Ánalisis de las encuestas de hodares con R

Cápitulos: Procesamiento longitudinal de las encuestas rotativas

CEPAL - Unidad de Estadísticas Sociales

Tabla de contenidos I

Procesamiento longitudinal de las encuestas rotativas

Introducción

¿Por qué pensar en procesamiento longitudinal?

- Algunas oficinas requieren estadísticas de seguimiento continuo.
- Se aprovecha el esquema rotativo para generar información longitudinal.
- Requiere estructura de ponderación específica.

¿Qué es una encuesta longitudinal?

- Recolecta información sobre los mismos elementos en múltiples momentos.
- Contrasta con levantamientos transversales.
- Un ejemplo: esquema rotativo 4(1)0 permite seguimiento anual de 25%.

Introducción

Según Lynn (2009), una encuesta longitudinal observa los mismos elementos a lo largo del tiempo. Muchas encuestas rotativas, como las de empleo, se pueden convertir en longitudinales si se estructura adecuadamente el seguimiento.

Estimación del cambio y la varianza

- El foco está en estimar cambios entre periodos consecutivos.
- Es necesario calcular:
  - Varianza del periodo 1
  - Varianza del periodo 2
  - Correlación entre ambos
- Esos elementos afectan CV y tamaño de muestra.

"Uno de los retos metodológicos clave es que las muestras no son independientes. Para estimar cambios correctamente, debemos considerar la varianza en cada ronda y la correlación entre observaciones repetidas."

Análisis posibles con datos longitudinales

Caracterización de transiciones individuales

- Identificación de hogares/personas que cambian de estatus.
- Análisis de las características de quienes *entran o salen* de situaciones como la pobreza extrema, incluso sin cambios netos agregados.

Estabilidad e inestabilidad en el tiempo

- Seguimiento prolongado permite detectar *trayectorias persistentes* o *fluctuaciones*.
- Comprensión más profunda de los factores que **explican la permanencia en condiciones como la pobreza extrema**.

Duración, eventos e impactos

Caracterización de eventos y duración

- Estudio de la *duración de estados*: cuánto tiempo se permanece desempleado, inactivo, fuera del sistema educativo, etc.
- Posibilidad de construir *indicadores de duración y persistencia*.

Evaluación de impactos y relaciones causales

- Estimación del **efecto de intervenciones** o choques externos (ej. COVID-19) sobre fenómenos como la **desocupación**.

Diseño de paneles rotativos en encuestas de hogares

- En América Latina, varias encuestas de hogares incorporan esquemas de panel rotativo que permiten la observación repetida de una misma unidad de análisis.
- Este diseño busca:
  - Capturar dinámicas intra-hogar e interpersonales a lo largo del tiempo.
  - Generar estimaciones robustas sobre cambios de estado (e.g., ocupación inactividad).
- La dimensión longitudinal se configura sobre los hogares que **respondieron efectivamente en más de un periodo**, permitiendo análisis más allá de los cortes transversales.

Traslapes muestrales en un esquema 4(0)1

- Una encuesta con diseño **4(0)1** realiza cuatro observaciones trimestrales consecutivas por vivienda antes de su rotación definitiva.
- La rotación de paneles genera **traslapes sistemáticos** entre periodos consecutivos:
  - **T1 vs T2 → 75%** de hogares compartidos.
  - **T1 vs T3 → 50%**
  - **T1 vs T4 → 25%**
  - **T1 vs T5 → 0%** (panel completamente renovado).
- Esta propiedad escalonada permite construir **bases longitudinales de corto, mediano y largo plazo**, dependiendo del objetivo analítico.

Traslapes muestrales en un esquema 4(0)1

Tabla 1: *Rotación de páneles para un diseño 4(0)1.*

Trimestre	Panel 1	Panel 2	Panel 3	Panel 4
T1	a_1	b_1	c_1	d_1
T2	b_1	c_1	d_1	a_2
T3	c_1	d_1	a_2	b_2
T4	d_1	a_2	b_2	c_2
T5	a_2	b_2	c_2	d_2
T6	b_2	c_2	d_2	a_3

Construcción de esquemas longitudinales en paneles rotativos

La figura ilustra tres esquemas longitudinales que pueden derivarse del diseño rotativo:

