Guía de Instrucciones para Uso de ELSOC 2016-2019

(versión wide)

*Roberto González, Matías Bargsted, Daniel Miranda, Héctor Carvacho,   
Edgardo Cerda, Monserratt Mella y Alejandro Plaza*

*14 de Octubre de 2020*

**Introducción**

El siguiente informe tiene el objetivo de presentar las instrucciones generales de uso de la base de datos de ELSOC 2016-2019. El equipo ELSOC se encargó de combinar las bases de datos y realizar las homologaciones y cambios pertinentes para facilitar su uso por investigadores del Centro. A su vez, este documento reseña brevemente algunas particularidades relacionadas a la estructura de los datos y su análisis.

**Formato de la Base de Datos**

Existen dos tipos de estructura de datos para estudios paneles o longitudinales: formato wide (o ancho) y formato long (o largo). En el formato wide, las filas representan individuos seguidos en la muestra, y las columnas representan distintas mediciones a dicho individuo. La Figura 1 provee un ejemplo de dicho formato, dónde *Country* permite identificar los 5 casos en la base de datos. Nótese que los atributos que son medidos varias veces en el tiempo (*college* y *jobs*) incluyen sufijos que permiten identificar el tiempo de medición. Las respuestas a un mismo ítem en años distintos se encuentran siempre en dicha fila (no puede haber dos filas para un mismo sujeto) y en columnas separadas. El formato de la base de datos combinada de ELSOC 2016-2019 es **wide** y el equipo ELSOC reordenó las variables de modo tal que dos mediciones de un mismo ítem aparecen de manera consecutiva y añade un sufijo para indicar la ola del estudio correspondiente (además, todas las etiquetas de variables señalan el año de la medición).



Figura 1: Ejemplo de Formato de Base de Datos Longitudinal Wide

Por otro lado, en el formato long cada fila es un sujeto en un punto del tiempo, de manera que cada sujeto puede tener múltiples filas. Aquí las variables reflejan atributos y pueden contener más de una medición en el tiempo para un individuo. Por este motivo, este formato requiere contar con una variable que localice tanto al sujeto en el tiempo como el folio identificador de éste. La Figura 2 ejemplifica el formato long, siendo *Country* el identificador de casos y *Year* la variable que señala el tiempo de medición. En contraste a la Figura 1 dónde existían 4 columnas para *college*, aquí hay solo una que combina las mediciones en el tiempo. En la actualidad, los softwares estadísticos pueden usar ambos formatos, y en caso de ser necesario pueden ser transformados. Los usuarios de ELSOC pueden transformar la base de datos a formato long si así lo requieren, siendo responsables por dicho procedimiento. En caso de que esté interesado en dicho cambio puede explorar:

**• En R**: comando pivot\_longer() en paquete *tidyr*, comando gather() en paquete *tidyverse*, to\_long() en paquete *sjmisc* y melt() en *reshape2*. Si desea ejemplos visite [**este vínculo**](http://www.cookbook-r.com/Manipulating_data/Converting_data_between_wide_and_long_format/). Sobre el comando to\_long entre [**aquí**](https://www.rdocumentation.org/packages/sjmisc/versions/2.6.3/topics/to_long).

• **En Stata**: el comando reshape. Para conocer un ejemplo entre [**aquí**](https://www.rdocumentation.org/packages/sjmisc/versions/2.6.3/topics/to_long)**.**

• **En SPSS**: si trabaja por Menú, vaya a Data/Restructure Data Wizard para la transformación del formato. Si tiene dudas puede revisar [**este vínculo**](https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/base/idh_rdwiz_c2v_step2.html). Aquí puedes encontrar[**un ejemplo**](https://stats.idre.ucla.edu/spss/modules/reshaping-data-wide-to-long-in-versions-11-and-up/).



Figura 2: Ejemplo de Formato de Base de Datos Longitudinal Long

Tal como se mencionó, el formato de ELSOC combinado 2016-2019 es **wide**. La Figura 3 presenta la estructura de los datos. En este ejemplo se toma como ilustración las variables de actitudes hacia la democracia **c01** (para conocer el listado completo de variable se recomienda revisar los documentos anexos de **Listado de Variables** y los **Manuales de Usuario**, los cuales contienen los Libros de Códigos asociados a cada ola del estudio). A cada variable se les agrega un sufijo indicando a qué medición corresponde cada medición. De esta manera **w01** corresponde a la primera ola (de 2016), **w02** a la segunda ola (2017), y así sucesivamente. Las variables en la base de datos están ordenadas de modo tal que las variables que representan la segunda (o tercera) medición de una pregunta está inmediatamente después de la primera medición (para revisar las variables que se repiten en el tiempo, ver documento de Listado de Variables). Adicionalmente, la base de datos **wide** contiene etiquetas de variables que señalan el año de la medición (aparece el año de medición entre paréntesis).



