

BASE DE DATOS ARMONIZADA

Manual de uso de los datos y consideraciones generales

(Versión 18 de noviembre 2016)



Copyright © [2016] Banco Interamericano de Desarrollo. Esta obra se encuentra sujeta a una licencia Creative Commons IGO 3.0 Reconocimiento-NoComercial-SinObras Derivadas (CC-IGO 3.0 BY-NC-ND) licencia Creative Commons IGO 3.0 Reconocimiento-NoComercial-SinObraDerivada (CC-IGO BY-NC-ND 3.0 IGO) (<http://creativecommons.org/licenses/by-nc-nd/3.0/igo/legalcode>) y puede ser reproducida para cualquier uso no-comercial otorgando el reconocimiento respectivo al BID. No se permiten obras derivadas.

Cualquier disputa relacionada con el uso de las obras del BID que no pueda resolverse amistosamente se someterá a arbitraje de conformidad con las reglas de la CNUDMI (UNCITRAL). El uso del nombre del BID para cualquier fin distinto al reconocimiento respectivo y el uso del logotipo del BID, no están autorizados por esta licencia CC-IGO y requieren de un acuerdo de licencia adicional.

Note que el enlace URL incluye términos y condiciones adicionales de esta licencia.



Manual de uso de los datos y consideraciones generales¹

La base armonizada contiene información de la Encuesta Longitudinal de Protección Social (ELPS) para los países de El Salvador, Colombia, Uruguay, Paraguay y Chile. Este documento es una guía para el usuario respecto del uso de la base armonizada, los archivos que acompañan la base y consideraciones generales al momento de analizar los datos.

I. Manual de uso de los datos de la base armonizada

A. Archivos que acompañan la base armonizada:

Los archivos que acompañan la base armonizada le ofrecen al usuario de los datos información relevante para comprender cómo se crearon las variables, nivel de comparabilidad de los países que forman parte de la base y principales resultados. A continuación se detallan los archivos correspondientes:

1. “*Comparativo_variables_paises.xls*”: presenta las variables utilizadas por país para la generación de la variable armonizada. Cada pregunta por país está marcada con distintos colores y muestra el nivel de comparabilidad de las variables. Cada color representa un código asignado a la variable armonizada. Por ejemplo, como se muestra a continuación la variable armonizada “*relacion_ci*”, que representa la relación de parentesco del entrevistado respecto al jefe de hogar, se le asigna el código 1 a los marcados con amarillo, el código 4 a los marcados con anaranjado, etc.

Variable	El Salvador 2013		Chile 2006	
relacion_ci	3. ¿Cuál es el parentesco que tiene (...) con el (la) jefe (a) del hogar?		a5. ¿Qué relación de parentesco tiene Ud. con el jefe de hogar?	
1 Jefe de hogar	1	Jefe(a)	1	jefe de hogar
2 Cónyuge o pareja	2	Esposa(o), compañera(o)	2	cónyuge o pareja
3 Hijo(a)	3	Hijo(a)	3	hijo(a)
4 Otro familiar	4	Padre/Madre	11	otro familiar
5 Otro no familiar	5	Hermano(a)	12	otro no familiar
	6	Yerno/Nuera		
	7	Nieto(a)		
	8	Suegro(a)		
	9	Otros familiares		
	10	Empleada(o) doméstica(o)		
	11	Jefatura Compartida		
	12	Otros		

El archivo está dividido por secciones que conforman la base de datos como:

- a. Identificación
- b. Demográficas
- c. Educación

¹ Documento elaborado por Sara Correa, consultor del BID para la ELPS Armonizada.

- d. Laboral
- e. Retrospectivas
- f. Capacitación
- g. Seguridad social
- h. Conocimiento de pensiones
- i. Educación financiera
- j. Ingreso
- k. Patrimonio
- l. Salud
- m. Vivienda

Algunas secciones presentan hojas adicionales que complementan la información de las variables armonizadas, por ejemplo, la hoja "CINE_ELPS" muestra cómo están construidas las variables de "niveleducativo_ci" según la Clasificación Internacional Normalizada de la Educación - CINE 2011 de la UNESCO. También, se muestra la hoja "PPA" donde se encuentra la fórmula de cálculo y fuentes de datos utilizadas para la generación de las variables de ingreso expresadas en dólares PPA 2013. La hoja "PET" muestra un cuadro resumen de la población en edad de trabajar (PET), edad de quien responde el módulo de historia laboral, periodo de recolección de los datos, etc.