1. **Bimestral (T1–T2):**
2. **Trimestral extendido (T1–T3):**
3. **Anual (T1–T4):**

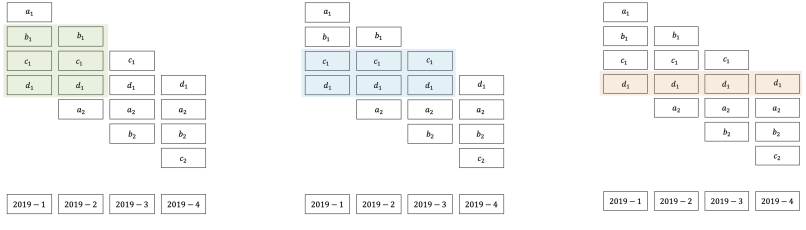


Figura 1: Tres escenarios longitudinales en un esquema rotativo 4(0)1.

Efectos del COVID-19 en el esquema rotativo – Año 2020

- El año 2020 representó un quiebre en la operatividad normal de los levantamientos debido a la emergencia sanitaria provocada por el COVID-19.
- Las *restricciones de movilidad* obligaron a:
  - Cambiar el modo de recolección a entrevistas **telefónicas**, reduciendo la cobertura y la tasa de respuesta efectiva.
  - *Repetir el diseño muestral* del primer trimestre en el tercer trimestre, lo que alteró la lógica original de rotación 4(0)1.

Esquema observado:

Año	Trimestre	Panel 1	Panel 2	Panel 3	Panel 4
2020	T1	a_1	b_1	c_1	d_1
	T2	b_1	c_1	d_1	a_2
	T3	b_1	c_1	d_1	a_2
	T4	c_1	d_1	a_2	b_2

Implicaciones en el traslape muestral:

- **T2 vs T3:** 100% de traslape.
- **T1 vs T3:** 75% de traslape.
- **T1 vs T4:** 50% de traslape.

Cargue de base de datos y librerias

```r
library(printr)     # Mejora la presentación de tablas en documentos RMarkdown
library(tidyverse)  # Conjunto de paquetes para manipulación y
                    # visualización de datos
library(tidyr)      # Manejo de estructuras anchas/largas
library(pROC)       # Curvas ROC
library(survey)     # Análisis de encuestas con diseño muestral complejo

# Carga de la base de datos a nivel de personas
base_personas <- readRDS(file.path(input, "base_personas.rds")) %>%
  ungroup()  # Elimina cualquier agrupamiento previo

# Creación de la base a nivel de hogares, extrayendo variables
# clave sin duplicados

base_hogares <- base_personas %>%
  distinct(upm, trimestre, id_hogar, fep)

# Visualización preliminar de la base de hogares
head(base_hogares, 10)
```

Cargue de base de datos y librerias

upm	trimestre	id_hogar	fep
1100100006	t1	262	19
1100100006	t2	262	19
1100100006	t1	265	16
1100100006	t2	265	16
1100100006	t1	277	16
1100100006	t2	277	16
1100100006	t1	288	19
1100100006	t2	288	19
1100100006	t1	289	30
1100100006	t2	289	30

Muestra de UPMs por trimestre

- Número de hogares por trimestre

```
base_hogares %>%   group_by(trimestre) %>%
  tally(name = "hogares")
```

trimestre	hogares
t1	5000
t2	4846

- Número de UPMs únicas por trimestre

```
base_hogares %>%   distinct(trimestre, upm) %>%
  group_by(trimestre) %>%   tally(name = "upm")
```

trimestre	upm
t1	500
t2	495

Número de UPMs en el traslape

Paso 1: identificar hogares que aparecen en ambos trimestres

```
# Identifica hogares que aparecen exactamente en dos trimestres
hogares_ambos <- base_hogares %>%
  group_by(id_hogar) %>%
  count() %>%  filter(n == 2) %>%  # Aparecen en dos trimestres
  pull(id_hogar)

# Extrae los pesos del trimestre 1 para los hogares que serán comparados
base_t1 <- base_hogares %>%
  filter(trimestre == "t1") %>%   select(id_hogar, fep_t1 = fep)

# Número total de hogares con traslape en dos trimestres
length(hogares_ambos)
```

```
[1] 3627
```

Fundamentos para la generación de bases longitudinales

- El análisis longitudinal permite observar transiciones individuales o de hogares entre estados (ocupación, pobreza, etc.) y **no es viable en todas las encuestas**, solo en aquellas con *esquemas rotativos* planificados.
- Es posible construir **tablas de transición** entre dos periodos a partir de observaciones empalmadas de la misma unidad.