Figura 3: Formato de ELSOC 2016-2018

**Homologación de Variables en ELSOC Longitudinal**

El equipo ELSOC se encargó de armonizar las variables que modificaron su escala de medición. A continuación, se detallan dichos cambios:

1) La variable c17 (Identificación con coaliciones políticas) cambió su escala de medición entre 2016 y 2017.

* En el año 2016 esta variable consulta sobre la coalición política que representa mejor sus intereses, creencias y valores. En esta oportunidad las categorías de respuestas corresponden a: 1. Chile Vamos; 2. Nueva Mayoría y 3. Otras. -No obstante, para el año 2017 con la emergencia política del Frente Amplio, el equipo de ELSOC decide ingresar esta coalición dentro de las respuestas posibles.
* A partir de esto, la recodificación final de esta pregunta para ambas olas queda como: 1. Chile Vamos; 2. Nueva Mayoría; 3. Frente Amplio; y 4. Otra.

2) La variable c20 (Identificación con movimiento social) agregó categorías de respuesta el año 2018 y 2019.

* En el año 2016 y 2017 esta variable consulta sobre el movimiento social que más valora el entrevistado. En dichas mediciones las categorías de respuestas corresponden a: 1. Movimiento de apoyo a la causa estudiantil; 2. Movimiento de apoyo a demandas laborales; 3. Movimiento de grupos ambientalistas; 4. Movimiento de apoyo a las demandas indígenas; 5. Movimiento de apoyo a la diversidad sexual; 6. Movimiento social provida o antiaborto; 7. Movimiento social antidelincuencia; 8. Otro. Especifique; 9. Ninguno (no leer).
* No obstante, tras la revisión de las respuestas de los dos primeros años y los cambios en la política nacional, el año 2018 el equipo de ELSOC decide incorporar como posibles respuestas (8) movimiento feminista y (9) movimiento por el cambio al sistema de pensiones.
* Asimismo, el año 2019 tras el estallido social que surgió en el país el equipo de ELSOC decide agregar una nueva categoría de respuesta: (11) movimiento social de Octubre 18/O.

**Dimensiones de la Base de Datos**

La base de datos ELSOC 2016-2019 en formato wide contiene 4.447 filas y 1.536 columnas. Las 4.447 filas corresponden a 4.447 entrevistados para los cuales hay respuestas entre las olas 2016 y 2019, incluyendo tanto a los sujetos que abandonaron el estudio (atrición muestral) como a los individuos incluidos como parte de la muestra de refresco. Los casos perdidos por atrición son identificados por medio de la variable **tipo\_atricion**.

Cabe mencionar que el diseño muestral de ELSOC es de panel repetido, incorporándose desde la tercera medición (año 2018) una muestra independiente que funciona como refresco. Hay 1.519 casos incorporados a ELSOC como parte de la muestra de refresco. Para más detalles sobre el diseño muestral de la muestra de refresco consulte el Manual de Usuario de Corte Transversal 2018.

A continuación, la Tabla Nº 1 presenta la descomposición de los casos incluidos en la base de datos:

|  |  |  |
| --- | --- | --- |
| Tabla N°1: Distribución tipo atrición | | |
| Tipo Atrición | Frecuencia | Porcentaje |
| 2016-2019 | 1.913 | 43,02 |
| 2016-2018 | 193 | 4,34 |
| 2016, 2017 y 2019 | 129 | 2,90 |
| 2016, 2018-2019 | 84 | 1,89 |
| 2016-2017 | 238 | 5,35 |
| 2016 y 2018 | 38 | 0,85 |
| 2016 y 2019 | 27 | 0,61 |
| Sólo 2016 | 305 | 6,86 |
| Sólo 2018 (M. original) | 1 | 0,02 |
| 2018-2019 | 1.264 | 28,42 |
| Sólo 2018 (M. refresco) | 255 | 5,73 |

De la Tabla Nº 1 se puede apreciar que 1.913 entrevistados fueron medidos en las cuatro olas. El tamaño muestral original (ola 2016) fue de 2.927 casos, por lo que el 65,36% de dichos casos ha sido entrevistado en cuatro ocasiones. Por su parte, 4,3% de los entrevistados fueron medidos en las tres primeras olas; mientras que 4,8% de los entrevistados fueron medidos también en tres olas, pero no consecutivas. Además, 35,23% de los entrevistados tienen dos mediciones, de los cuales el 28,42% pertenecen a la muestra refresco que se incluyó en el estudio el año 2018, por lo que se podría afirmar que estos últimos han participado de todas las olas que se les ha contactado.

