas) presenta valores faltantes. En el caso de una encuesta a hogares, la no respuesta puede deberse a:

- Rechazo de un hogar a responder la encuesta.
- Imposibilidad de contactar al hogar por ausencias reiteradas, aún luego de varias visitas.
- Valores incoherentes: aún cuando se implementen rigurosos controles informáticos, es imposible eliminar la presencia de valores incoherentes entre las respuestas. Esto da en general origen a valores considerados 'missing'.
- Imposibilidad de realizar el trabajo de campo, por ejemplo por problemas climáticos. Las viviendas que no fueron encuestadas entran en la categoría de no respuesta.
- El encuestado desconoce la respuesta: dado que en la EPH (como en la mayoría de las encuesta de su tipo) un entrevistado puede contestar por otro miembro del hogar, el respondente puede ignorar la respuesta. O si se trata de un autorespondente, puede no recordar la respuesta exacta, dando origen a una no respuesta.

Por lo tanto no siempre la 'No Respuesta' se debe a la falta de cooperación de los entrevistados con la encuesta.

En una encuesta a hogares como la EPH hay varios niveles de no respuesta:

- A nivel de vivienda
- A nivel de hogar
- A nivel de persona

Y en cada uno de estos niveles la no respuesta puede ser total o parcial:

- No respuesta total: ausencia total de información en la unidad de muestreo (o de análisis en la práctica).
- No respuesta parcial: ausencia de información en algunas variables relevadas.

Dependerá de cada encuesta y los objetivos analíticos la cantidad y cualidad de variables cuya presencia/ausencia hará que la no respuesta sea parcial o total.

No respuesta en las variables de ingreso

En una encuesta a hogares las variables de ingreso son, en general, las que presentan mayores dificultades para la captación y complejidad para la corrección de la no respuesta. Esto se debe a varios motivos:

- **Asimetría de la variable con presencia de valores extremos:** las variables de ingreso son en todos los países variables asimétricas, caracterizadas por presentar muchos valores concentrados en la cola izquierda (valores bajos) de la distribución y pocos valores extremos a la derecha.

- **Dificultad de determinar el carácter de los valores extremos y su tratamiento:** las variables de ingreso son utilizadas para construir diferente indicadores: medias o medianas de ingreso, medidas de desigualdad, e.g coeficiente de Gini, ratios, estimaciones de pobreza, etc.. Para cada uno de estos indicadores, hay varias formas de tratar los valores extremos:

  - Asignarles ponderación igual a 1, suponiendo que son valores totalmente atípicos, que no 'representan' a otros individuos.
  - Modificar el valor de la variable ('trimming'), asignándole el primer valor inferior considerado no 'extremo'.
  - Eliminarlo del análisis, etc.

- **Temática sensible:** el ingreso de los hogares y las personas es una de las dimensiones de la EPH más utilizadas por los usuarios para calcular medidas de desigualdad, pobreza, índices de indexación, etc. De ahí la importancia de un tratamiento cuidadoso de la no respuesta.

No respuesta de ingreso en la EPH

Se presentan algunos cuadros que permiten dimensionar la no respuesta de ingreso en la EPH. Se tabularon estos porcentajes sin ponderar.

No respuesta de ingreso en el ingreso de la ocupación principal (sin ponderar)

Periodo	% NR Ingreso de la ocupación principal
1° T 2004	15.3
2° T 2004	14.9
3° T 2004	13.6
4° T 2004	12.8
1° T 2005	13.2
2° T 2005	11.0
3° T 2005	11.6
4° T 2005	10.4
1° T 2006	10.4
2° T 2006	10.0
3° T 2006	9.6
4° T 2006	10.1

No respuesta de ingreso en la ocupación principal según categoría ocupacional (sin ponderar)

Período	Patrón	Cta Propia	Asalariado
1° T 2004	37.9	21.4	10.8
2° T 2004	34.8	20.2	10.7
3° T 2004	35.2	19.6	9.4
4° T 2004	32.8	17.8	9.3
1° T 2005	31.4	17.1	9.9
2° T 2005	32.1	14.0	8.1
3° T 2005	27.0	15.4	8.7
4° T 2005	30.0	13.4	7.8
1° T 2006	27.2	13.3	7.9
2° T 2006	24.9	12.9	7.6
3° T 2006	25.6	12.8	7.3
4° T 2006	24.7	13.2	7.9

No respuesta de ingreso en el total familiar (sin ponderar)

Período	% NR ingreso total familiar
1° T 2004	21.9
2° T 2004	20.4
3° T 2004	19.6
4° T 2004	18.0
1° T 2005	18.3
2° T 2005	15.4
3° T 2005	16.8
4° T 2005	15.1
1° T 2006	14.6
2° T 2006	14.1
3° T 2006	13.9
4° T 2006	14.3

No respuesta al ingreso total familiar según nivel de educación del jefe de hogar (sin ponderar)

Periodo	Sin inst /Prim inc	Prim comp /Sec inc	Sec comp /Univ inc	Univ comp
1° T 2004	16.9	19.3	24.6	30.7
2° T 2004	17.5	17.9	22.4	27.4
3° T 2004	13.5	18.0	21.3	27.4
4° T 2004	13.7	16.3	18.6	26.6
1° T 2005	13.0	16.0	19.4	28.7
2° T 2005	11.9	13.4	16.3	23.2
3° T 2005	13.0	15.0	17.8	24.1
4° T 2005	11.6	12.7	16.4	22.9
1° T 2006	11.0	12.4	15.5	23.3
2° T 2006	10.2	11.5	15.2	23.6
3° T 2006	9.8	12.1	15.3	21.0
4° T 2006	11.2	12.4	15.6	20.5

Hasta el año 2003, en la EPH no se efectuaban correcciones específicas de las variables de ingreso. Para la elaboración de tabulados según las escalas decílicas de la población, ya fuera según el ingreso total individual, el ingreso de la ocupación principal o el ingreso total familiar (en hogares) se trabajaba con los respondentes, eliminando del cálculo los no respondentes, y conservando el factor de expansión original (con los correspondientes ajustes por no respuesta total). A partir de 2003, con la entrada en vigencia de la modalidad continua, se incorporó provisoriamente una metodología de corrección, reponderando las variables de ingreso.

Se creaba para cada variable de ingreso -P21 (ingreso de la ocupación principal), P47t (ingreso total individual) e ITF (ingreso total familiar)- un ponderador ad-hoc que se aplicaba a la sub-base de respondentes, y con valor cero para los no respondentes: Pondiio, Pondii, Pondih.
Para analizar estas tres variables había que trabajar, de hecho, con tres sub-bases. Esto traía aparejado problemas de coherencia en las estimaciones, pues surgían de tres bases diferentes, y severas limitaciones para cualquier análisis multivariado ya que el usuario debía trabajar con cuatro factores de expansión distintos.

Otra consecuencia del tratamiento de reponderación para la corrección de la no respuesta parcial es la 'amplificación' del porcentaje de no respuesta al pasar de los individuos a los hogares, ya que la falta de información parcial de algún individuo determina como no respondente al hogar al cual pertenece.