Enfoque de estimación de flujos brutos Feinberg y Stasny (1983):

- Considera las diferencias entre pesos de muestreo en dos momentos como producto de la dinámica poblacional (ingresos y salidas del marco).
- Supongamos que un individuo fue clasificado como **empleado** tanto en el periodo $t-1$ como en el periodo t.

Ejemplo:

- Asuma que su peso muestral en el primer periodo fue $w_k^{t-1} = 300$ y en el segundo $w_k^t = 305$:

1. Se asignan **300** (el mínimo entre ambos pesos) a la celda *(Empleado $\rightarrow$ Empleado)*.
2. La diferencia **5** se atribuye a la celda *(Fuera $\rightarrow$ Empleado)*, asumiendo que provienen de entradas netas a la población de interés.

- Inversamente, si $w_k^{t-1} = 305$ y $w_k^t = 300$:

1. Se asignan **300** a la celda *(Empleado $\rightarrow$ Empleado)*.
2. La diferencia **5** se asigna a la celda *(Empleado $\rightarrow$ Fuera)*, representando salidas netas del marco poblacional.

Procedimiento para generación de pesos longitudinales**

Metodología Verma, Betti, y Ghellini (2006):

1. Pesos iniciales (transversales):

Basados en el diseño muestral, ajustados por:

- Probabilidad de selección por panel.
- No respuesta y cobertura.

2. Pesos longitudinales (dos periodos):

Ajustes aplicados:

- **Población longitudinal efectiva** (hogares presentes en ambos periodos).
- **Atrición muestral** (pérdida de casos por falta de seguimiento).
- **Calibración final** para alinear con totales poblacionales conocidos.

Consideraciones técnicas

- El tamaño de la base longitudinal **disminuye** a medida que se incrementa el número de periodos integrados (máximo 4 en diseño 4(0)1).
- **Agrega mediciones, pero reduce unidades únicas**: más observaciones por individuo, menos individuos distintos.

Consolidación de bases longitudinales

El primer paso en la generación de pesos longitudinales consiste en consolidar las bases de datos correspondientes a los periodos de interés. Esta integración produce bases de distintos tamaños según la cantidad de periodos combinados —por ejemplo, dos, tres o cuatro trimestres consecutivos.

- En términos generales:

1. Cuantos menos periodos se integren, mayor será el número de unidades observacionales disponibles.
2. En el caso específico de un esquema rotativo 4(0)1, no es posible consolidar cinco periodos consecutivos, ya que la rotación garantiza traslape máximo en solo cuatro trimestres consecutivos.

Implicaciones de la integración de paneles

La consolidación de paneles implica dos efectos clave:

- *Agregación de información:* Se repiten las observaciones de los mismos individuos en múltiples periodos, lo que enriquece el análisis longitudinal y permite estimar dinámicas de cambio.
- *Reducción de unidades observacionales:* Al requerir presencia continua en todos los periodos seleccionados, se pierde cobertura muestral frente al total de la muestra transversal en cada periodo.

Por tanto, se debe balancear el análisis de cambios individuales con la representatividad estadística que puede verse comprometida al aumentar la exigencia de continuidad en los datos.

Creación de los pesos longitudinales iniciales

El proceso inicia con la definición de los **periodos consecutivos** que se desean combinar.

En el caso del año 2020, se deben considerar los cambios operativos causados por la pandemia:

- A partir del segundo trimestre, los levantamientos dejaron de ser presenciales.
- Esto impactó tanto la tasa de respuesta como la comparabilidad entre trimestres.

Solo se incluirán las unidades muestrales que **respondieron en todos los periodos** seleccionados.

Determinación de pesos básicos

Los pesos longitudinales iniciales se derivan a partir de los **pesos básicos ajustados por cobertura** del primer periodo de combinación.

Ejemplos:

- Para combinar **T1 y T2 de 2020**, se parte de los pesos ajustados del **primer trimestre**.
- Para combinar **T2 y T3**, se toman los pesos del **segundo trimestre**.