Cabe mencionar que hay entrevistados que sólo fue posible contactar un año, ya sea el año 2016 (305 casos) o el año 2018 (256 casos). Los del año 2018, 255 casos son de la muestra refresco que no fue posible su re-contacto en la segunda oportunidad (año 2019), pero hay un caso perteneciente a la muestra original (Código de Disposición Final = 120) que no completó la entrevista durante 2016, no pudo ser contactado durante 2017, sí fue contactado y se completó la entrevista en 2018, pero no se pudo contactar el año 2019. Esto puede ocurrir porque la estrategia de campo de ELSOC implica que cada año se intenta re-contactar a todos los participantes seleccionados en la muestra, independiente de su participación en la medición anterior.

Las 1.536 columnas corresponden a las variables incluidas en el estudio. No todas las preguntas incluidas en el cuestionario de ELSOC son medidas todos los años (*permanentes*), ya que algunas sólo se miden en una ocasión (*atributos fijos* en el tiempo) y otras son alternadas según olas (ítems *intercalados* o *alternados*). En la documentación complementaria puede encontrar detalles sobre las mediciones de cada uno de los ítems del cuestionario.

**Análisis de Consistencia**

Como todos los años, el Equipo ELSOC el año 2019 se encargó de realizar un pormenorizado análisis de consistencia de los casos entre olas[[1]](#footnote-1), lo que sumado a los controles de calidad efectuados por Centro MicroDatos garantizan altos estándares en los datos. A finales del año 2018, a petición del equipo ELSOC, Centro MicroDatos se encargó de recontactar a cada uno de los casos incluidos en ELSOC para determinar si se habían seguido adecuadamente los protocolos de reentrevista: si en el mismo hogar se reentrevistó al mismo sujeto seleccionado. Tras dichos chequeos se corrigió la información de algunos entrevistados, cambios que fueron incorporados en las versiones 3.00 (en adelante) de ELSOC 2016 y 2.00 (en adelante) de ELSOC 2017.

Adicionalmente, fueron eliminados de la base de datos aquellos entrevistados que presentaban respuestas inconsistentes en atributos sociodemográficos claves y que, tras la revisión de MicroDatos, se determinó que correspondían a casos con errores de validación o con fallas en el proceso de reentrevista. Mayores detalles de estos cambios son documentados en el informe “*Reporte de Correcciones ELSOC 2016-2017*”.

ELSOC sigue altos estándares de calidad, incluyendo exigentes criterios de supervisión de campo y validación de datos. No obstante, sigue siendo una encuesta que recolecta información reportada por individuos. No es posible anular el error de medición ni tampoco impedir que existan inconsistencias en las respuestas de los entrevistados a lo largo del cuestionario o entre distintas mediciones. Dada la magnitud de la base de datos, no es posible para el equipo ELSOC revisar cada una de estas potenciales inconsistencias. Son los usuarios de los datos quiénes deben ponderar el modo más adecuado de abordar estos aspectos.

Sin embargo, el equipo ELSOC el año 2019 sí analizó la consistencia intertemporal de tres atributos sociodemográficos clave (sexo, edad y educación), creando la variable **tipo\_caso,** que clasifica a los sujetos según los patrones de consistencia en dichas variables sociodemográficas. Este análisis se volvió a hacer el año 2020, detectando que 4.055 casos (tamaño muestral 4.447) no presentan inconsistencias en sexo, edad y/o educación[[2]](#footnote-2). En otras palabras, para un 91,2% de los casos seguidos desde ELSOC 2016 (o 2018 para muestra refresco) no hay diferencias anómalas en sus reportes de sexo, edad y/o educación a lo largo del tiempo. En el otro extremo, hay 12 entrevistados (representan un porcentaje inferior al 1%) en la base de datos cuyas respuestas presentan notorias inconsistencias intertemporales en estos atributos sociodemográficos clave. Por último, hay 362 casos (8,1% de casos) que presentan inconsistencias menores en uno o dos de estos atributos. Estos casos son catalogados como *pendientes*, ya que en la medida que se recolecte más información (nuevas olas del estudio), más factible será discernir la calidad de las respuestas provistas por estos entrevistados. Los usuarios deberán determinar cómo trabajar con dichos casos. En base a la información entregada pueden “reparar” los datos por medio de técnicas de imputación u optar por descartarlos de su análisis. La base de datos de ELSOC contiene preguntas relativas a nacionalidad (m45), estado civil (m36), religión (m38) y pertenencia a pueblos originarios (m53) que pueden ser usadas por los usuarios para discernir sobre los niveles de inconsistencias de las respuestas de los encuestados.