Para ejemplificar esto, se adaptó un ejemplo tomado del trabajo de Fernando Medina y Marco Galván[5] .

Se supone que a partir de una base de personas se desea construir una base de hogares, sumarizando el ingreso de los individuos para obtener el ingreso familiar.

Se parte de la base de individuos

Hogar	Individuo	Recibe ingreso	Declara su ingreso?
A	1	Si	Si
A	2	Si	Si
A	3	Si	Si
B	4	Si	No
B	5	Si	Si
B	6	No	-
C	7	Si	Si
C	8	Si	Si
C	9	Si	No

[5] Medina, F. , Galván, M. (2007). *Imputación de datos: teoría y práctica*. CEPAL. Serie Estudios Estadísticos y Prospectivos.

Porcentaje de no respuesta en el ingreso individual: 100*2/8 = 2.5%

Se obtiene la base de hogares

Hogar	Ingreso total familiar
A	I1 + I2 + I3
B	?
C	?

Porcentaje de no respuesta en el ingreso total familiar: 100*2/3=66.7%

Para corregir esto, se implementa en la EPH a partir del tercer trimestre de 2009 el método de imputación, también ampliamente utilizado en los institutos de estadística.

La metodología de imputación en el tratamiento de valores faltantes

La imputación consiste en reemplazar un valor faltante ('missing') por un valor válido. Según menciona E. Rancourt[6] la primera utilización de la imputación en las estadísticas oficiales se debe a Hansen, Hurvitz y Madow, en 1948.

La mayoría de estos métodos define en primer término las celdas de imputación, para luego dentro de ellas imputar según diversas técnicas:

- Imputación por la media
- Imputación por regresión
- Imputación mediante hot-deck, etc.

Asimismo cada uno de estos métodos puede presentar diversas variantes (aleatorio, no aleatorio, etc.)

Imputación mediante hot-deck

El método de hot-deck es muy sencillo:

> Consiste en seleccionar, dentro de la clase de imputación correspondiente, un *donante* al azar, entre los que presentan valores válidos, y asignar este valor válido a la celda con el valor faltante.

Algunas de las variantes más usuales del hot- deck son: aleatorio, jerárquico, métrico, secuencial. La selección del donante puede ser también realizada, dentro de cada celda de imputación, mediante probabilidades proporcionales al factor de expansión.
El método se hizo rápidamente popular principalmente en la depuración de grandes volúmenes de datos (censos), con la variante 'secuencial', que consiste en imputar el valor más cercano en el orden del archivo dentro de cierta clase de imputación.

---

[6] Rancourt, E., "Edit and imputation: from suspicious to scientific techniques", Actes, l'Association internationale des statisticiens d'enquête, pp 605-633, (2001)

La imputación por hot-deck aleatorio conduce a estimadores insesgados (de totales) si:

- La probabilidad de respuesta en cada celda de imputación es uniforme.
- Los donantes son seleccionados con probabilidad proporcional a su factor de expansión.

Un resumen de la teoría subyacente en la imputación puede verse en Haziza (2002)[7]. Para un tratamiento detallado de la no respuesta en las encuestas por muestreo, un panorama de los métodos modernos de imputación y simulaciones comparando la eficiencia de los mismos mediante una programación en Stata, recomendamos al lector el trabajo de Fernando Medina y Marco Galván: "*Imputación de datos: teoría y práctica*" CEPAL. Serie Estudios Estadísticos y Prospectivos.

Respecto a la imputación por la media, el hot-deck presenta estas ventajas:

- Reproducir la variabilidad que hay en la población: la imputación por la media imputa el mismo valor a todos los casos faltantes de la celda de imputación. Esto es una ventaja en la estimación de medidas de desigualdad.
- Imputar un valor observado: El valor imputado es un valor efectivamente observado, a diferencia de la imputación por la media o por regresión.

Una desventaja del hot-deck aleatorio es que produce un aumento de la varianza, al introducir un nuevo mecanismo aleatorio: qué valor se imputa. Para reducir esta nueva fuente de variación se recomienda la construcción de 'celdas de imputación' con valores lo más homogéneos posibles.

Imputación por hot-deck e imputación por regresión

En las encuestas donde hay variables 'explicativas' continuas (encuestas a empresas, o encuestas de gastos), una alternativa usual es la imputación mediante un modelo de regresión. En el caso del ingreso en una encuesta a hogares, las variables auxiliares disponibles no permiten ajustar un modelo con suficiente fuerza explicativa, menos aún para las categorías ocupacionales de *Cuenta Propia* y *Patrones*.

Una ventaja de la imputación por regresión es su capacidad de imputar aún en celdas con pocos valores efectivos siempre que el modelo sea válido y pueda ser ajustado correctamente, mientras que en el hot-deck puede suceder que no haya suficientes donantes para una celda con pocos valores efectivamente observados. Por otro lado, el hot-deck tiene la ventaja de poder 'predecir' mejor en las celdas con suficientes valores efectivos.

---

[7] Haziza, David (2002). Inference en presence d'imputation: un survol. Actes des JMS 2002. INSEE.

Para un detallado análisis de diversas alternativas de imputación remitimos al lector al minucioso trabajo de Medina-Galván, donde además de las consideraciones teóricas hay una serie de simulaciones realizadas con bases de microdatos de la EPH de Argentina. En el mismo se demuestra que no hay un método que sea superior a los demás, aún considerando una sola encuesta. Pero como dice un adagio célebre en el campo de la imputación, el mejor método es el que no se aplica. La mejor corrección a la no respuesta es evitarla. Con esto quiere decirse que ningún método estadístico, por sofisticado que sea puede suplir la información faltante.

Tratamiento de la no respuesta de ingreso en la EPH a partir del tercer trimestre de 2009

En la EPH se implementa a partir del tercer trimestre de 2009 la imputación por hot-deck aleatorio, cuyas características son:

- Se imputan cada una de las componentes elementales de las preguntas de ingreso. Luego, se sumariza para obtener P21, P47t e ITF (ingreso de la ocupación principal, ingreso total individual e ingreso total familiar).

- Para cada variable a imputar, se determinan un conjunto de 'variables explicativas', que definirán las 'celdas de imputación' (sexo, tramo de edad, calificación, horas trabajadas, nivel educativo, condición de actividad, aglomerado, etc.).

- Se genera un orden aleatorio en la base previo a la imputación.

- Un valor faltante es reemplazado por un valor válido del registro de una persona perteneciente a la misma 'celda de imputación'. Este registro seleccionado al azar se denomina 'donante'.

- La selección de los donantes se realiza sin reposición. Un 'donante' no es utilizado dos veces.

- El mecanismo es iterativo: Si en cierta celda quedan valores sin imputar, se retira una de las variables explicativas y se repite el procedimiento.

- Identificación de casos con variables imputadas: Se creó una variable que indica para cada caso, la presencia o ausencia de imputación en cada tipo de variable elemental de P21 o P47T.