Según LaRoche (2003), los pesos básicos se corrigen por la probabilidad de selección de paneles:

$$d_{1,k}^{bsico} = \frac{d_{1,k}}{\Pr(\text{selección de paneles})}$$

Consideraciones y validaciones

- En el ejemplo de **T1 y T2**, si tres paneles coinciden de cuatro posibles, entonces: $\Pr(\text{selección de paneles}) = \frac{3}{4}$.
- En **T2 y T3**, con la muestra replicada debido a la pandemia: $\Pr(\text{selección de paneles}) = \frac{4}{4} = 1$.

Validación clave:

La suma de pesos básicos debe aproximar el tamaño poblacional:

$$\sum_{s^{(1)}} d_{1,k}^{bsico} \approx N$$

Además, como propone la metodología de la *Survey of Labour and Income Dynamics* (Naud 2002; LaRoche 2003), este ajuste inicial incorpora la **probabilidad de traslape**.

Creación de los pesos longitudinales

Paso 2: crear variable dicotómica de respuesta en ambos trimestres

```
base_t1_t2 <- base_hogares %>%
  mutate(respboth = if_else(id_hogar %in% hogares_ambos, 1, 0)) %>%
  inner_join(base_t1, by = "id_hogar")

head(base_t1_t2 %>% filter(respboth == 1), 6)
```

upm	trimestre	id_hogar	fep	respboth	fep_t1
1100100006	t1	262	19	1	19
1100100006	t2	262	19	1	19
1100100006	t1	265	16	1	16
1100100006	t2	265	16	1	16
1100100006	t1	277	16	1	16
1100100006	t2	277	16	1	16

Creación de los pesos longitudinales

Paso 3: asignar peso básico (solo a hogares que respondieron en ambos)

```
prob_panel <- 2/4
base_t1_t2 <- base_t1_t2 %>%
  mutate( fep_long = ifelse(respboth == 1, fep_t1/prob_panel,0)
  )
sum(base_t1_t2$fep_long)
```

```
[1] 1181104
```

```
base_personas %>% group_by(trimestre) %>%
  summarise(pob_estimada = sum(fep))
```

trimestre	pob_estimada
t1	1239768
t2	1143222

Creación de los pesos longitudinales

```
head(base_t1_t2 %>%
       select(trimestre, id_hogar,
              respboth, fep_t1, fep_long), 10)
```

trimestre	id_hogar	respboth	fep_t1	fep_long
t1	262	1	19	38
t2	262	1	19	38
t1	265	1	16	32
t2	265	1	16	32
t1	277	1	16	32
t2	277	1	16	32
t1	288	1	19	38
t2	288	1	19	38
t1	289	1	30	60
t2	289	1	30	60

Ajuste por ausencia de respuesta

- Sobre los **pesos básicos**, se debe realizar un ajuste por **no respuesta**.
- Este ajuste debe basarse en:
  - Covariables auxiliares disponibles en el marco de muestreo.
  - Registros administrativos o rondas previas.
- Las unidades que no respondieron deben ser excluidas:

$$d_{1,k}^{bsico} = 0, \quad \forall k \notin s_r^{(1)}$$

donde $s_r^{(1)}$ representa el conjunto de respondientes efectivos del primer periodo.

Modelo de propensión de respuesta

- Se modela la **probabilidad de respuesta**:

$$\phi_{1,k} = \Pr(D_{1,k} = 1 \mid I_{1,k} = 1) = f(\mathbf{x}_1, \beta)$$

donde $D_{1,k}$ indica si la persona del hogar respondió la encuesta y $I_{1,k}$ si la persona pertenece a la muestra del primer periodo.

- Comúnmente se usa un **modelo logístico**:

$$\hat{\phi}_{1,k} = \frac{\exp(\mathbf{x}_1'\hat{\beta})}{1 + \exp(\mathbf{x}_1'\hat{\beta})}$$

- Requiere que $\mathbf{x}_1$ esté disponible para todos los seleccionados (respondan o no).