**Diseño Muestral de ELSOC**

Si desea conocer detalles sobre el diseño muestral de ELSOC se recomienda revisar los Manuales de Usuario de Corte Transversal 2016, 2017, 2018 y 2019, en que se describen los procedimientos para la selección de la muestra.

La base de datos incluye variables para análisis de muestras complejas:

* estrato: indica el tipo de ciudades que se definieron para la selección muestral.
* segmento: manzana censal (código ficticio para garantizar secreto estadístico).
* ponderadores: pesos muestrales que ajustan según probabilidad de selección y otros factores la distribución muestral manteniendo fijo el tamaño muestral.
* factores de expansión: pesos muestrales que ajustan según probabilidad de selección y otros factores la distribución muestral y expanden las frecuencias al tamaño de la población objetivo.

**Los ponderadores y factores de expansión incluidos en ELSOC 2016-2019 son pesos muestrales diseñados con una lógica de corte transversal**. A la fecha, no se encuentran disponibles los ponderadores longitudinales, los cuales serán elaborados por expertos en análisis estadístico.

La variable **idencuestador** no se encuentra homologada entre ambas olas. Varios de los encuestadores trabajaron en las distintas olas de ELSOC, pero a la fecha no ha sido posible armonizar dicha variable.

**Variables incluidas en la Base de Datos**

Dada la naturaleza panel de la encuesta ELSOC, las variables que se presentan en la base de datos tienen ciertas características particulares a razón de mantener consistencia y claridad en la estructura de la información. A continuación, se presentan algunas características que es necesario tomar en cuenta:

1. **Variables recodificadas**: Las siguientes variables fueron recodificadas.

* **c17** (identificación con coaliciones políticas).
* **c20** (movimiento social más valorado).
* **idencuestador** (folio identificador de encuestadores). Los folios de encuestadores fueron construidos con una lógica de corte transversal, con valores numéricos consecutivos. Sin embargo, buena parte de los encuestadores no trabajó en ambas versiones del estudio. A la fecha, no ha sido posible homologar la información de encuestadores. Para evitar posibles confusiones en los análisis, el equipo ELSOC decidió agregar un prefijo de centenares de miles que permitan diferenciar entre las olas. De esta forma la identificación para un encuestador de la ola 1, está en una escala de 100.000 y un encuestador de la ola 2 está en una escala de 200.000. En la medida de que la encuesta ELSOC tenga más año, este número tendría la misma lógica.

2. **Variables Especiales:** la base de ELSOC 2016-2019 contiene variables específicas para describir la naturaleza de los datos. Éstas son:

* idencuesta: folio identificador de los encuestados.
* tipo\_atricion: cuáles son las olas incluidas en la versión de la base de datos combinada.
* tipo\_caso: clasifica los casos según su consistencia intertemporal en los atributos sexo, edad y educación.
* version: versión de la base de datos combinada.
* muestra: indica la muestra del panel ELSOC.
* cuestion\_mig: indica la nacionalidad de los migrantes que se les mencionó en las preguntas r05, r06, r07, r08, r09, r10, r11, r12, r16, r17, r18, d01\_04, f01\_05. Con valores (1) Peruanos, (2) Venezolanos y (3) Haitianos.

3. **Naturaleza de las variables**: en vista de que ELSOC 2016-2019 está disponible en tres formatos disponibles (R, Stata y SPSS), se optó por un criterio ecuánime al momento de definir las variables. La gran mayoría de éstas son almacenadas en formatos numéricos o entero. **Cada usuario debe preocuparse de manipular las variables si desea usarlas como variables categóricas** según los procedimientos acorde a cada software estadístico.

Por otra parte, se tienen variables de cadena o carácter, que en su gran mayoría son variables en las que se le pide especificar al encuestador un atributo que no está disponible en el abanico de posibles respuestas. Para reducir problemas de compatibilidad entre los distintos softwares, todas las etiquetas de variables están codificadas sin tildes, diéresis y sin la letra “ñ”. La codificación empleada para este proceso corresponde a UTF-8.

4. **Valores Perdidos “No Sabe” y “No Responde”**: A partir de las especificidades propias que comprende un estudio panel en relación a la homologación de datos entre olas, el equipo ELSOC decidió como política para los casos perdidos asignar un código único de -999 a “No Responde” y -888 “No Sabe”. Dentro de la base de datos panel los códigos se encuentras asignados de esta forma y en las siguientes olas se mantendrá esta codificación.

