Introducción al Análisis de Datos Panel

Estudio Longitudinal Social de Chile

Alejandro Plaza Reveco
Centro de Estudios en Conflicto y Cohesión Social
Enero, 2021

Contenidos

- Introducción general de ELSOC
- Disponibilidad de las Bases de datos
- Características generales respecto a su uso
- Características especiales
- Bases conceptuales de los datos panel
- Empezando a análizar ELSOC en R

Introducción general de ELSOC

Objetivos General y Específicos

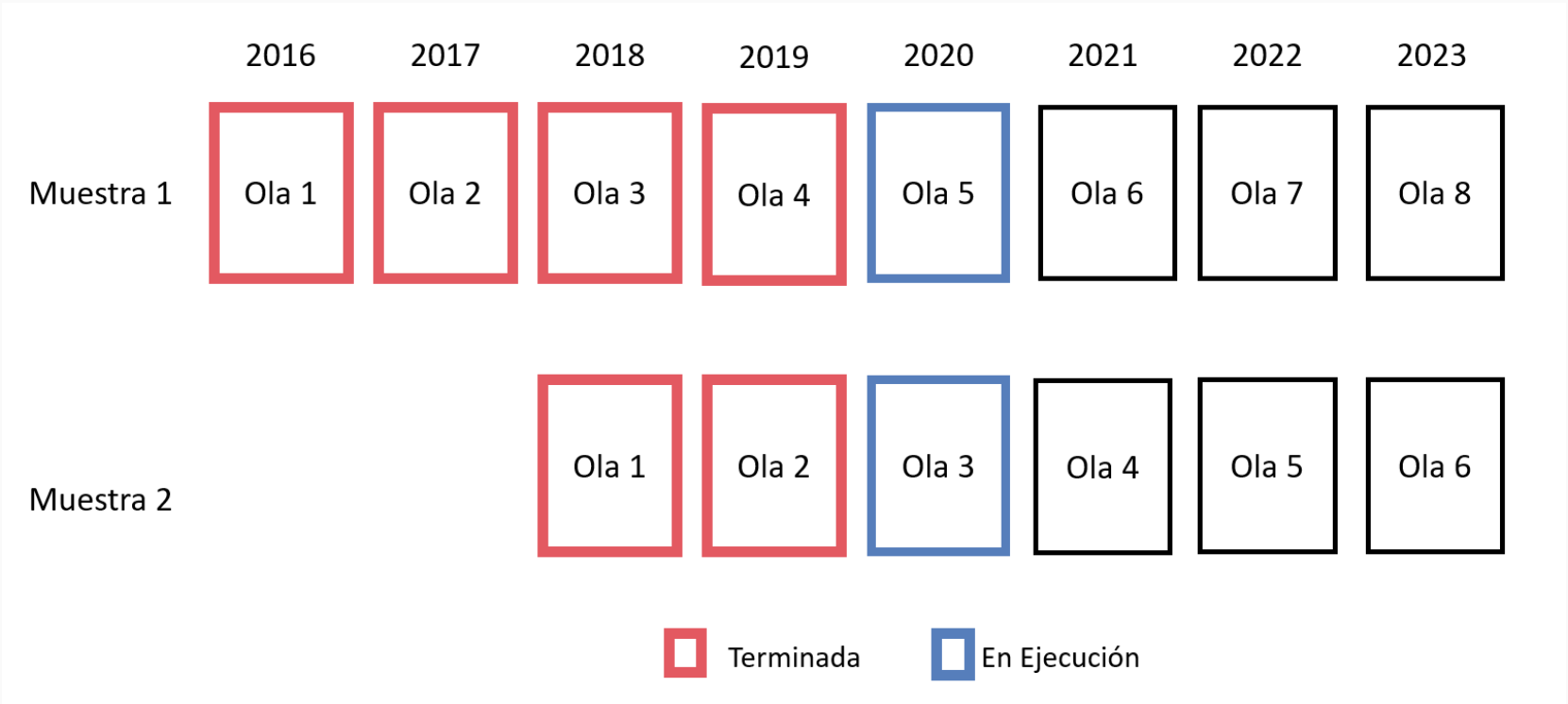
El Estudio Longitudinal Social de Chile (ELSOC) es una encuesta panel, representativa de la población nacional urbana, que analiza la estabilidad y cambio de las creencias, actitudes y percepciones que tenemos los chilenos y chilenas respecto de la convivencia y del conflicto en nuestra sociedad a lo largo del tiempo. Los principales temas de interés analítico aborados por este estudio corresponden a los módulos en los cuales se estructura:

1. Ciudadanía y Democracia
2. Redes sociales e interacciones inter-grupales.
3. Legitimidad y desigualdad social.
4. Conflicto social.
5. Dimensión barrial y territorial.
6. Salud y bienestar.
7. Caracterización Sociodemográfica.