Ajuste final del peso inicial

- Se ajustan los pesos por el inverso de la probabilidad estimada:

$$d_{1,k}^{inicial} = \frac{d_{1,k}^{bsico}}{\hat{\phi}_{1,k}}$$

- En ausencia de información auxiliar: Usar la **tasa media de respuesta** como imputación para $\hat{\phi}_{1,k}$.
- Si la unidad es nueva en el panel: Imputar el **peso del hogar** al que pertenece.
- Verificar:
  - **Balanceo** entre respondientes y no respondientes.
  - **Soporte común** entre distribuciones de propensión (evitar extremos 0 y 1).

Creación de los pesos longitudinales

Paso 4: Identicar las personas en los hogares respondieron en ambos trimestre

Para construir una base de datos longitudinal consistente, es necesario identificar a las personas que respondieron en ambos trimestres (T1 y T2), dentro de los hogares previamente identificados como comunes.

```
base_personas_t1_t2 <- base_personas %>%
  filter(id_hogar %in% hogares_ambos) %>%
  mutate(id_llave = paste0(id_hogar, id_pers))

hogares_personas <- base_personas_t1_t2 %>%
  group_by(id_llave) %>%   count() %>%
  filter(n == 2) %>% pull(id_llave)

base_personas_t1_t2 <- base_personas_t1_t2 %>%
  mutate(respboth_per = if_else(id_llave %in% hogares_personas, 1, 0))

base_personas_t1_t2 <- base_personas_t1_t2 %>%
  inner_join(base_t1_t2 %>%
               select(id_hogar, fep_long, trimestre),
             by = c("id_hogar","trimestre"))
```

Ajuste por falta de respuesta de personas

Paso 5: estimar modelo logístico de probabilidad de respuesta

Se utiliza un modelo de regresión logística para estimar la probabilidad de que una persona haya respondido en ambos trimestres, en función de características observables.

```
modelo_logit <- glm(respboth_per ~ pobreza + area + etnia + sexo + edad,
                    data = base_personas_t1_t2,
                    family = binomial(link = "logit"))
```

Estimación de la probabilidad de respuesta

Paso 6: predecir la probabilidad de respuesta.

Se predicen las probabilidades individuales de respuesta, las cuales luego serán utilizadas para ajustar los pesos longitudinales.

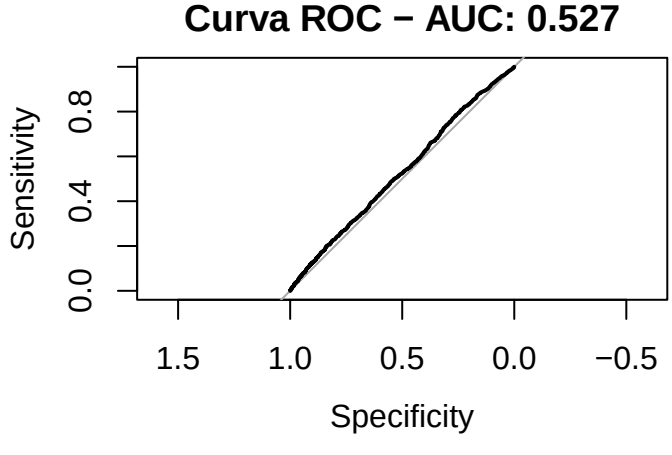
```
prob_resp = predict(modelo_logit, type = "response")
base_personas_t1_t2$prob_resp = prob_resp

roc_obj <- roc(base_personas_t1_t2$respboth_per,
               base_personas_t1_t2$prob_resp)
```

Evaluación del modelo: Curva ROC

La curva ROC permite verificar la capacidad predictiva del modelo de propensión. El área bajo la curva (AUC) debe acercarse a 1 para un buen ajuste.

```
plot(roc_obj, main = paste("Curva ROC - AUC:", round(auc(roc_obj), 3)))
```



Ajuste del peso longitudinal inicial

Paso 7: ajustar el peso longitudinal inicial usando el inverso de la probabilidad

El factor de expansión longitudinal se ajusta dividiendo el peso original entre la probabilidad de respuesta estimada.

```
base_personas_t1_t2 <- base_personas_t1_t2 %>%
  mutate(fep_aj = fep_long / prob_resp)
```