Anexo a este documento introductorio se provee, dentro de cada Manual de Usuario de corte transversal (2016, 2017, 2018 y 2019), un Libro de Códigos que describe cada una de las variables incluidas en ELSOC 2016, 2017, 2018 y 2019, incluyendo información relativa a los valores y etiquetas de valores de cada variable.

**Recomendaciones para el Análisis**

Algunas consideraciones prácticas para los usuarios de la base de datos:

• La pregunta **m40** consulta por el nombre del último colegio en que asistió en educación y la comuna dónde se localiza. El ítem relativo a la comuna fue entregado como variable de caracteres (cadena). En 2017 la base de datos para la pregunta **m50** (misma pregunta sobre educación secundaria) la variable de comuna fue codificada. El equipo ELSOC se encargó de la homologación de la información de 2016 (variables m40\_comuna y m40\_pais), pero los usuarios interesados en dicha información pueden revisar la codificación.

• Las variables cod\_m03, cod\_m04, cod\_m22 y cod\_m23 fueron elaboradas por Centro MicroDatos y aluden a la codificación de ocupación y giro de la empresa para el entrevistado y jefe de hogar). Para las mediciones de la ola 2016[[3]](#footnote-3) se utilizaron los códigos CIUO 88 y CIIU (3ra Versión), mientras que las mediciones de la ola 2018 utilizan los códigos CIUO 08 y CIIU (4ta Versión). Se recomienda cautela al trabajar intertemporalmente con dichas variables.

• La información relativa a caracterización del hogar en 2016 (variables miembroXX\_sexo y miembroXX\_ edad) **no tiene una correspondencia inequívoca** con la información del hogar en 2017 (m46\_ocup\_XX y m46\_educ\_XX). En primer lugar, ELSOC es una encuesta de individuos y es factible que los hogares de éstos varíen su composición a lo largo del tiempo. Luego, la información de 2016 fue obtenida dentro del proceso de empadronamiento (Tabla Kish). Es posible construir indicadores agregados que caractericen al hogar, pero no existe correspondencia entre miembro01\_sexo y m46\_educ\_01: **no necesariamente corresponden al mismo sujeto.** Micro Datos almacena los nombres de pila pero por motivos de confidencialidad dicha información no es liberada. El equipo ELSOC evaluará la posibilidad de armonizar dicha información en versiones futuras de la base de datos.

• La versión 1.00 de ELSOC 2016-2019 no contiene variables territoriales. El equipo ELSOC liberará dicha información en versiones futuras.

**Datos en R**

La base de ELSOC 2016-2019 se libera en tres formatos: SPSS, Stata (versiones 13 y 14) y R. Los usuarios de los primeros dos softwares están acostumbrados al uso de etiquetas de variables y valores. Sin embargo, en el entorno de R no es tan frecuente su uso. El equipo ELSOC desarrolló etiquetas de variables y valores con los paquetes *haven* y *sjlabelled*, por lo cual en conjunto las variables tienen incorporadas como atributos las etiquetas. Si los usuarios están interesados en utilizar las etiquetas se recomienda utilizar los paquetes *sjlabelled, sjmisc* y *sjPlot*.

**Dudas y Consultas**

En caso de dudas y consultas sobre la base de datos longitudinal se recomienda revisar la documentación asociada al estudio. Adicionalmente, los interesados pueden contactarse con el equipo ELSOC en el correo electrónico encuestacoes@gmail.com.

**Citación de Base de Datos**

La base de datos debe ser citada con el siguiente formato:

Centro de Estudios de Conflicto y Cohesión Social (2020). Estudio Longitudinal Social de Chile, Versión Panel Combinada 2016-2019 (ELSOC\_Long\_2016\_2019\_v1.00). [Archivo de datos]. Santiago, Chile: Centro de Estudios de Conflicto y Cohesión Social (COES). www.coes.cl

Centro de Estudios de Conflicto y Cohesión Social (2020). Estudio Longitudinal Social de Chile, Versión Panel Combinada 2016-2019 (ELSOC\_Wide\_2016\_2019\_v1.00). [Archivo de datos]. Santiago, Chile: Centro de Estudios de Conflicto y Cohesión Social (COES). www.coes.cl

1. Este análisis se realizó para la muestra original. [↑](#footnote-ref-1)
2. En esta categoría también se clasifican individuos que cambia su nivel educativo, pero dicho cambio es plausible de ser verídico. A su vez, también incluye casos que abandonan la muestra (atrición) y no hay evidencia que señale inconsistencia entre olas. [↑](#footnote-ref-2)
3. Se incluyen las correcciones propuestas por el equipo del investigador COES Vicente Espinoza. [↑](#footnote-ref-3)