Guía de Instrucciones para Uso de ELSOC 2016-2018 (versión wide y long)

Roberto González, Matías Bargsted, Daniel Miranda, Hector Carvacho, Benjamín Muñoz y Alejandro Plaza

23 de Abril de 2019

Introducción

El siguiente informe tiene el objetivo de presentar las instrucciones generales de uso de la base de datos de ELSOC 2016-2018 en formato wide y long, la cual es distribuida a todos los miembros de la comunidad COES. 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 fortmato 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 item en años distintos se encuentran siempre en dicha fila (no pueden haber dos filas para un mismo sujeto) y en columnas separadas. El formato de la base de datos combinada de ELSOC 2016-2018 es wide y el equipo ELSOC reordenó las variables de modo tal que dos mediciones de un mismo item 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).

	County	LandArea	NatAmenity	College 1970	College1980	College 1990	College2000	Jobs 1970	Jobs 1980	Jobs 1990	Jobs2000
1	Autauga	599	4	.064	.121	.145	.180	6853	11278	11471	1628
2	Baldwin	1578	4	.065	.121	.168	.231	19749	27861	40809	7024
3	Barbour	891	4	.073	.092	.118	.109	9448	9755	12163	1519
4	Bibb	625	3	.042	.049	.047	.071	3965	4276	5564	609
5	Blount	639	4	.027	.053	.070	.096	7587	9490	11811	1650

Figure 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 mutiples 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 al sujeto en el tiempo además del 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. Sobre el comando to_long

entre aquí.

- En Stata: el comando reshape. Para conocer un ejemplo entre aquí
- En SPSS: si trabaja por Menú, vaya a Data/Restructure Data Wizard para la transformación del formato. Si tiene dudas puede revisar este vínculo. Aquí puedes encontrar un ejemplo.

	County	LandArea	NatAmenity	Year	College	Jobs
1	Autauga	599	4	1970	.064	6853
2	Autauga	599	4	1980	.121	11278
3	Autauga	599	4	1990	.145	11471
4	Autauga	599	4	2000	.180	16289
5	Baldwin	1578	4	1970	.065	19749
6	Baldwin	1578	4	1980	.121	27861
7	Baldwin	1578	4	1990	.168	40809
8	Baldwin	1578	4	2000	.231	70247
9	Barbour	891	4	1970	.073	9448
10	Barbour	891	4	1980	.092	9755
11	Barbour	891	4	1990	.118	12163
12	Barbour	891	4	2000	.109	15197
13	Bibb	625	3	1970	.042	3965
14	Bibb	625	3	1980	.049	4276
15	Bibb	625	3	1990	.047	5564
16	Bibb	625	3	2000	.071	6098
17	Blount	639	4	1970	.027	7587
18	Blount	639	4	1980	.053	9490
19	Blount	639	4	1990	.070	11811
20	Blount	639	4	2000	.096	16503

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

Tal como se mencionó, el formato de ELSOC combinado 2016-2018 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 del 2016, w02 a la segunda ola del 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).

	idencuesta	version	olas	muestra	casos_sosp~s	c01_w01	c01_w02
1	1101011	Version Lon	2016-2017	Muestra Ori…	Sin problem	Nada satisf	Poco satisf.
2	1101012	Version Lon	2016-2017	Muestra Ori…	Sin problem	Nada satisf	Nada satisf.
3	1101013	Version Lon	2016-2017	Muestra Ori…	Sin problem	Nada satisf	Nada satisf.
4	1101021	Version Lon	2016-2017	Muestra Ori…	Sin problem	Nada satisf	Algo satisf.
5	1101022	Version Lon	2016-2017	Muestra Ori…	Sin problem	Poco satisf	Poco satisf.
6	1101023	Version Lon	2016-2017	Muestra Ori…	Sin problem	Nada satisf	Bastante sa.
7	1101031	Version Lon	2016-2017	Muestra Ori…	Sin problem	Algo satisf	Poco satisf.
8	1101032	Version Lon	2016-2017	Muestra Ori…	Sin problem	Algo satisf	Nada satisf
9	1101033	Version Lon	2016-2017	Muestra Ori…	Sin problem	Algo satisf	Algo satisf
10	1101041	Version Lon	2016-2017	Muestra Ori…	Sin problem	Nada satisf	Nada satisf
11	1101042	Version Lon	2016-2017	Muestra Ori…	Sin problem	Bastante sa	Nada satisf
12	1101043	Version Lon	2016-2017	Muestra Ori	Sin problem	Algo satisf	Algo satisf.

Figure 3: Formato de ELSOC 2016-2018

En esta nueva entrega se ha incluido una nueva versión en formato **long**, la cual es bastante sencilla de usar. Básicamente, la base se estructura como base de datos en que los individuos son identificados por la conjunción del folio (variable **idencuesta**) con el año de medición (variable **ola**). A su vez, la base de datos tiene los valores perdidos, válidos y etiquetas homologados de modo tal que no sean necesarios transformaciones para comparar las respuestas entre olas.