Visualización del ajuste de pesos

Se revisan los valores originales y ajustados para verificar consistencia y posibles extremos.

```
head(base_personas_t1_t2 %>% select(id_hogar, id_pers ,trimestre,
                                    fep_long, fep_aj))
```

id_hogar	id_pers	trimestre	fep_long	fep_aj
262	1	t1	38	39.94584
262	1	t2	38	39.94584
262	2	t1	38	39.98534
262	2	t2	38	39.98534
265	1	t1	32	33.74785
265	1	t2	32	33.74785

Creación de los pesos longitudinales finales

Definición de la población longitudinal

- La población longitudinal está compuesta por las personas que permanecen en la población objetivo entre dos periodos consecutivos.
- Ejemplo: en una encuesta del 2020, la población longitudinal del primer semestre incluye personas presentes tanto en el primer como en el segundo trimestre.

Creación de los pesos longitudinales finales

Exclusión por salida de la población

- Entre los dos periodos, pueden haber personas que salen de la población objetivo (muerte, migración, institucionalización, etc.).
- Estas personas no deben formar parte del análisis longitudinal, aunque sí están en la población objetivo del segundo periodo.

Creación de los pesos longitudinales finales

Muestra longitudinal y asignación de pesos

- La muestra longitudinal se define como:

$$s^{(2)} = s^1 \cap s^2$$

- El peso longitudinal inicial se transfiere directamente:

$$d_k^{\text{final}} = d_{1,k}^{\text{inicial}}$$

Se recomienda calibrar con totales auxiliares si están disponibles.

Ausencia de respuesta y atrición

- No todas las unidades responden en ambos periodos.
- Se identifican tres subconjuntos:
  a. Respondieron en t1 pero no en t2.
  b. No respondieron en t1 pero sí en t2.
  c. No respondieron en ninguno.
- Solo se mantienen las unidades que respondieron en ambos.

Asignación de peso nulo a los no respondientes

- Para unidades que no están en ambos periodos:

$$d_{2,k}^{inicial} = \begin{cases} d_{1,k}^{inicial}, & \text{si } k \in s_r^{(2)} \\ 0, & \text{si } k \notin s_r^{(2)} \end{cases}$$

- $s_r^{(2)}$ representa a las unidades respondientes en ambos periodos.

Ajuste por respuesta en segundo periodo

- Se estima la probabilidad de respuesta en el segundo periodo:

$$\phi_{2,k} = Pr(D_{2,k} = 1 \mid I_{2,k} = 1) = f(\mathbf{x}_2, \beta)$$

- Utilizar covariables auxiliares $\mathbf{x}_2$ disponibles para todos los seleccionados.

Peso longitudinal ajustado

- El ajuste final del peso longitudinal se hace mediante el inverso de la probabilidad de respuesta en el segundo periodo:

$$d_{2,k}^{longitudinal} = \frac{d_{2,k}^{inicial}}{\hat{\phi}_{2,k}}$$

- Este peso se utiliza para representar adecuadamente la población longitudinal.

Calibración de los pesos longitudinales

- Tras el ajuste por no respuesta, se recomienda calibrar los pesos.
- La calibración busca alinear los pesos con totales poblacionales conocidos:
  - Proyecciones nacionales.
  - Distribución por sexo, edad, región, área, etc.

Restricción de calibración

- Las variables auxiliares se denotan como $\mathbf{z}_k$.
- Se impone la siguiente condición:

$$\sum_{s_r^{(2)}} w_{2,k}^{calibrado} \mathbf{z}_k = \sum_{U} \mathbf{z}_k$$

- Donde $s_r^{(2)}$ representa a los respondientes en ambos periodos.

Forma funcional del peso calibrado

- Los pesos calibrados se definen como:

$$w_{2,k}^{calibrado} = g_k \cdot d_{2,k}^{longitudinal}$$

- g_k: factor de calibración, idealmente cercano a 1.
- $d_{2,k}^{longitudinal}$: peso longitudinal ajustado por no respuesta.

Consideraciones adicionales

- Las restricciones deben referirse a la población del **primer periodo**.
- La muestra longitudinal **no representa entradas posteriores**, solo a quienes estaban presentes al inicio.