- Valores extremos: Las variables de ingreso tienen en general una distribución asimétrica, y pueden presentar valores 'extremos', i.e. valores muy elevados que se supone no

representan a otras unidades. O que al aparecer con baja frecuencia en el tiempo, pueden dificultar el análisis de la evolución de las medidas de desigualdad. Imputar estos valores (seleccionarlos como donantes) introduciría un aumento de la varianza. Para evitar esto, se programó un mecanismo para no considerar como donantes a los valores extremos. Esto es algo usual en la práctica de la imputación cuando se trabaja con variables muy asimétricas ('trade-off' entre sesgo y varianza).

- Reproducibilidad: Uno de los motivos para adoptar el método hot-deck es su sencillez de implementación, por lo que cualquier usuario, si lo desea, puede programarlo.

Asimetría de las variables de ingreso. Ingreso de la ocupación principal

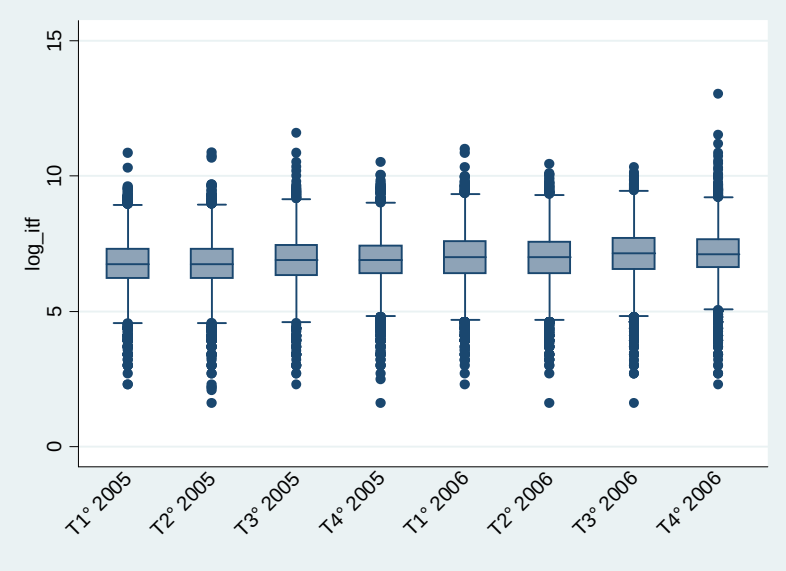
El ingreso es una variable muy asimétrica a la derecha. El siguiente cuadro muestra mediana, media, percentil 99 y valor máximo del ingreso de la ocupación principal efectivamente declarado (estimación ponderada):

Periodo	Mediana	Media	Percentil 99	Máximo	Desvío Standard
1° T 2004	400	562	3,100	54,000	1,071
2° T 2004	400	573	3,000	18,000	717
3° T 2004	400	578	3,000	25,000	717
4° T 2004	450	597	3,000	15,000	691
1° T 2005	500	641	3,800	30,000	764
2° T 2005	500	678	4,000	15,000	785
3° T 2005	576	741	3,800	100,000	1,511
4° T 2005	600	758	4,000	22,000	834
1° T 2006	600	829	4,000	60,000	1,188
2° T 2006	700	848	4,200	50,000	962
3° T 2006	730	900	4,200	30,000	907
4° T 2006	800	958	5,000	300,000	2,223

La asimetría se puede ver gráficamente tomando logaritmo y graficando un Box-Plot. Aún mediante esta transformación, la variable transformada presenta ‘valores extremos’. Se observa además cómo esta transformación ‘simetriza’ a estas variables.

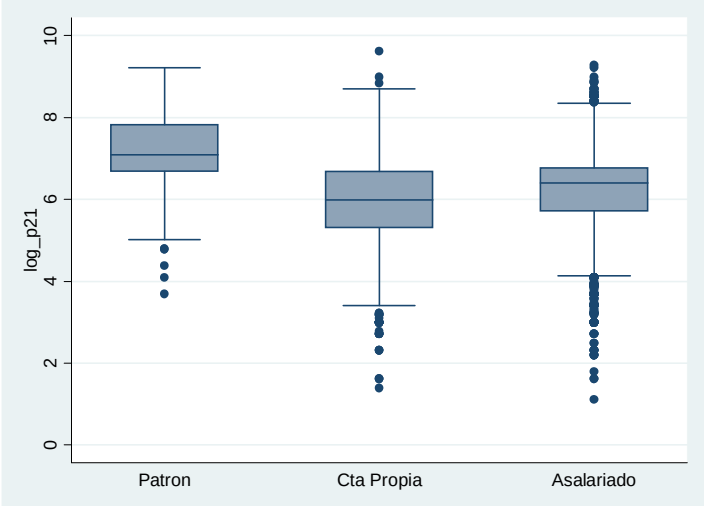
Se ejemplifica con los trimestres de los años 2005-2006, para las variables log (Ingreso total Familiar) y log (Ingreso de la ocupación Principal).