Ficha Técnica

- Diseño: Estudio cuantitativo por medio de un cuestionario estructurado.
- Periodicidad: Anual.
- Diseño Longitudinal: panel repetido (misma encuesta se aplica a dos muestras independientes). Segunda muestra se implementó a partir del tercer año de medición (2018).
- Período de Aplicación: entre Julio y Noviembre de cada año. Cuarta medición se aplico entre el 21 de noviembre de 2019 y el 9 de marzo de 2020
- Instrumento: Cuestionario compuesto por preguntas cerradas de carácter simple y múltiple junto a algunas preguntas abiertas. Combina módulos de preguntas permanentes (medidas en todas las olas) y otras intercaladas entre olas.
- Cobertura Temática: Contiene siete módulos temáticos: Territorio, Redes y actitudes sociales, Ciudadanía y democracia, Desigualdad y legitimidad, Conflicto social, Salud y bienestar y Caracterización sociodemográfica.

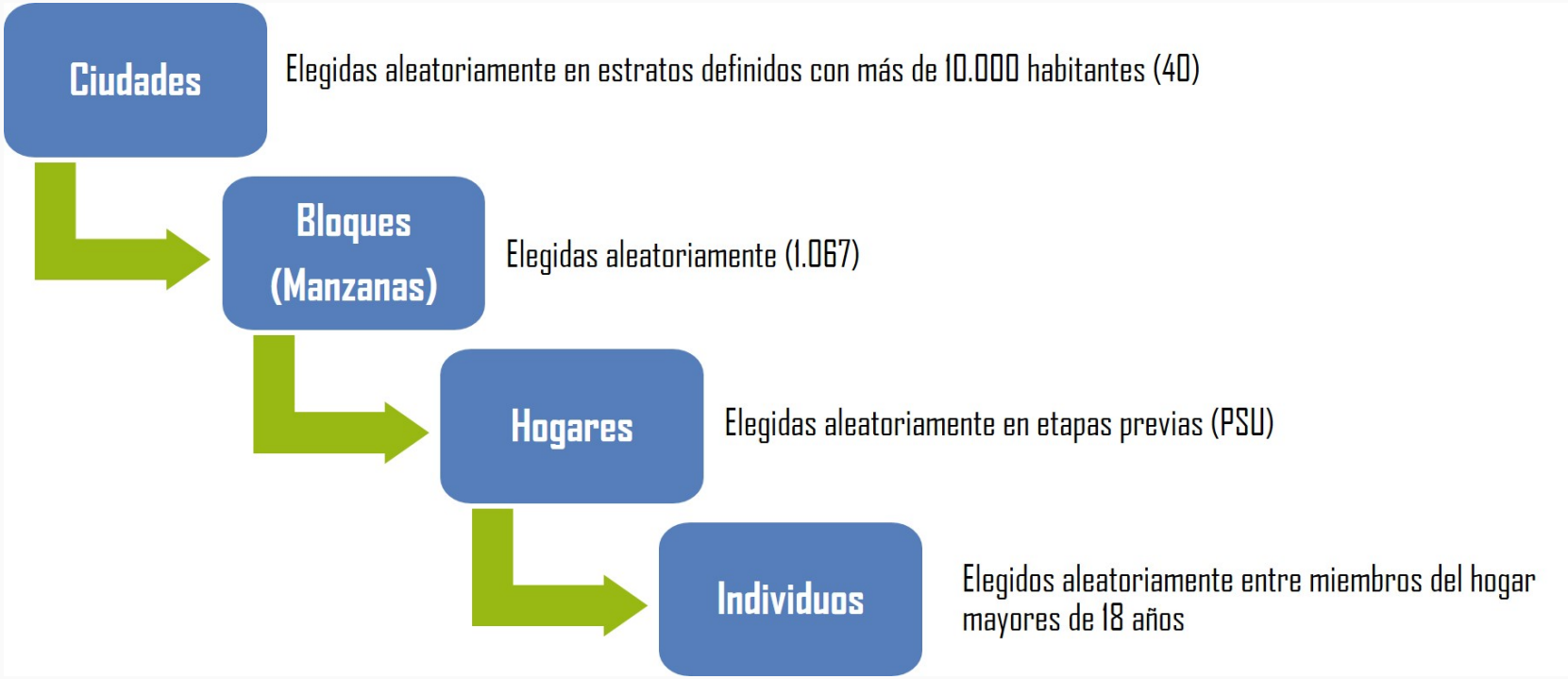
Dimension longitudinal del Diseño



Ficha Técnica

- Unidad de Análisis: Individuos.
- Población Objetivo: Hombres y mujeres de 18 a 75 años, residentes habituales de viviendas particulares ocupadas en zonas urbanas, localizadas en 40 ciudades (92 comunas, 13 regiones) del país.
- Marco Muestral: Marco de muestreo de manzanas del pre-censo 2011, trabajo elaborado por el Centro de Inteligencia Territorial (CIT) de la Universidad Adolfo Ibáñez.
- Diseño Muestral: Probabilístico, estratificado (por tamaño de ciudades), por conglomerados y multietápico.
- Unidades de Muestreo: Primero se eligen ciudades (UPM), luego manzanas (USM), y sub-bloques y viviendas (UTM). La unidad final de selección es la persona.

Etapas de Selección en el Diseño



Acceso a Base de Datos

Acceso a Base de Datos

ELSOC tiene un compromiso con los más altos estándares científicos en términos de producción y análisis de datos. Dentro de esta visión global, ELSOC se guía por las principales pautas de Transparencia y Apertura en la investigación científica.

Con este propósito, las bases de datos y documentación correspondientes a las cuatro primeras mediciones de ELSOC se encontrarán disponibles, de manera libre y gratuita, en un repositorio de datos, al cual se podrá acceder en el siguiente enlace .

Descripción de Atrición

Tamaño muestral