Creación de los pesos longitudinales

Paso 8: Definir totales poblacionales conocidos

```
# Totales poblacionales por área y sexo
total_pob <- c(
  area1 = 1076892,
  area2 = 162876,
  sexoMujer = 615528
  # sexoHombre = 1076892 + 162876 - 615528
)
```

Creación de los pesos longitudinales

Paso 9: Crear diseño muestral con pesos ajustados

```
library(survey)
library(srvyr)
design_long <- svydesign(
  ids = ~upm,             # Asumiendo muestreo sin conglomerados
  strata = ~estrato,      # Variable de estrato
  data = base_personas_t1_t2,
  weights = ~fep_aj       # Peso ajustado por no respuesta
)
```

Calibración de los pesos longitudinales

Paso 10: Ajuste por calibración por área y sexo

```
design_cali <- calibrate(
  design = design_long,
  formula = ~0 + area + sexo,  # Efectos fijos (sin intercepto)
  population = total_pob
) %>%   as_survey()
summary(weights(design_long)); summary(weights(design_cali))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
6.281105	67.67086	128.0161	171.6724	210.8998	6489.454

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.086495	22.74004	42.54663	57.26676	70.14263	2183.58

Calibración de los pesos longitudinales

Validación de la calibración

```
options(survey.lonely.psu="adjust")
# Validación por área
design_cali %>%   group_by(area) %>%
  cascade(total = survey_total(), .fill = "Nacional")
```

area	total	total_se
1	1076892	0
2	162876	0
Nacional	1239768	0

Calibración de los pesos longitudinales

Validación de la calibración

```
# Validación por sexo
design_cali %>%   group_by(sexo) %>%
  cascade(total = survey_total(), .fill = "Nacional")
```

sexo	total	total_se
Hombre	624240	0
Mujer	615528	0
Nacional	1239768	0

Creación de los pesos longitudinales

Paso final: Asignar pesos calibrados a la base