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:

- 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.
- La variable c20 (Identificación con movimiento social) cambió su escala de medición entre 2017 y 2018.
- 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 equipo de ELSOC decide incorporar como posibles respuestas al movimiento por el cambio al sistema de pensiones y al movimiento feminista.
- A partir de esto, la recodificación final de esta pregunta para ambas olas queda como: 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. Movimiento social feminista o de apoyo a la igualdad de género; 9. Movimiento por el cambio al sistema de pensiones; 10. Otro. Especifique; 11. Ninguno (no leer).

Dimensiones de la Base de Datos

La base de datos ELSOC 2016-2018 en formato wide contiene 4447 filas y 1053 columnas Las 4447 filas corresponden a 4447 entrevistados para los cuales hay respuestas en las olas 2016 y 2018, 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**. A continuación, la Tabla Nº 1 presenta la descomposición de los casos incluidos en la base de datos:

Tipo de Caso	Frecuencia	Porcentaje
2016-2018	2,106	47.36
2016 y 2017	367	8.25
2016 y 2018	122	2.74
Solo 2016	332	7.47
Solo 2018 (M. Original)	1	0.02
Solo 2018 (M. Refresco)	1,519	34.16

De la Tabla Nº 1 se puede apreciar que 2106 entrevistados fueron medidos en las tres olas. El tamaño muestral original (ola 2016) fue de 2927 casos, por lo que el 71.95% de dichos casos ha sido entrevistado en tres ocasiones. Adicionalmente, hay entrevistados que fueron medidos sólo en 2016 (332 casos), en 2016 y 2017 (367 casos) y en 2016 y 2018 (122 casos). Esto se debe a que la estrategia de campo de ELSOC implica que cada año se intenta recontactar a todos los participantes seleccionados en la muestra, independiente de su participación en la medición anterior. Producto de lo anterior, hay un caso perteneciente a la muestra original que no completó la entrevista durante 2016 (Código de Disposición Final=120) y no pudo ser contactado durante 2017, pero sí fue contactado y se completó la entrevista en 2018. Producto de lo anterior, sólo en esta tercera medición es incluido.

El diseño muestral de ELSOC es de panel repetido, incorporándose desde la tercera medición (ola 2018) una muestra independiente que funciona como refresco. Hay 1519 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.

Las 1052 columnas corresponden a las variables incluidas en el estudio. No todas las preguntas incluidas en el cuestionario de ELSOC son medidas todos los años (son *permanentes*), ya que algunas sólo se miden en una ocasión (atributos fijos en el tiempo) y otras son alternadas según olas (items *intercalados* o *alternados*). En la documentación complementaria puede encontrar detalles sobre las mediciones de cada uno de los itemes del cuestionario.

Análisis de Consistencia

El Equipo ELSOC se encargó de realizar un pormenorizado análisis de consistencia de los casos entre olas lo que sumado a los controles de calidad efectuados por Centro MicroDatos garantizan altos estándares en los datos. A finales del año 2018 y petición del equipo ELSOC, Centro MicroDatos se encargó de recontactar cada uno de los casos incluidos en ELSOC y determinó 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.

¹Evidentemente, este análisis es válido únicamente para la muestra original.

Adicionalmente, fueron eliminados de la base de datos áquellos 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".

No obstante ELSOC sigue altos estándares de calidad, incluyendo exigentes criterios de supervisión del campo y validación de datos, 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 sí analizó la consistencia intertemporal de tres atributos sociodemográficos clave (sexo, edad y educación). En base a dicho análisis se creó la variable **tipo_caso** que clasifica a los sujetos según los patrones de consistencia en dichas variables sociodemográficas. Para 2711 casos de la muestra original (tamaño muestral 2928) no se detectaron inconsistencias en sexo, edad y educación². En otras palabras, para un 92.59% de los casos seguidos desde ELSOC 2016 no hay diferencias anómalas en sus reportes de sexo, edad y 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 205 casos (7% muestra) 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 y 2018, dónde 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 2016-2018 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. En futuras versiones de la base de datos serán incluidos dichos pesos muestrales.

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:

 $^{^2}$ 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.

- 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 2016-2017 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 casos: 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.
- 3. Naturaleza de las variables: en vista de que ELSOC 2016-2018 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. Por otro lado, 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 y 2018), un Libro de Códigos que describe cada una de las variables incluidas en ELSOC 2016, 2017 y 2018, 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 item relativo a la comuna fue entregada 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 2016³ 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 2016-2018 no contiene variables territoriales. El equipo ELSOC liberará dicha información en versiones futuras.

Datos en R

La base de ELSOC 2016-2018 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 (2019). Estudio Longitudinal Social de Chile, Versión Panel Combinada 2016-2018 (ELSOC_Wide_2016_2018_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 (2019). Estudio Longitudinal Social de Chile, Versión Panel Combinada 2016-2018 (ELSOC_Long_2016_2018_1.00). [Archivo de datos]. Santiago, Chile: Centro de Estudios de Conflicto y Cohesión Social (COES). www.coes.cl

³Se incluyen las correcciones propuestas por el equipo del investigador COES Vicente Espinoza.