El diseño de ELSOC contempló entrevistar a 3.000 personas en su primera medición, reconociendo que año tras año, se reduciría el número de participantes, dado que algunos optarían voluntariamente por dejar de participar en el estudio y otras personas no podrían ser recontactadas o incluso algunas fallecerían.

Este fenómeno es conocido como atrición, y pueden tener efectos nocivos sobre la utilidad de los datos longitudinales. Para cada año se planifica obtener un número de entrevistados (Muestra Objetivo) considerando una proyección de la atrición definida al momento de diseñar el estudio.

Tamaño muestral

Medición	Muestra Objetivo	Muestra Lograda	Porcentaje de Logro
Muestra Original 2016	3000	2927	97.6%
Muestra Original 2017	2536	2473	97.5%
Muestra Original 2018	2131	2229	104.6%
Muestra Original 2019	1790	2153	120.3%
Muestra Refresco 2018	1500	1519	101.3%
Muestra Refresco 2019	1275	1264	99.1%

Atrición

Medición	Muestra Lograda	Porcentaje Recuperado	Atrición
Muestra Original 2016	2927	-	
Muestra Original 2017	2473	84.5%	15.5%
Muestra Original 2018	2229	90.1%	9.9%
Muestra Original 2019	2153	96.6%	3.4%
Muestra Refresco 2018	1519	-	
Muestra Refresco 2019	1264	83.2%	16.8%

Características generales respecto a su uso

Temas de investigación por módulos

Módulos	Temas
Ciudadanía (c)	Actitudes hacia la democracia, confianza institucional e interpersonal, comportamiento prosocial, participación e interés en política
Conflicto Social (f)	Justificación de la violencia, punitividad, fuerza y aversión al conflicto
Desigualdad y Legitimidad (d)	Estatus subjetivo, percepción y justificación de desigualdad, conflictos de clase, percepción de trato justo
Género (g)	Sexismo benevolente y hostil, identidad roles y normas de género
Redes y Actitudes (r)	Redes lejanas, redes cercanas, relaciones entre chilenos e inmigrantes
Salud y Bienestar (s)	Satisfacción vital, estado de salud, sintomatología depresiva, conductas saludables
Sociodemográfica (m)	Educación, clase social, ingresos, calidad de trabajo, previsión social

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.

Estructura longitudinal según módulos

Cada aplicación (por ola y por muestra) de ELSOC tiene en promedio 300 ítems. A continuación se presentan características de la estructura longitudinal de los ítems:

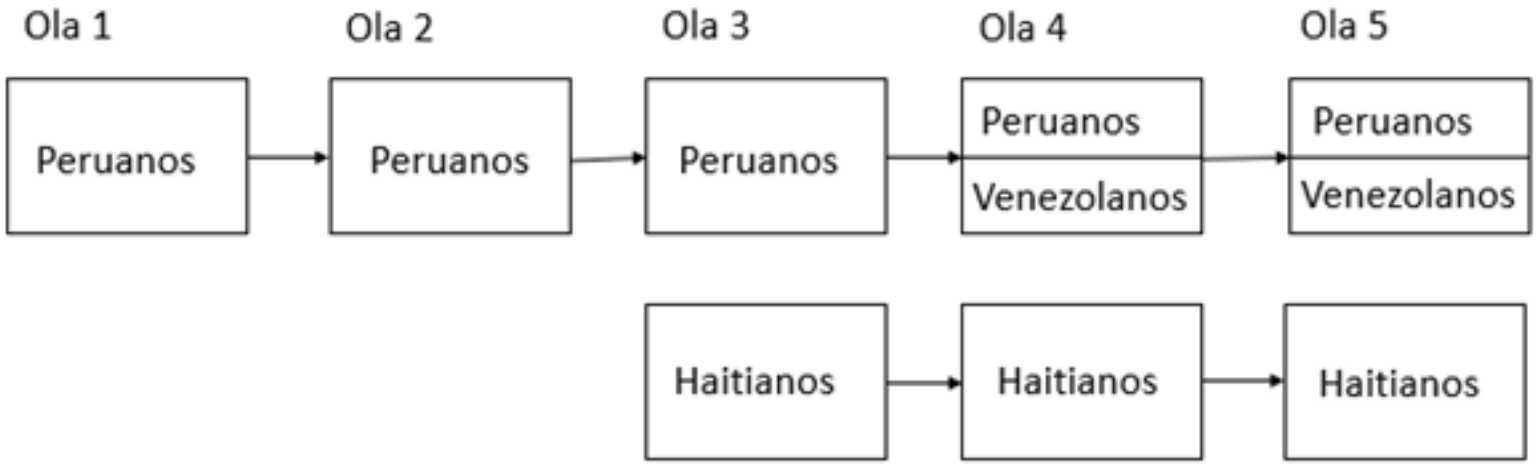
- Permanente: Se preguntan permanentemente a lo largo del estudio.
- Atributo fijo: Se preguntan una sóla vez a lo largo del estudio (p.e: Educación del padre)
- Intercalados: Se preguntan año por medio, en años pares o años impares.

Ambas muestras en promedio comparten alrededor del 65% de los ítems.

Características especiales

Características especiales de ELSOC

- Inclusión de dos baterías para estudiar el **entorno y la influencia social** desde una perspectiva de redes egocéntricas. El generador de posiciones (batería r01) y el generador de nombres (r13).
- Diseño para estudiar relaciones con tres poblaciones de inmigrantes.



- Posibilidad de vincular ELSOC con variables a nivel de manzana. *Base de datos y documentación será liberada durante el 2021.*

Bases conceptuales de los datos panel

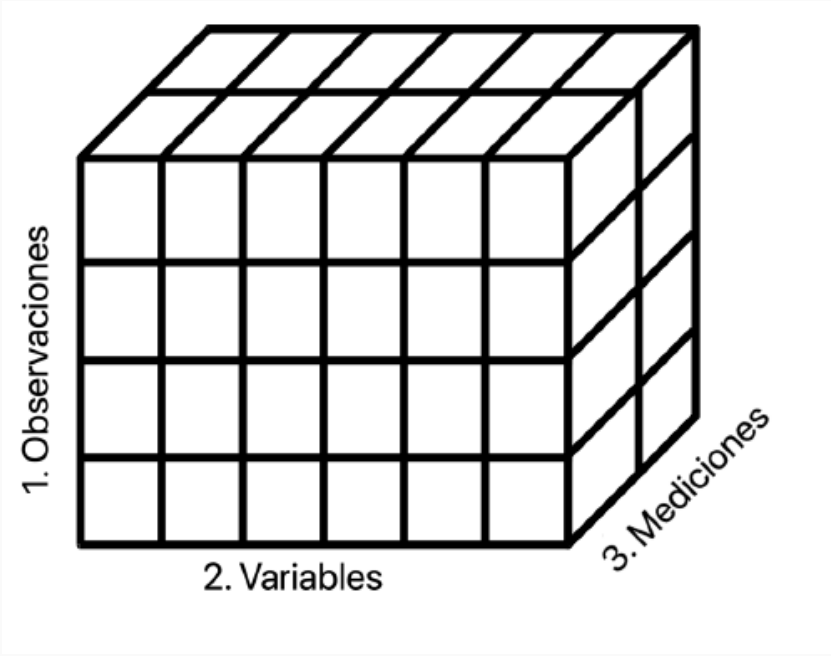
Estructura Básica de Datos Panel

Los datos de panel son tridimensionales:

Contienen **n** unidades $i = 1, 2, .. n$.

Contienen **V** variables $v = 1, 2, .. V$.

Contienen **T** Mediciones $t = 1, 2, .. T$.