2. *"Descripcion_variables_ELPS.xls"*: describe la definición y opciones de respuesta de la variable armonizada destacando el universo de la variable y posibles diferencias entre países de acuerdo a cómo fue planteada la pregunta o temas generales a considerar. Además, indica el país y año para los cuales está disponible la variable.
3. *"Estadistica_descriptiva_ELPS.xls"*: ofrece resultados de las variables armonizadas en términos de la media, desviación estándar, mínimo y máximo. Las estadísticas se generan omitiendo la no respuesta representada con el código -4 y aplicando factores de expansión donde se ve modificada la media y desviación estándar. En la última columna se realiza un conteo de la no respuesta para tener noción de los casos omitidos. La interpretación de las estadísticas mencionadas deben ser vistas junto con la descripción de las variables que se encuentran en el archivo *"Descripcion_variables_ELPS.xls"*.
4. *"Documento Metodologico_ELPS"*: se detalla la metodología de armonización de los datos y generación de base armonizada.

B. Variables de interés

La ELPS se diferencia de otras encuestas porque registra información de las historias laborales y previsionales de los entrevistados y sobre todo es una encuesta panel lo cual permite darle seguimiento al entrevistado en temas de salud, previsión social, laboral y condiciones de vida en general.

La base armonizada entrega información de variables retrospectivas que permiten conocer cuál es la situación al momento de la entrevista, hace un 1 año y 2 años desde la fecha de levantamiento de la encuesta reportada por país. Las variables que entregan información a este nivel son:

1. **condocup_ci**: Variable categórica que identifica la condición de ocupación del individuo en relación a la última actividad laboral.
2. **categoinac_ci**: Variable categórica que identifica la condición de inactividad de los entrevistados.
3. **antiguedad_mes_ci**: Variable continua que indica el número de meses de trabajo en la actividad principal actual de la persona ocupada.
4. **categopri_ci**: Variable categórica que indica la categoría ocupacional de la actividad principal para los ocupados.
5. **rama_ci**: Variable categórica que indica la actividad laboral de la ocupación principal según la Clasificación Industrial Uniforme (CIIU) a un dígito según revisión 2.
6. **spublico_ci**: Variable dicotómica que indica con valor 1 si el entrevistado ocupado lleva a cabo su actividad laboral principal en el sector público y 0 para el resto de ocupados. El universo de la variable son los ocupados.
7. **tamemp_ci**: Variable categórica que indica el tamaño de la empresa donde el individuo realiza su actividad laboral principal.
8. **ocupa_ci**: Variable categórica que indica la ocupación laboral de los ocupados en la actividad principal. Se utiliza la Clasificación Internacional Uniforme de ocupaciones CIUO.
9. **horaspri_ci**: Variable continua que indica el número de horas semanales trabajadas en la actividad principal en la semana de referencia.
10. **cotiza_ci**: Variable dicotómica que indica con valor 1 si el entrevistado ocupado cotiza a la seguridad social respecto a su trabajo actual, y con 0 si no lo está. El espíritu de esta variable es capturar la formalidad del empleo.
11. **ylmpri_qmean_ppa_ci**: Variable continua que indica el monto mensual líquido del ingreso laboral de los ocupados provenientes de la actividad principal.
12. **ylmpri_qmean_ppa_ci**: Variable continua que indica el monto mensual líquido del ingreso laboral de los ocupados provenientes de la actividad principal. Monto expresado en dólares PPA 2013.

Las variables terminadas en *_ci*, por ejemplo *condocup_ci*, corresponden al momento más reciente, *1a_ci* se refiere a un año antes (*condocup_1a_ci*) y *2a_ci* se refiere a dos años antes (*condocup_2a_ci*). Información acerca de las fechas de referencia exactas para cada encuesta revisar el documento “*Descripcion_variables_ELPS.xls*”.

C. Indicaciones de uso de variables armonizadas y comandos de STATA

Esta sección presenta una guía del uso de las variables de la base armonizada, haciendo hincapié en aquellas variables que permitan al usuario sacar provecho de la información contenida.