Logaritmo del ingreso total familiar

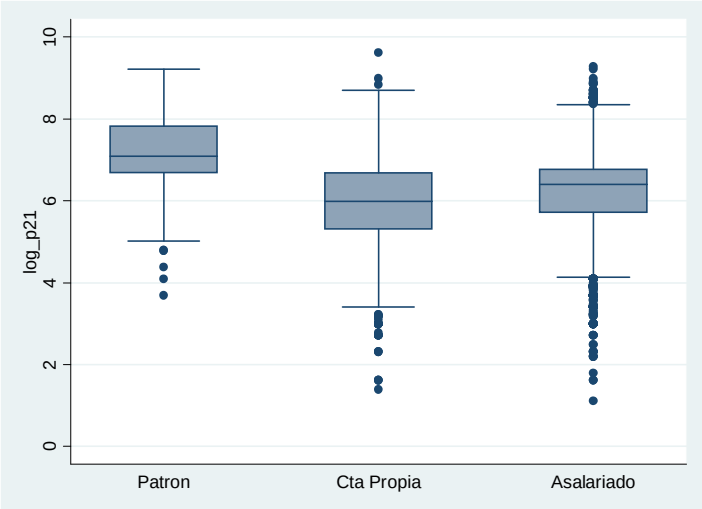


Logaritmo del ingreso de la ocupación principal

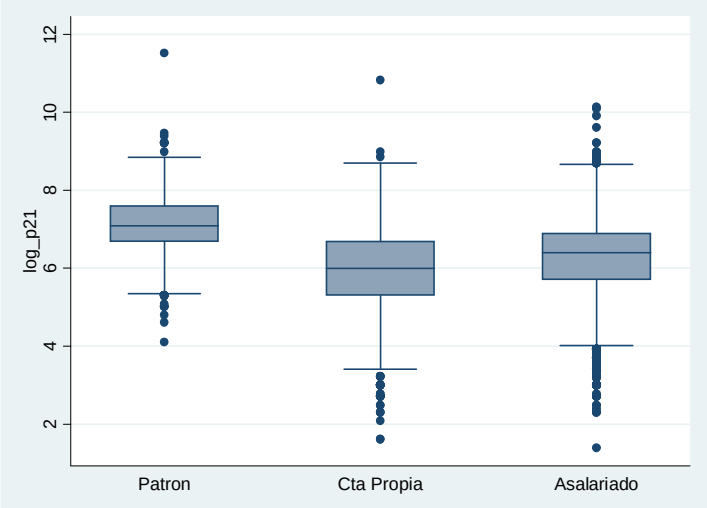
1º Trimestre 2005



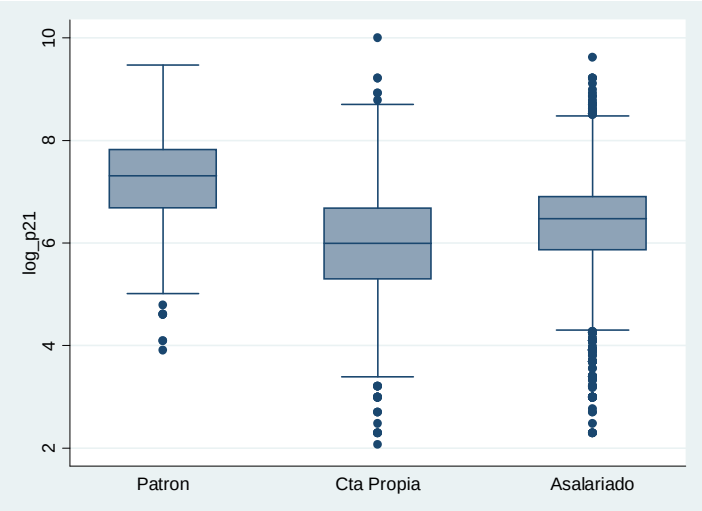
2º Trimestre 2005



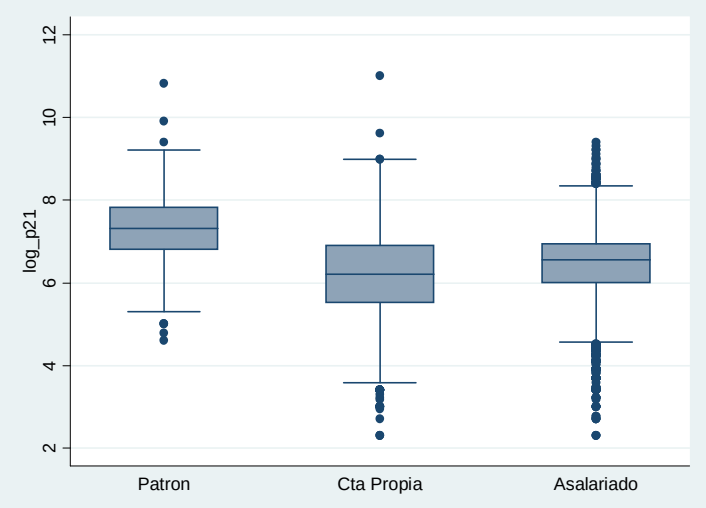
3º Trimestre 2005



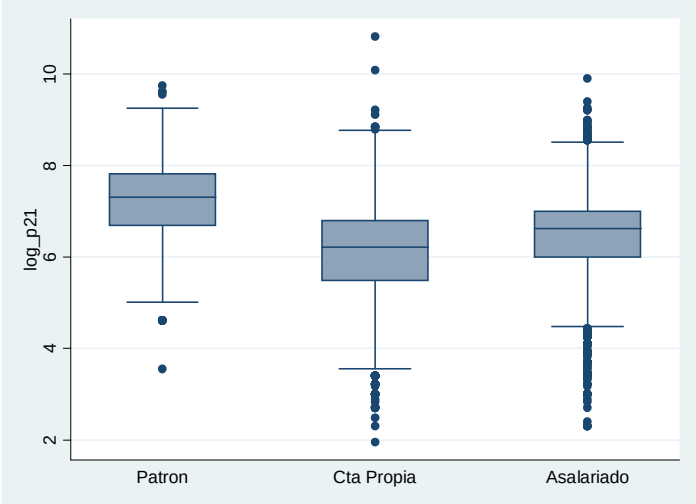
4º Trimestre 2005



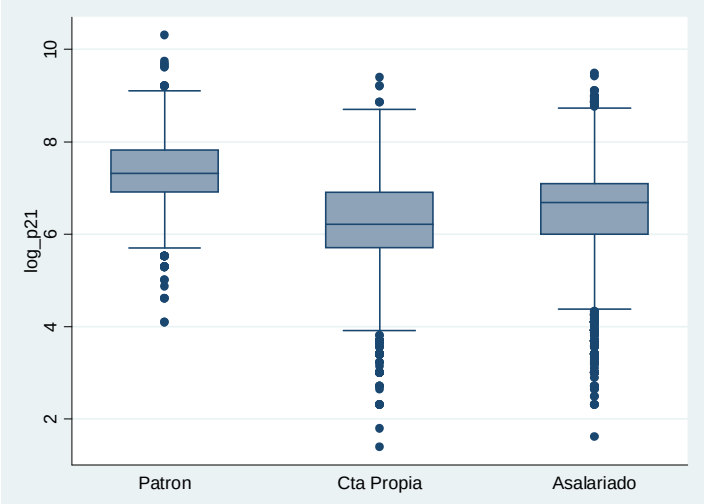
1º Trimestre 2006



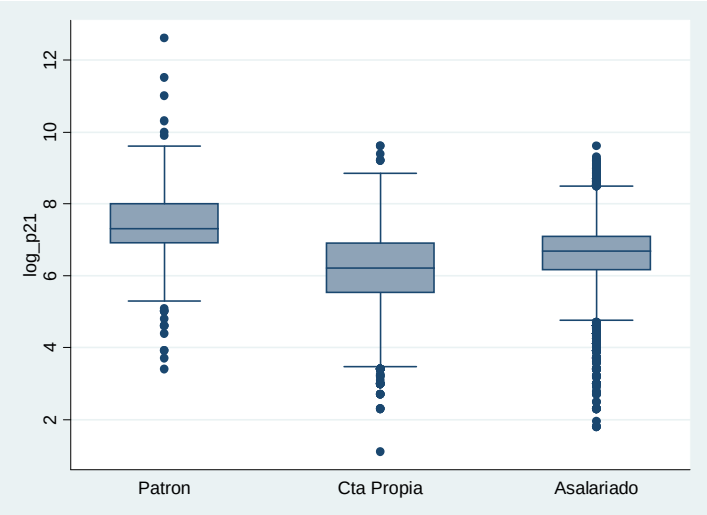
2º Trimestre 2006



3º Trimestre 2006



4º Trimestre 2006



Valores declarados y valores imputados

La no respuesta no se distribuye aleatoriamente entre los ocupados (o perceptores); esto se verifica al producir la imputación un leve aumento en la media del ingreso a lo largo de los trimestres y comprobar que la no respuesta de ingreso es más alta en las categorías con mayor ingreso.