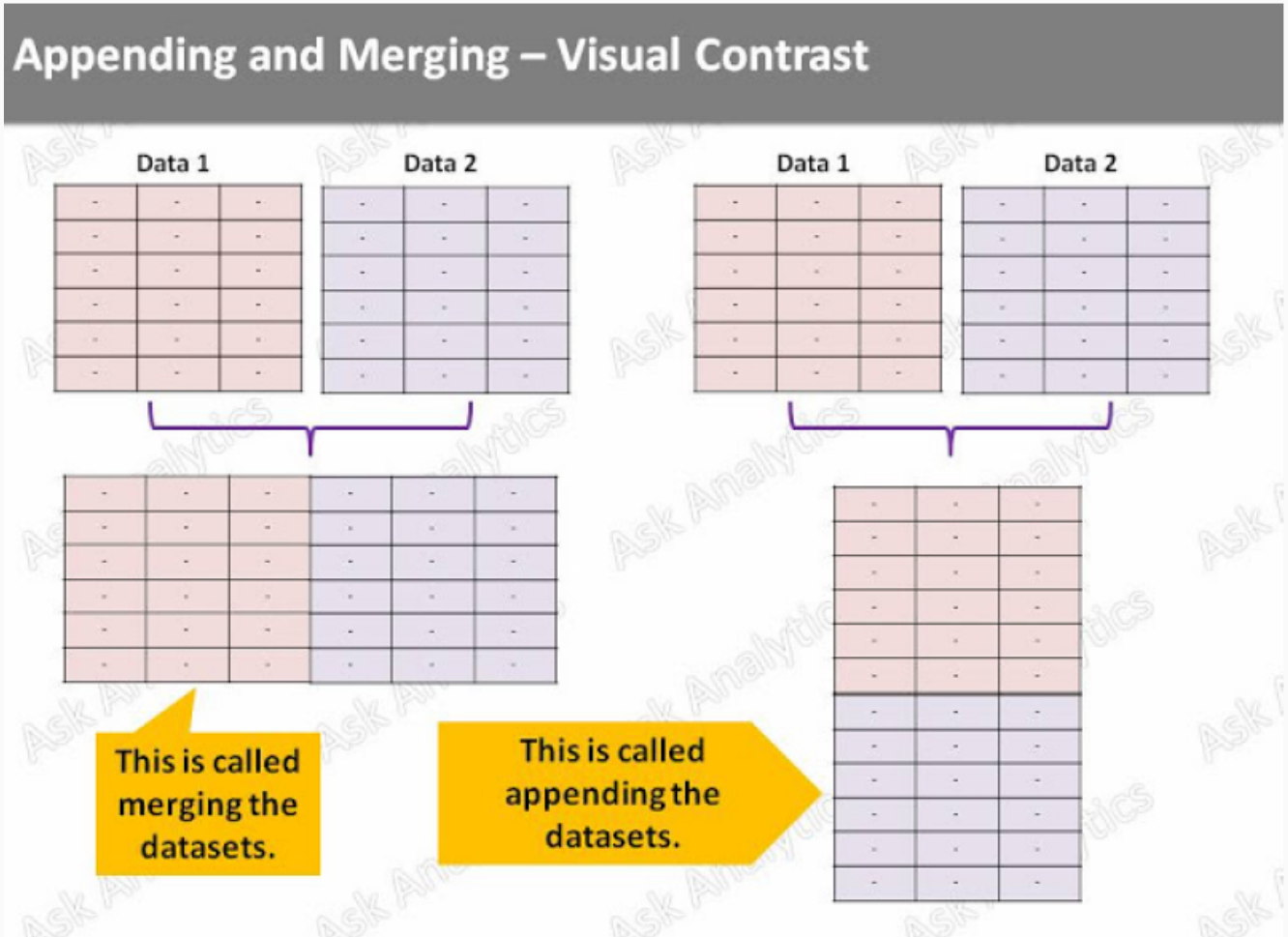


El desafío es transformar dicha estructura en un formato bidimensional

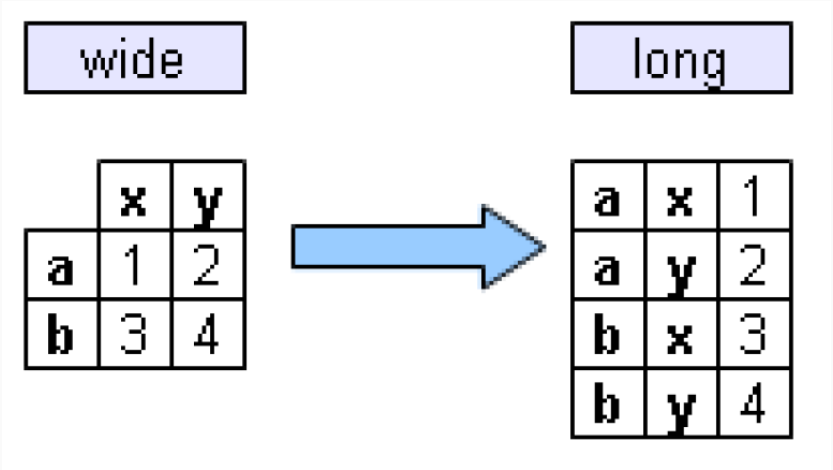
Formato Wide y Long

- En formato **Wide** cada observación es una fila (hay n filas). Las columnas representan variables-mediciones (T*V columnas)
- En formato **Long** cada medición es una fila (hay n*T filas). Las columnas representan variables (V columnas)
- Para incorporar el tiempo, en el formato wide se usan los nombres de las variables, mientras en long se incorpora una variable indicadora. Esto se hace porque se necesitan (a lo menos) dos variables para identi◎car las unidades.
- El formato long es más eficiente (no respuesta temporal reduce tamaño base de datos). El formato wide facilita análisis correlacional (cada medición es una variable) y re◎eja verdaderas dimensiones de los datos (n unidades observadas independientes entre sí).

Manipulación de Datos



Manipulación de Datos



Empezando a análizar ELSOC en R

Paso 0: Entorno de trabajo

Cargar paquetes y base de datos

```
#Instalar paquetes (de ser necesario)
#install.packages(c("car","lmtest","glmmML","panelView","plm","pglm","sjmisc","sjPlot

#Activar paquetes
library(car)
library(lmtest)
library(glmmML)
library(panelView)
library(plm)
library(pglm)
library(sjmisc)
library(sjPlot)
library(stargazer)
library(tidyverse)
library(panelr)

load("Directorio/ELSOC_Wide_2016_2019_v1.00_R.RData")
#cambiar nombre
els←elsoc_wide_2016_2019
```

Paso 1: Revisión de base de datos

Revisamos las dimensiones de la base de datos

```
dim(els)
```

```
## [1] 4447 1536
```

Revisamos los primeros 6 casos de las variables especiales

```
head(els[1:6])
```

```
##   idencuesta tipo_atricion tipo_caso version muestra cuestion_mig
## 1    1101011             1         0 1619100       1            2
## 2    1101012             1         0 1619100       1            2
## 3    1101013             1         0 1619100       1            1
## 4    1101021             1         0 1619100       1            1
## 5    1101022             5         0 1619100       1            1
## 6    1101023             1         0 1619100       1            2
```