1. Los datos se encuentran en versión STATA 13. Si el usuario cuenta con una versión del STATA menor a la mencionada, se sugiere descargar el comando “use13” mediante la pestaña Help que ofrece STATA. Una vez descargado el comando puede abrir la base usando “use13” en lugar de “use” seguido del directorio de la base. Cuando abra el archivo puede guardarlo en la versión de interés.
2. Para saber el número de observaciones para cada ronda y país disponible, se recomienda utilizar la variable *pais_anio_c*.

```
. tab pais_anio_c
```

Identificad or de la encuesta (País y año)	Freq.	Percent	Cum.
CHL2006	16,443	17.50	17.50
CHL2009	14,463	15.39	32.89
COL2012	14,407	15.33	48.22
PRY2015	15,173	16.15	64.37
SLV2013	15,051	16.02	80.39
URY2013	18,428	19.61	100.00
Total	93,965	100.00	

3. Si se quiere analizar la ELPS de un año específico se puede utilizar “anio_c” o de un país en particular usar “pais_c”. Por ejemplo, las encuestas en 2013 como se indica:

```
. tab pais_c if anio_c==2013
```

País de la encuesta	Freq.	Percent	Cum.
SLV	15,051	44.96	44.96
URY	18,428	55.04	100.00
Total	33,479	100.00	

- La base armonizada contiene la información de las rondas 3 y 4 de Chile (EPS 2006 y EPS2009)². Los datos de las rondas 1 y 2 (EPS2002 y EPS2004) son utilizadas para completar información de algunas variables las cuales se especifican en la descripción de la variable.
- Los datos de la EPS Chile 2006 y 2009 recogen información de los entrevistados que responden en la respectiva ronda. Si se quiere saber de los entrevistados del panel, es decir, si respondieron al menos dos rondas, se utiliza la variable “idpanel” que muestra los años en que ha participado el entrevistado.
- “idpanel” es una variable string que sirve para identificar los años en que el entrevistado ha participado en las diferentes rondas de la ELPS. A continuación se presentan los resultados de la variable para Chile EPS2006 y EPS2009, dado que el resto de países no han implementado rondas posteriores.

² La ronda 5 EPS2012 no se incluye en la base armonizada porque de acuerdo a lo expresado por la Subsecretaría de Previsión Social de Chile (entidad encargada de los datos) no recomiendan que se usen estos datos con fines de inferencia estadística y por tanto podría generar ruido al ser utilizado con el resto de países que no tienen problemas al respecto. A continuación el comunicado.

<https://www.previsionsocial.gob.cl/sps/investigadores/encuesta-de-proteccion-social/bases-de-datos-eps/>

Chile EPS2006

Identificador del entrevistado	Obs.	Porcentaje
2002-2004-2006	1,936	11.77
2002-2004-2006-2009	9,431	57.36
2002-2006	601	3.66
2002-2006-2009	1,146	6.97
2004-2006	535	3.25
2004-2006-2009	2,794	16.99
Total	16,443	100

Fuente: elaboración propia

Chile EPS2009

Identificador del entrevistado	Obs.	Porcentaje
2002-2004-2006-2009	9,431	65.21
2002-2004-2009	579	4
2002-2006-2009	1,146	7.92
2002-2009	333	2.3
2004-2006-2009	2,794	19.32
2004-2009	180	1.24
Total	14,463	100

- a. Si se quiere trabajar con los entrevistados que respondieron en dos rondas consecutivas, por ejemplo, 2006 y 2009 para ver los cambios en las variables de interés, entonces se debe filtrar por las opciones “2002-2004-2006-2009”, “2002-2006-2009”, “2004-2006-2009”.
 - b. Si se quiere identificar al panel “puro”, es decir, aquellos que respondieron en todas la rondas desde 2002 hasta 2009, entonces se considera las opciones “2002-2004-2006-2009” y “2004-2006-2009”, este último caso corresponde a la muestra que se incorporó a la EPS el 2004 correspondiente a nuevos afiliados y no afiliados a la seguridad social.
7. Si se quiere analizar a los pensionados, la base contiene la variable dicotómica “edadlegaljub_ci” que indica con valor 1 a los entrevistados que se encuentran en la edad legal de jubilación o por encima de esta y con cero a quienes se encuentren por debajo de la edad legal de jubilación. Las edades varían por país y género. En el archivo “Descripcion_variables_ELPS.xls” se especifica los criterios para la generación de la variable.
 8. Las variables dicotómicas pension_ej_ci y pensionsub_ej_ci separan a los pensionados que reciben pensiones contributivas y no contributivas respectivamente. Estas variables incluyen a los entrevistados que se encuentran en la edad legal de jubilación o por encima de esta. Si no se quiere limitar por la edad de jubilación se puede utilizar las variables pension_ci y pensionsub_ci.
 9. Para reproducir en la base de datos armonizada las estadísticas mostradas de media y desviación estándar reportadas en el archivo de “Estadística descriptiva_ELPS.xls” se sugiere utilizar el siguiente comando para el caso de la variable estudia_ci que significa “Variable dicotómica que toma el valor de 1 si el entrevistado está estudiando, 0 si no estudia al momento de la entrevista.”

```
. bysort pais_anio_c: sum estudia_ci if estudia_ci!=-4 [w=factor_ci]
```

```
-> pais_anio_c = CHL2006  
(analytic weights assumed)
```