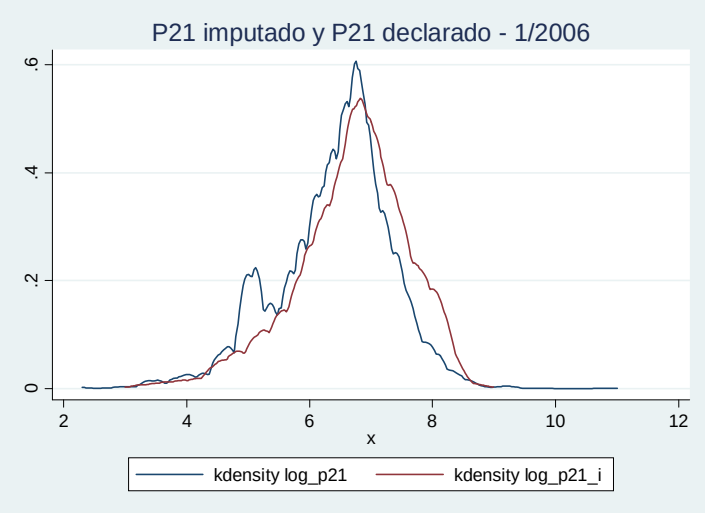
Los gráficos que se presentan a continuación muestran las diferencias entre:

- Distribución de Valores declarados
- Distribución de Valores imputados
- Distribución de Variable completa (valores declarados+ valores imputados)

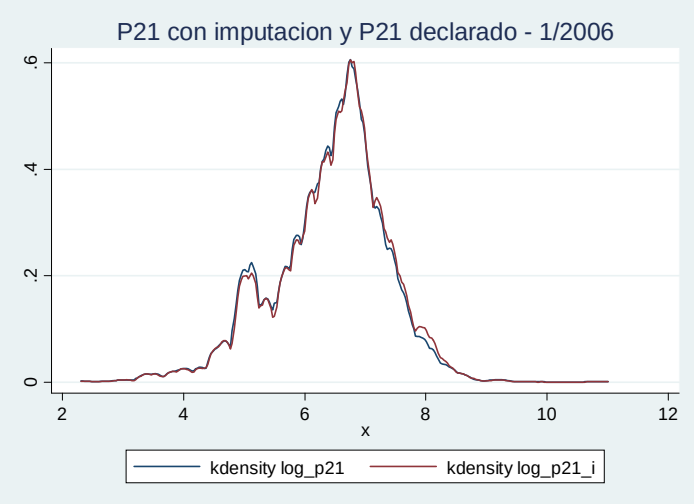
Para visualizar la diferencia entre la distribución del ingreso de la ocupación principal de los declarantes y de sólo los imputados, y entre los ocupados que declaran el ingreso de la ocupación principal y el conjunto de ocupados, se graficó la densidad de ambas variables mediante el soft Stata (utilizando el *kernel* de Epanechnikov). Se observa que la densidad del conjunto de valores imputados (que eran faltantes) está corrida a la derecha, pero que el conjunto de valores, incluyendo los imputados, no se diferencia del conjunto de valores antes de la imputación. Esto se debe a la baja tasa de no respuesta a nivel general.

Densidad del logaritmo del ingreso de la ocupación principal (log_p21)

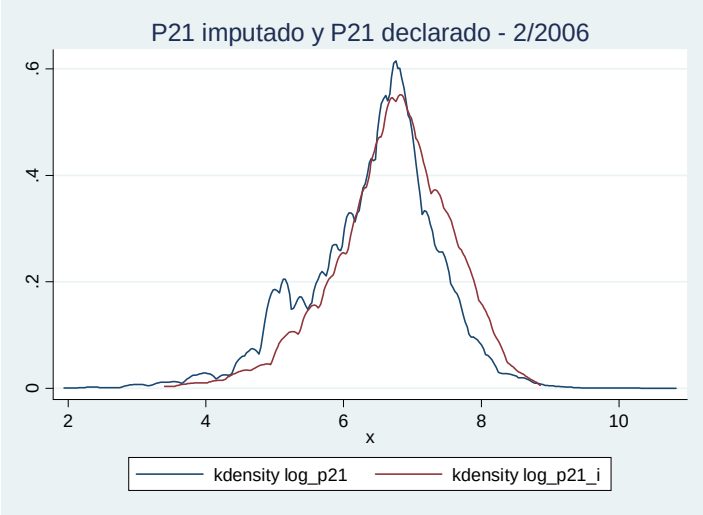
Valores declarados y valores imputados



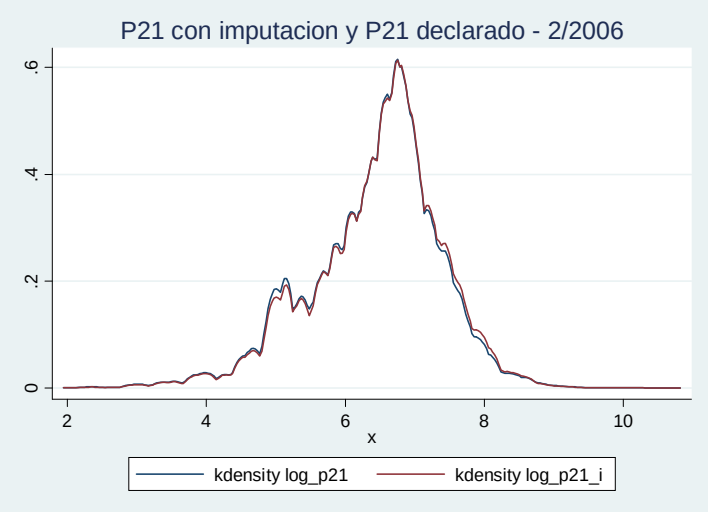
Valores declarados y valores declarados más imputados