Paso 2: Revisión de variables especiales

Con la variable tipo_atricion podemos observar la estructura longitudinal de los datos

```
sjmisc::frq(els$tipo_atricion)
```

```
##
## Participacion de encuestado en olas de ELSOC (x) <numeric>
## # total N=4447  valid N=4447  mean=5.06  sd=4.15
##
## Value |                    Label |    N | Raw % | Valid % | Cum. %
## ---------------------------------------------------------------------
##     1 |                2016-2019 | 1913 | 43.02 |   43.02 |  43.02
##     2 |                2016-2018 |  193 |  4.34 |    4.34 |  47.36
##     3 |       2016, 2017 y 2019 |  129 |  2.90 |    2.90 |  50.26
##     4 |         2016, 2018-2019 |   84 |  1.89 |    1.89 |  52.15
##     5 |                2016-2017 |  238 |  5.35 |    5.35 |  57.50
##     6 |             2016 y 2018 |   38 |  0.85 |    0.85 |  58.35
##     7 |             2016 y 2019 |   27 |  0.61 |    0.61 |  58.96
##     8 |               Solo 2016 |  305 |  6.86 |    6.86 |  65.82
##     9 | Solo 2018 (M. Original) |    1 |  0.02 |    0.02 |  65.84
##    10 |                2018-2019 | 1264 | 28.42 |   28.42 |  94.27
##    11 | Solo 2018 (M. Refresco) |  255 |  5.73 |    5.73 | 100.00
##  <NA> |                    <NA> |    0 |  0.00 |    <NA> |    <NA>
```

Paso 2: Revisión de variables especiales

Con la variable muestra podemos observar la distribución para M1 y M2

```
sjmisc::frq(els$muestra)
```

```
##
## Muestra de Participantes en ELSOC (x) <numeric>
## # total N=4447  valid N=4447  mean=1.34  sd=0.47
##
## Value |            Label |    N | Raw % | Valid % | Cum. %
## ------------------------------------------------------------
##     1 | Muestra Original | 2928 | 65.84 |   65.84 |  65.84
##     2 | Muestra Refresco | 1519 | 34.16 |   34.16 | 100.00
##  <NA> |             <NA> |    0 |  0.00 |    <NA> |   <NA>
```