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
estudia_ci	16443	12426437	.1121782	.3155951	0	1

```
-> pais_anio_c = CHL2009
```

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
estudia_ci	14453	12755961	.0623807	.2418541	0	1

```
-> pais_anio_c = COL2012
```

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
estudia_ci	14407	13452558.1	.0389156	.1934006	0	1

```
-> pais_anio_c = PRY2015
```

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
estudia_ci	15173	4547133.55	.1840709	.3875548	0	1

```
-> pais_anio_c = SLV2013
```

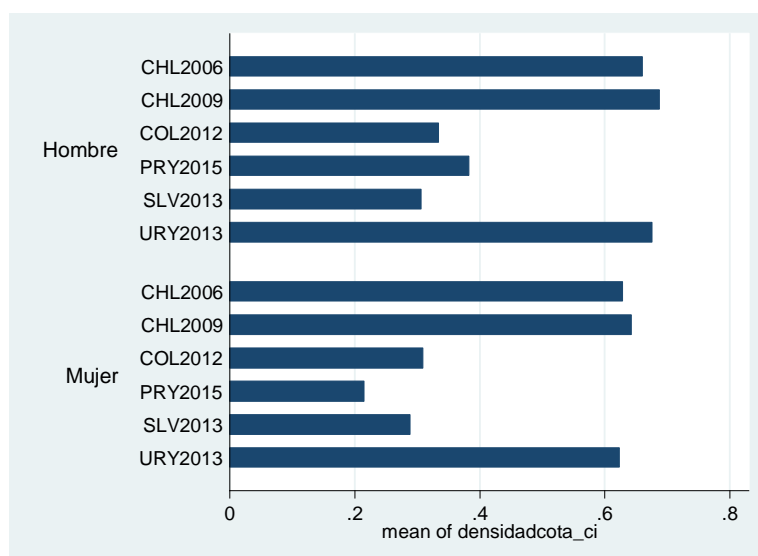
Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
estudia_ci	15051	4079881	.0539151	.2258575	0	1

```
-> pais_anio_c = URY2013
```

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
estudia_ci	18427	2757182	.1152042	.3192769	0	1

10. Si se quiere conocer la densidad de cotización promedio de los entrevistados que se encuentran por debajo de la edad legal de jubilación controlando por sexo y país participante de la ELPS entonces podemos generar lo siguiente:

```
graph hbar (mean) densidadcota_ci [pweight = factor_ci] if edadlegaljub_ci==0 & densidadcota_ci!=-4,  
over(pais_anio_c) over(sexo_ci)
```



11. Si se quiere conocer el nivel de educación de las personas que se encuentran por encima de la edad legal de retiro entonces podemos generar lo siguiente:

```

gen nivel_educ = .
replace nivel_educ=1 if niveleducativo_ci==0 | niveleducativo_ci==1 | niveleducativo_ci==2 |
niveleducativo_ci==3 | niveleducativo_ci==9
replace nivel_educ=2 if niveleducativo_ci==5 | niveleducativo_ci==6 | niveleducativo_ci==7

label define nv 1 "Baja" 2 "Alta"
label values nivel_educ nv
table pais_anio_c nivel_educ if pension_ej_ci==1 [pweight = factor_ci], row col

```

Identificador de la encuesta (País y año)	nivel_educ		
	Baja	Alta	Total
CHL2006	878,729	64,157	942,886
CHL2009	924,107	71,624	995,731
COL2012	490,634	204,878	695,512
PRY2015	49,300.2	21,175.9	70,476.1
SLV2013	124,931	41,835	166,766
URY2013	340,185	51,237	391,422
Total	2807886	454,907	3262793

II. Consideraciones generales:

1. Los datos de ingresos y gastos registran valores extremos en ambas colas de la distribución y distorsionan la media de las variables. Se analizó la no respuesta ya que en algunas variables no resultó clara la identificación de estos casos por la ausencia de labels en la base de datos. Además, se comparó la media de los ingresos de la ELPS con la media de ingresos de encuestas de corte transversal representativos de cada país de

tal manera de verificar que los resultados de la ELPS armonizada sean cercanos a la distribución de ingresos oficiales de estos países.