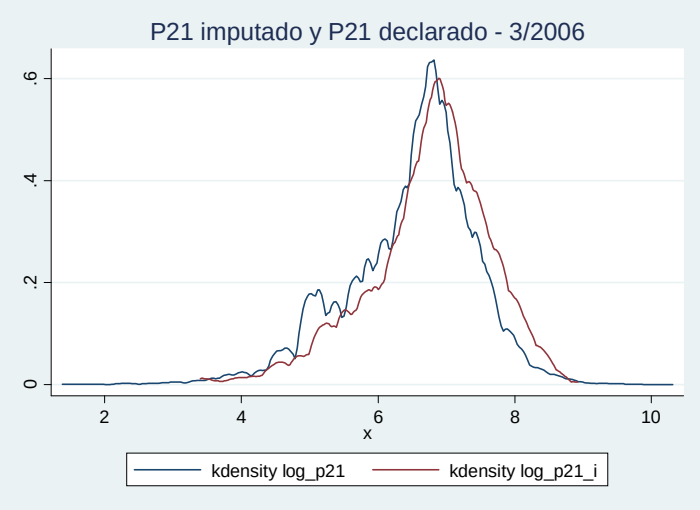
Valores declarados y valores imputados



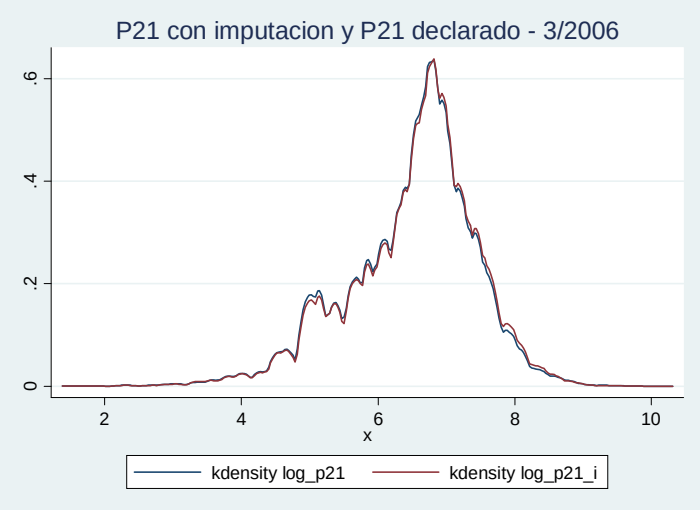
Valores declarados y valores declarados más imputados



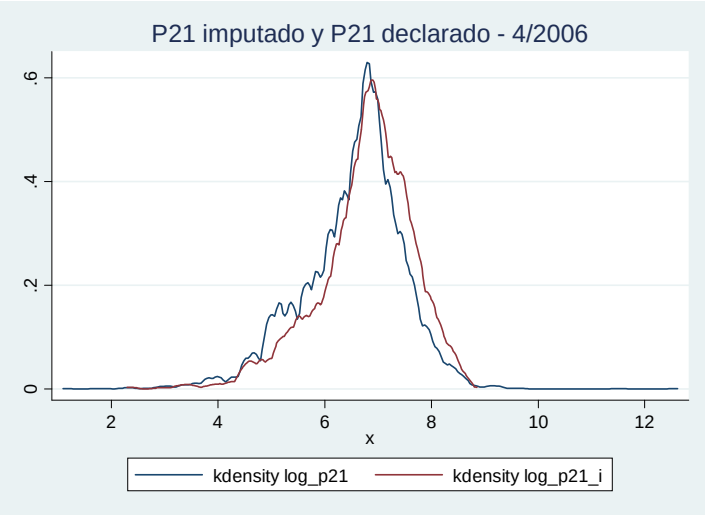
Valores declarados y valores imputados



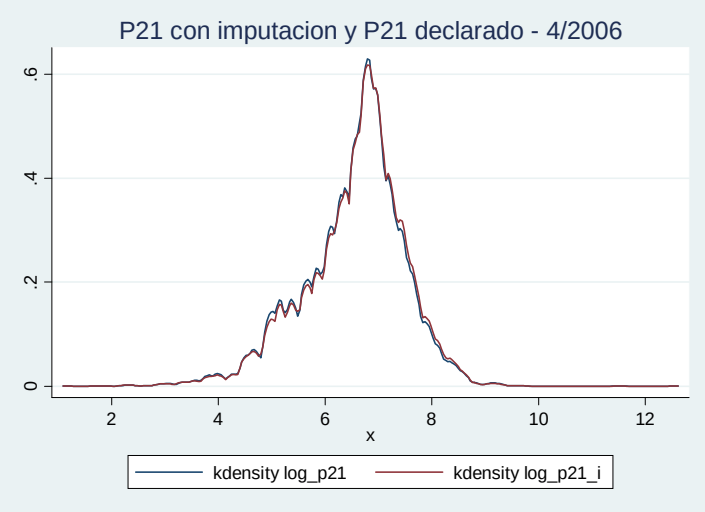
Valores declarados y valores declarados más imputados



Valores declarados y valores imputados

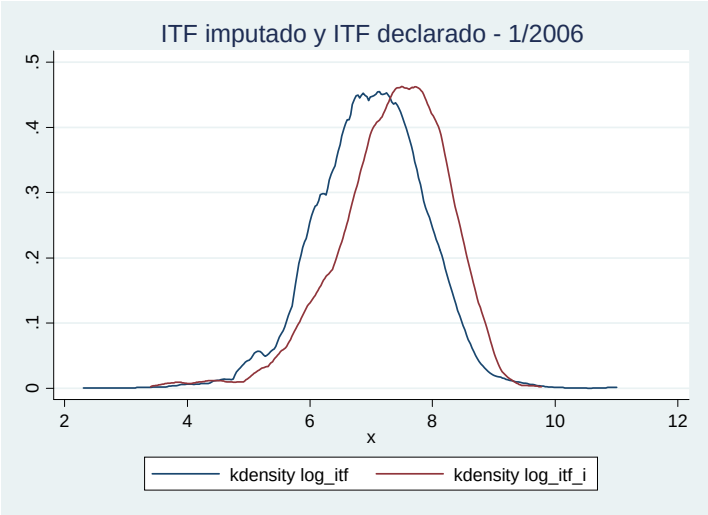


Valores declarados y valores declarados más imputados

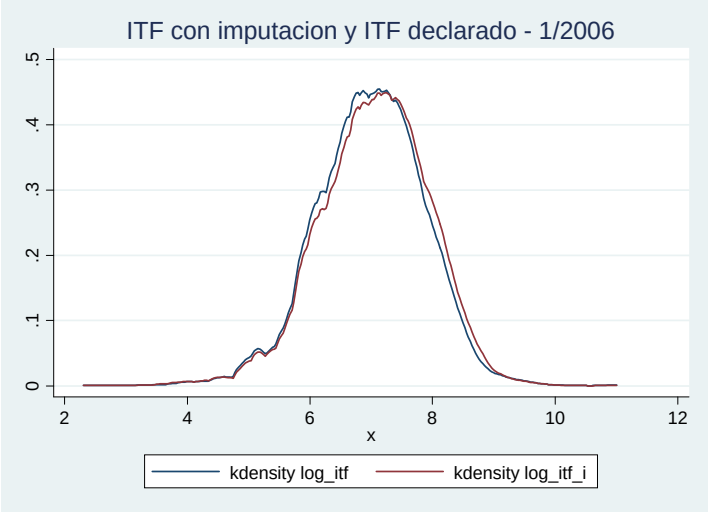


Densidad del logaritmo del ingreso total familiar (log_itf)