Paso 2: Revisión de variables especiales

Con la variable tipo_caso podemos ver casos inconsistentes marcados por el equipo ELSOC (decisión del investigador qué hacer con estos casos)

```
sjmisc::frq(els$tipo_caso)
```

```
##
## Clasificacion de encuestado segun consistencia temporal (x) <numeric>
## # total N=4447  valid N=4447  mean=0.09  sd=0.32
##
## Value |                         Label |    N | Raw % | Valid % | Cum. %
## ------------------------------------------------------------------------
##     0 |            Casos Consistentes | 4055 | 91.19 |   91.19 |  91.19
##     1 | Casos Inconsistencias Menores |  362 |  8.14 |    8.14 |  99.33
##     2 | Casos Inconsistencias Mayores |   30 |  0.67 |    0.67 | 100.00
##  <NA> |                         <NA> |    0 |  0.00 |    <NA> |   <NA>
```

Paso 3: filtros y selección de variables

En primer lugar hacermos un filtro indicandonos que nos deje en la base de datos los casos completos del 2016 al 2019, y sin casos inconsistentes mayores

```
els←els %>% dplyr::filter(tipo_atricion==1 & tipo_caso !=2)
```

Luego podemos seleccionar las variables de interés.

```
el_wide←els %>% dplyr::select(idencuesta, #folio
                      m0_sexo_w01,m0_sexo_w02,m0_sexo_w03,m0_sexo_w04, #Sexo
                      m0_edad_w01,m0_edad_w02,m0_edad_w03,m0_edad_w04, #Edad
                      m01_w01,m01_w02,m01_w03,m01_w04, #Educación
                      c01_w01,c01_w02,c01_w03,c01_w04, #Satisfacción con la democraci
                      c15_w01,c15_w02,c15_w03,c15_w04, #Escala Izquierda Derecha
                      c20_w01,c20_w02,c20_w03,c20_w04, # Valoración de movimientos so
                      d01_01_w01,d01_01_w02,d01_01_w03,d01_01_w04) # Estatus social s
```

Paso 4: Wide-Long

Ocupando las funciones del paquete `panelr` podemos hacer la transformación de las bases de datos de manera relativamente sencilla

```
#wide a long
el_long ← long_panel(data = el_wide, prefix = "_w0", begin = 1, end = 4, label_locati
                                     id = "idencuesta", wave = "ola")
dim(el_long)
```

```
## [1] 7544    9
```

```
head(el_long[1:6], n=3)
```

```
## # Panel data:    3 x 6
## # entities:      idencuesta [1]
## # wave variable: ola [1, 2, 3 (3 waves)]
##   idencuesta   ola m0_sexo m0_edad   m01   c01
##   <fct>      <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1 1101011        1       2      64     2     1
## 2 1101011        2       2      65     2     2
## 3 1101011        3       2      66     3     1
```

Paso 5: Long-Wide

Se recomienda revisar en detalle la documentación de panelr

```
el_wide2 ← widen_panel(el_long, separator = "_w0",ignore.attributes = FALSE, varying=
dim(el_wide2)
```

```
## [1] 1886   57
```

```
head(el_wide2[1:6], n=3)
```

```
## # A tibble: 3 x 6
##   idencuesta m0_sexo_w01 m0_edad_w01 m01_w01 c01_w01 c15_w01
##   <fct>            <dbl>       <dbl>   <dbl>   <dbl>   <dbl>
## 1 1101011              2          64       2       1      11
## 2 1101012              2          60       4       1      12
## 3 1101013              2          26       4       1      12
```

Sacar los casos perdidos

```
el_long[el_long==-999 | el_long==-888] ← NA
el_wide[el_wide==-999 | el_wide==-888] ← NA
```

Paso 6: Estadística Descriptiva

```
sjmisc::frq(el_long$d01_01)
```

```
##
## Estatus Social Subjetivo: Ubica usted (2016) (x) <numeric>
## # total N=7544  valid N=7512  mean=4.36  sd=1.58
##
## Value |                  Label |    N | Raw % | Valid % | Cum. %
## ---------------------------------------------------------------
##  -999 | No Responde (no leer) |    0 |  0.00 |    0.00 |   0.00
##  -888 |     No Sabe (no leer) |    0 |  0.00 |    0.00 |   0.00
##     0 |   0 El nivel mas bajo |  147 |  1.95 |    1.96 |   1.96
##     1 |                     1 |  225 |  2.98 |    3.00 |   4.95
##     2 |                     2 |  428 |  5.67 |    5.70 |  10.65
##     3 |                     3 | 1075 | 14.25 |   14.31 |  24.96
##     4 |                     4 | 1693 | 22.44 |   22.54 |  47.50
##     5 |                     5 | 2787 | 36.94 |   37.10 |  84.60
##     6 |                     6 |  693 |  9.19 |    9.23 |  93.82
##     7 |                     7 |  259 |  3.43 |    3.45 |  97.27
##     8 |                     8 |  131 |  1.74 |    1.74 |  99.01
##     9 |                     9 |   16 |  0.21 |    0.21 |  99.23
##    10 |  10 El nivel mas alto |   58 |  0.77 |    0.77 | 100.00
##  <NA> |                  <NA> |   32 |  0.42 |    <NA> |   <NA>
```