2. Por lo anterior se reemplazó por missing el 1% en ambas colas de la distribución de ingresos y gastos de la ELPS de tal forma que las medidas de tendencia central resultantes se ajusten a los valores obtenidos en las encuestas de corte transversal que los países típicamente usan para sus análisis de distribución de ingresos y generación de políticas.
3. Las variables de ingresos y gastos se reportan en quintiles y promedio por quintiles para obtener una caracterización más completa de la distribución de ingresos. Por ejemplo, la variable de ingreso laboral de los ocupados provenientes de la actividad principal se expresa en tres variables:
 - a. ylmpri_qmean_cte_ci: monto promedio del quintil al que pertenece el entrevistado expresado en moneda local corriente.
 - b. ylmpri_qui_ci: variable categórica que indica los quintiles de ingreso de la ocupación principal.
 - c. ylmpri_qmean_ppa_ci: monto promedio del quintil al que pertenece el entrevistado expresado en dólares PPA 2013.

Este mismo esquema se aplica a las variables retrospectivas de ingreso hace 1 año y 2 años:

1. ylmpri_1a_qmean_cte_ci
2. ylmpri_1a_qui_cte_ci
3. ylmpri_1a_qmean_ppa_ci

4. PARAGUAY:

- a. La base armonizada (versión 18 de noviembre 2016) cuenta con información actualizada de la base de datos de Paraguay 2015. Esta nueva versión incluye un set de variables adicionales creadas por los encargados de los datos en Paraguay tales como 1) escolaridad, 2) última situación ocupacional, 3) razón de inactividad de la última historia laboral, 4) categoría de ocupación de la última historia laboral, 5) historia laboral de enero 2013, 6) historia laboral de enero 2015 y 7) densidad de cotización.
- b. Estas variables también son calculadas en la base armonizada. Las variables 2), 3) y 4) coinciden con los cálculos realizados en la base armonizada. Las variables 5) y 6) corresponden a la situación ocupacional a enero 2013 y enero 2015 y no son comparables con las variables calculadas en la base armonizada debido a que se consideró como periodo de referencia 12 y 24 meses antes de la situación laboral registrada en el módulo de historia laboral. Las variables 1) y 7) discrepan de los cálculos de la base armonizada y no se cuenta con mayores antecedentes de la forma de cálculo de estas variables.

5. URUGUAY:

- a. La base armonizada (versión 18 de noviembre 2016) también cuenta con información actualizada de la base de datos de Uruguay 2013. La principal modificación de la base está relacionada con los factores de expansión. Mayores detalles de los cambios realizados a la base ver documento “cambios_base_actual_ELPS_URY2013.doc” disponible en el OSS.

- b. Uruguay realiza el proceso imputación de sus datos a través de la metodología de imputación múltiple que consiste en generar múltiples imputaciones ($m=5$) a las variables de interés.
- c. La presentación de los datos de la ELPS se realizan mediante 2 formatos: a) mlong que contiene las múltiples imputaciones ($m=5$) en un solo archivo y por lo tanto el tamaño de la base es extensa y b) flongsep que guarda una imputación en cada archivo, es decir, se cuenta con 5 bases imputadas ($m=1, m=2, \dots m=5$) y 1 base con los datos originales sin imputar ($m=0$).
- d. Para trabajar la base de Uruguay es necesario que el usuario se familiarice con los comandos disponibles en el software estadístico de STATA puesto que los comandos que usualmente se usan para inferencia estadística no aplican a este tipo de formato de datos. Se recomienda descargar el manual de STATA en la pestaña Help, luego elegir "PDF Documentation". Se descargará el archivo en formato pdf y después seleccionar en la barra izquierda del menú "[MI] Multiple Imputation".
- e. Adicional a este apoyo, se encuentra disponible el documento metodológico de imputación múltiple donde además de explicar el proceso llevado a cabo se adjunta el Anexo III que contiene una sesión de trabajo con la ELPS imputada.
- f. Para efectos de la base armonizada, el tratamiento que se aplicó a las variables imputadas fue generar una sola variable con el promedio de los cinco valores reportados. Por ejemplo, el monto del ingreso de la ocupación principal reporta 5 montos imputados y se resumen en una sola variable como el promedio de estos cinco valores. Luego de realizar este proceso a las variables imputadas de interés para la base armonizada se actualiza los cambios realizados con el comando `mi update` y luego se transforma la base en su versión original con el comando `"mi extract 0, clear"` para proceder a la generación de las variables armonizadas.