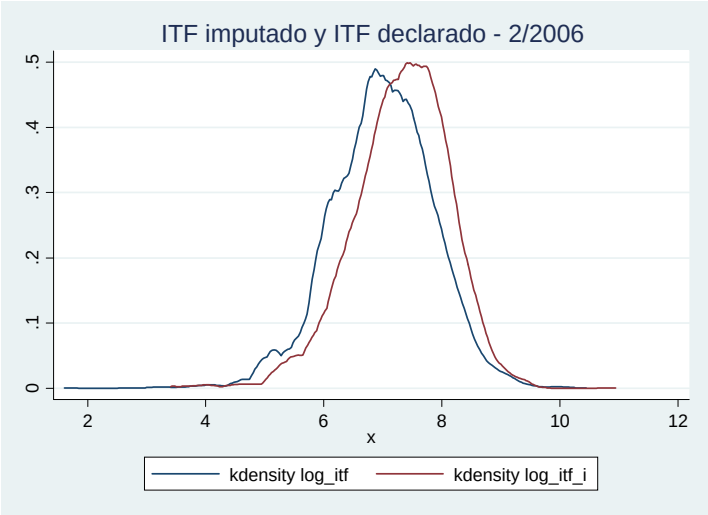
Valores declarados y valores imputados



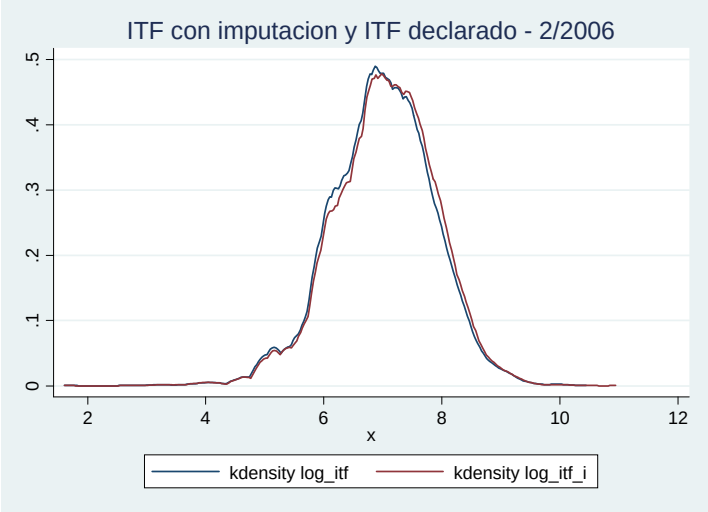
Valores declarados y valores declarados más imputados



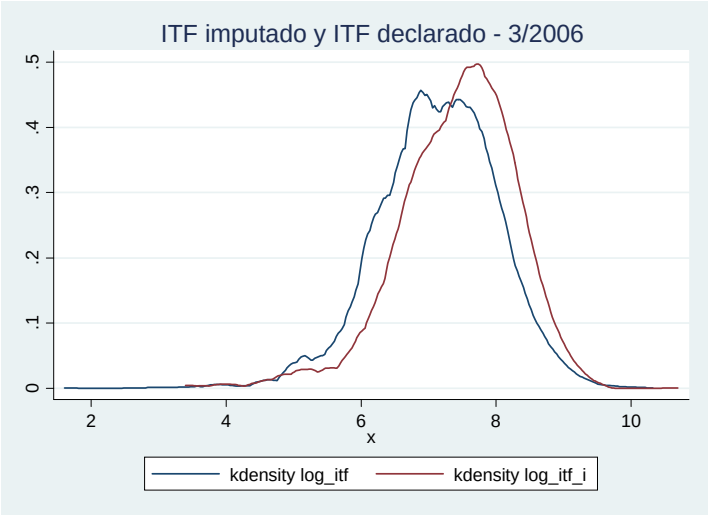
Valores declarados y valores imputados



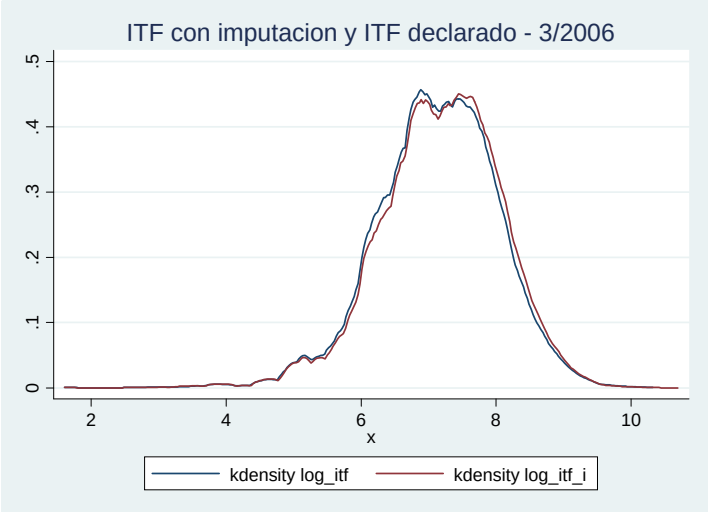
Valores declarados y valores declarados más imputados



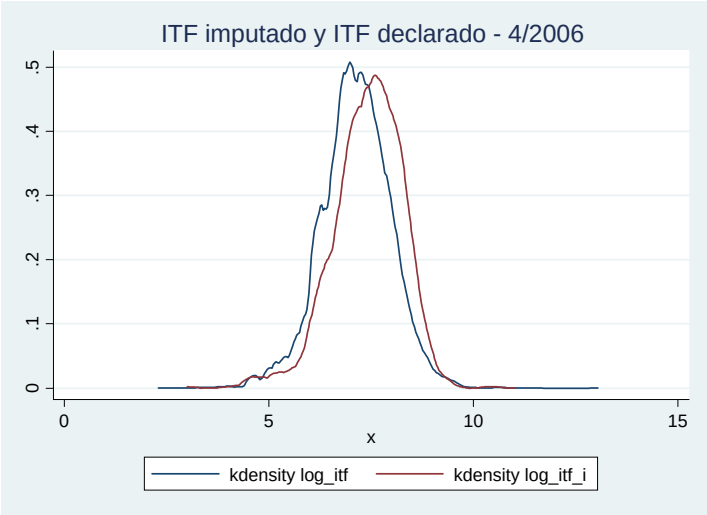
Valores declarados y valores imputados



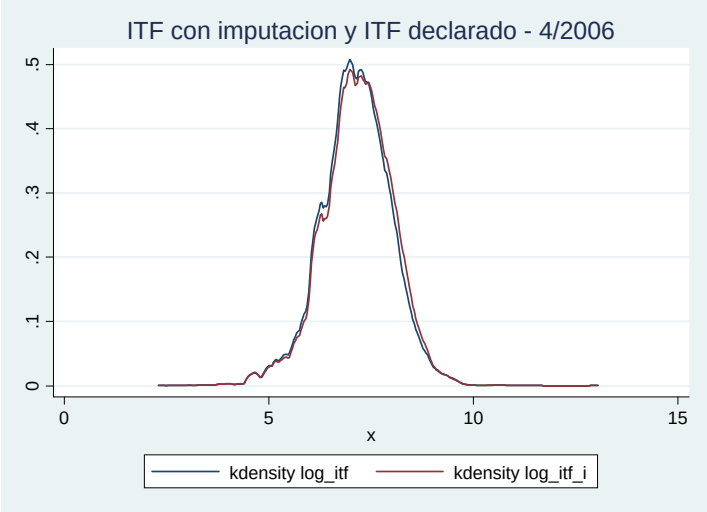
Valores declarados y valores declarados más imputados



Valores declarados y valores imputados



Valores declarados y valores declarados más imputados



El efecto de la imputación en las escalas de la distribución del ingreso, tanto para el ingreso de la ocupación principal como para el ingreso total familiar, es limitado dado la baja tasa de no respuesta a nivel general.
En los cuadros siguientes se presentan los percentiles 1, 5, 10, 25, 50, 75, 90 ,95 y 99 del ingreso de la ocupación principal para el universo de ocupados, y del ingreso total familiar para el universo de hogares, antes de la imputación con la ponderación general sin corrección por no respuesta parcial, y con la ponderación general luego de la imputación:

Ingreso de la ocupación principal sin imputación

Percentiles calculados sobre el universo muestral (ponderado con el factor de expansión general) de los ocupados que responden el ingreso de la ocupación principal.