Paso 6: Estadística Descriptiva

Medias por ola

```
#También es posible realizar el mismo cálculo por medio de la base de datos wide
el_wide %>% dplyr::summarise(estatus_w01 = mean(d01_01_w01, na.rm = TRUE),
                             estatus_w02 = mean(d01_01_w02, na.rm = TRUE),
                             estatus_w03 = mean(d01_01_w03, na.rm = TRUE),
                             estatus_w04 = mean(d01_01_w03, na.rm = TRUE))
```

```
##   estatus_w01 estatus_w02 estatus_w03 estatus_w04
## 1    4.294305    4.532198    4.431927    4.431927
```

Correlaciones entre olas

```
cor(el_wide$d01_01_w01, el_wide$d01_01_w02, use = "complete.obs")
```

```
## [1] 0.3158896
```

```
cor(el_wide$d01_01_w02, el_wide$d01_01_w03, use = "complete.obs")
```

```
## [1] 0.3627323
```

```
cor(el_wide$d01_01_w01, el_wide$d01_01_w03, use = "complete.obs")
```

Paso 7: Análisis de trayectorias

- Se busca examinar la pertenencia/presencia de un fenómeno o atributo a lo largo del tiempo. Se clasi◎can a los individuos según sus trayectorias de estabilidad o cambio en dichos atributos.
- A nivel agregado también son interesantes las tasas de permanencia, salida y reingreso. Muchas veces se usan técnicas básicas de **event-history analysis**.

```
#Recodificar de manera más simple estatus social subjetivo
el_wide$estatus1← car::recode(el_wide$d01_01_w01, "0:4='Baja'; 5='Media';6:10='Alta''
el_wide$estatus2← car::recode(el_wide$d01_01_w02, "0:4='Baja'; 5='Media';6:10='Alta''
el_wide$estatus3← car::recode(el_wide$d01_01_w03, "0:4='Baja'; 5='Media';6:10='Alta''
el_wide$estatus4← car::recode(el_wide$d01_01_w04, "0:4='Baja'; 5='Media';6:10='Alta''
```

```
#Generación de trayectorias
el_wide$estatus_tray←paste0(el_wide$estatus1,el_wide$estatus2,el_wide$estatus3,el_wi
```

Paso 7: Análisis de trayectorias

```
frq(el_wide$estatus_tray, sort.frq = "desc")
```

```
##
## x <character>
## # total N=1886  valid N=1886  mean=54.46  sd=24.23
##
## Value                 |   N | Raw % | Valid % | Cum. %
## ----------------------------------------------------
## BajaBajaBajaBaja      | 285 | 15.11 |   15.11 |  15.11
## MediaMediaMediaMedia  | 102 |  5.41 |    5.41 |  20.52
## BajaBajaMediaBaja     |  89 |  4.72 |    4.72 |  25.24
## BajaMediaBajaBaja     |  88 |  4.67 |    4.67 |  29.90
## MediaBajaBajaBaja     |  73 |  3.87 |    3.87 |  33.78
## BajaMediaMediaMedia   |  58 |  3.08 |    3.08 |  36.85
## BajaMediaMediaBaja    |  57 |  3.02 |    3.02 |  39.87
## BajaBajaMediaMedia    |  54 |  2.86 |    2.86 |  42.74
## MediaMediaMediaBaja   |  52 |  2.76 |    2.76 |  45.49
## BajaBajaBajaMedia     |  51 |  2.70 |    2.70 |  48.20
## MediaMediaBajaBaja    |  44 |  2.33 |    2.33 |  50.53
## AltaMediaMediaMedia   |  42 |  2.23 |    2.23 |  52.76
## AltaAltaAltaAlta      |  40 |  2.12 |    2.12 |  54.88
## MediaBajaMediaMedia   |  36 |  1.91 |    1.91 |  56.79
## BajaMediaBajaMedia    |  34 |  1.80 |    1.80 |  58.59
```