Ingreso de la ocupación principal con imputación

Percentiles calculados sobre el universo muestral (ponderado con el factor de expansión general) de los ocupados, con ingreso de la ocupación imputado en caso de no respuesta.

Ingreso total familiar sin imputación

Percentiles calculados sobre el universo muestral (ponderado con el factor de expansión general) de hogares sin componentes con alguna no respuesta en alguna variable de ingreso.

Ingreso total familiar con imputación

Percentiles calculados sobre el universo muestral (ponderado con el factor de expansión general) de hogares, con ingreso total familiar como suma de ingresos de sus componentes, estos últimos imputados en caso de haber sido no respuesta.

Ingreso de la ocupación principal
Estadísticos antes y después de la imputación

Periodo	tipo	p1	p5	p10	p25	p50	p75	p90	p95	p99
1° T 2004	SI	0	0	8	150	400	700	1,200	1,653	3,200
1° T 2004	CI	0	0	50	160	400	732	1,200	1,800	3,200
2° T 2004	SI	0	0	60	170	400	720	1,200	1,600	3,000
2° T 2004	CI	0	0	80	200	450	800	1,200	1,702	3,000
3° T 2004	SI	0	0	60	160	400	750	1,200	1,600	3,000
3° T 2004	CI	0	0	80	200	450	800	1,250	1,800	3,000
4° T 2004	SI	0	0	60	180	450	800	1,200	1,800	3,050
4° T 2004	CI	0	0	80	200	495	800	1,300	1,850	3,000
1° T 2005	SI	0	0	80	200	500	800	1,300	1,800	3,800
1° T 2005	CI	0	0	100	240	500	800	1,450	2,000	3,500
2° T 2005	SI	0	0	100	200	500	850	1,400	2,000	4,000
2° T 2005	CI	0	0	100	250	500	900	1,500	2,000	3,850
3° T 2005	SI	0	0	100	250	580	922	1,500	2,000	3,800
3° T 2005	CI	0	10	100	270	600	1,000	1,500	2,000	4,000
4° T 2005	SI	0	0	100	250	600	1,000	1,500	2,000	4,000
4° T 2005	CI	0	0	120	300	600	1,000	1,600	2,100	4,000
1° T 2006	SI	0	0	100	300	625	1,000	1,700	2,200	4,000
1° T 2006	CI	0	0	120	300	700	1,100	1,800	2,500	4,000
2° T 2006	SI	0	0	120	300	700	1,000	1,800	2,300	4,200
2° T 2006	CI	0	30	150	300	700	1,100	1,800	2,500	4,250
3° T 2006	SI	0	20	145	300	750	1,200	1,880	2,500	4,200
3° T 2006	CI	0	48	150	350	800	1,200	2,000	2,500	4,500
4° T 2006	SI	0	0	120	340	800	1,200	2,000	2,500	5,000
4° T 2006	CI	0	10	150	380	800	1,200	2,000	2,700	5,000

SI: Sin imputación
CI: Con imputación

Ingreso total familiar
Estadísticos antes y después de imputación

Periodo	Tipo	p1	p5	p10	p25	p50	p75	p90	p95	p99
1° T 2004	SI	0	150	210	400	700	1,250	2,100	2,930	5,700
1° T 2004	CI	0	150	230	400	770	1,361	2,300	3,150	5,700
2° T 2004	SI	0	0	127	340	660	1,200	2,000	2,800	5,000
2° T 2004	CI	0	0	150	385	720	1,300	2,200	2,940	5,150
3° T 2004	SI	0	150	240	420	800	1,400	2,350	3,100	5,600
3° T 2004	CI	0	150	250	450	850	1,530	2,537	3,390	5,900
4° T 2004	SI	0	150	250	450	800	1,350	2,200	3,000	5,300
4° T 2004	CI	0	150	280	480	850	1,500	2,420	3,300	5,600
1° T 2005	SI	0	150	280	500	850	1,500	2,450	3,270	6,300
1° T 2005	CI	0	182	300	516	920	1,650	2,624	3,500	6,500
2° T 2005	SI	0	150	298	500	850	1,500	2,400	3,356	6,310
2° T 2005	CI	0	180	300	500	900	1,598	2,600	3,500	6,300
3° T 2005	SI	0	190	300	550	1,000	1,720	2,850	3,825	7,000
3° T 2005	CI	0	200	300	590	1,050	1,860	3,050	4,100	7,000
4° T 2005	SI	0	200	350	580	1,000	1,700	2,800	3,600	6,500
4° T 2005	CI	25	240	350	600	1,050	1,850	3,000	3,800	6,500
1° T 2006	SI	0	200	350	600	1,100	1,950	3,150	4,150	7,600
1° T 2006	CI	0	220	370	630	1,200	2,100	3,450	4,500	7,715
2° T 2006	SI	0	210	370	600	1,100	1,920	3,100	4,100	7,800
2° T 2006	CI	0	240	390	650	1,190	2,000	3,240	4,300	7,800
3° T 2006	SI	0	240	400	700	1,260	2,260	3,600	4,900	8,270
3° T 2006	CI	0	260	400	750	1,357	2,400	3,800	5,100	8,500
4° T 2006	SI	0	280	430	720	1,220	2,110	3,490	4,650	8,900
4° T 2006	CI	50	300	450	780	1,300	2,300	3,700	4,937	8,850

SI: Sin imputación
CI: Con imputación

Referencias Bibliográficas

Sarndal, C.-E., Swensson B., Wretman¸ J. (1992). *Model Assisted Survey Sampling*. Springer.

G. Kalton, Kasprzyk, D (1986). *The treatement of missing survey data*. Survey Methodology 12, 1-16.

Tillé, Y.(2001). *Théorie des sondages*. Dunod.

Haziza, David (2002). *Inference en presence d'imputation: un survol*. Actes des JMS 2002. INSEE.

Medina, F. , Galván, M. (2007). *Imputación de datos: teoría y práctica*. CEPAL. Serie Estudios Estadísticos y Prospectivos.

Singh, Avi , Mohl, Chris (1996). *Understanding calibration estimators in survey sampling*. Survey Methodology, vol. 22 no. 2.

Anders Christianson (Julio 2002). *Forum de discussion*. Statisticien d'Enquêtes, AISE.