```
base_personas_t1_t2$fex_cali <- weights(design_cali)
```

Análisis de flujos brutos y matrices de transición

Análisis de flujos brutos

¿Por qué usar datos longitudinales?

- Permiten conocer el **estado de una unidad** en distintos periodos.
- Posibilitan el análisis de **flujos brutos** entre categorías (por ejemplo, ocupación).
- Se pueden descomponer los **cambios netos** estimados en cortes transversales.

> *"La principal ventaja de los datos longitudinales es la posibilidad de estimar los flujos brutos." - Lynn (2009)*

Análisis de flujos brutos

Ejemplo conceptual

- Una encuesta rotativa permite estimar cuántas personas **cambiaron** de estado ocupacional.
- El análisis muestra si los "nuevos ocupados" en realidad ya lo eran en el trimestre anterior.

Ejemplo:

- A partir del seguimiento en dos trimestres, se puede saber si quienes estaban empleados, desempleados o inactivos, **mantuvieron o cambiaron** su estado.

Importancia del diseño y la no respuesta

1. ¿Qué puede sesgar los flujos brutos?
   - La **ausencia de respuesta** raramente es aleatoria.
   - Por ejemplo, los desempleados tienden a no responder más frecuentemente.
2. Riesgos:
   - Si se **ignora la ausencia de respuesta**, los flujos brutos estarán sesgados.
   - Si se corrige por no respuesta, pero se ignora el **diseño muestral**, también se introduce sesgo.

Matrices de transición

- Permiten estudiar **cambios de estado** entre dos periodos (ej. de ocupado a desempleado).
- Se construyen con unidades **respondientes en ambos periodos**.
- Fundamentales para entender dinámicas del mercado laboral.

Diseño complejo y flujos

- Las encuestas de hogares usan **muestreo complejo**: estratificado, multietápico, con probabilidades desiguales.
- Los **factores de expansión deben reflejar** este diseño para obtener estimaciones válidas.

Ausencia de respuesta y sesgos

- La no respuesta puede depender del **estado ocupacional** (ej. más no respuesta entre desempleados).
- Por eso, al estimar matrices de transición, se deben usar **pesos longitudinales ajustados y calibrados**.

Modelos de Markov

- Se utilizan para modelar los **cambios entre categorías** en dos periodos consecutivos.
- Asumen que los individuos pueden transitar entre **G estados mutuamente excluyentes**.

Transiciones en el tiempo

- El interés se centra en la **estimación de flujos brutos**: cambios reales entre estados.
- Solo se observa una muestra, pero el objetivo es **inferir la matriz poblacional de transición**.

Estructura de la matriz de transición

Sea X_{ij} la **cantidad de personas** que pasaron del estado i al j.

Tabla 14: Distribución no observable de los flujos brutos en una población.

Estado (T1/T2)	1	2	$\cdots$	j	$\cdots$	G
1	X_{11}	X_{12}	$\cdots$	X_{1j}	$\cdots$	X_{1G}
2	X_{21}	X_{22}	$\cdots$	X_{2j}	$\cdots$	X_{2G}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
i	X_{i1}	X_{i2}	$\cdots$	X_{ij}	$\cdots$	X_{iG}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
G	X_{G1}	X_{G2}	$\cdots$	X_{Gj}	$\cdots$	X_{GG}

Supuestos y limitaciones

- Se asume que la matriz de transición es **homogénea** para la población.
- La **ausencia de respuesta no ignorable** puede sesgar la estimación si no se corrige adecuadamente.
- Es fundamental **ajustar pesos y considerar el diseño muestral**.

Modelos de Markov con no respuesta

- Suponen que cada individuo pertenece a uno de G estados.
- Se modela el cambio de estado entre $t-1$ y t mediante:
  - η_i: probabilidad de estar en el estado i en $t-1$.
  - p_{ij}: probabilidad de transición de i a j.

Restricciones de los parámetros

- Suma total de probabilidades iniciales:

$$\sum_i \eta_i = 1$$

- Para cada fila i en la matriz de transición:

$$\sum_j p_{ij} = 1$$

Clasificación según respuesta

Tres grupos de individuos:

1. Respondieron en ambos periodos $\rightarrow$ matriz $G \times G$.
2. Solo respondieron en uno $\rightarrow$ complemento fila o columna.
3. No respondieron en ninguno $\rightarrow$ celda M.

Distribución observable con no respuesta

T1 / T2	Formal	Informal	Desocupado	Inactivo	Comp. fila
Formal	N_{11}	N_{12}	N_{13}	N_{14}	R_1
Informal	N_{21}	N_{22}	N_{23}	N_{24}	R_2
Desocupado	N_{31}	N_{32}	N_{33}	N_{34}	R_3
Inactivo	N_{41}	N_{42}	N_{43}	N_{44}	R_4
Comp. col.	C_1	C_2	C_3	C_4	M

Segunda etapa: proceso de respuesta

Cada individuo puede responder o no en cada periodo. Se introducen:

1. $\psi(i,j)$: probabilidad de respuesta en $t-1$ estando en celda ij.
2. $\rho_{RR}(i,j)$: probabilidad de seguir respondiendo en t.
3. $\rho_{MM}(i,j)$: probabilidad de seguir sin responder en t.

Modelos reducidos de no respuesta

Para estimar los parámetros de respuesta, se proponen modelos con supuestos específicos sobre la dependencia entre la respuesta y el estado de clasificación laboral.

Modelo A: Probabilidades constantes

- $\psi(i,j) = \psi$: misma probabilidad de respuesta en $t-1$ para todos.
- $\rho_{RR}(i,j) = \rho_{RR}$: misma probabilidad de mantenerse respondiente.
- $\rho_{MM}(i,j) = \rho_{MM}$: misma probabilidad de mantenerse ausente.
- **No depende del estado laboral.**

Modelos reducidos de no respuesta

Modelo B: Respuesta inicial según estado

- $\psi(i,j) = \psi(i)$: depende del estado en $t-1$.
- $\rho_{RR}(i,j) = \rho_{RR}$, $\rho_{MM}(i,j) = \rho_{MM}$: constantes.
- **Distingue entre formales, informales, etc. al inicio.**

Modelo C: Transiciones según estado en $t-1$

- $\psi(i,j) = \psi$: constante.
- $\rho_{RR}(i,j) = \rho_{RR}(i)$: depende del estado en $t-1$.
- $\rho_{MM}(i,j) = \rho_{MM}(i)$.
- **Transiciones diferenciadas por estado inicial.**

Modelos reducidos de no respuesta

Modelo D: Transiciones según estado en t

- $\psi(i,j) = \psi$: constante.
- $\rho_{RR}(i,j) = \rho_{RR}(j)$, $\rho_{MM}(i,j) = \rho_{MM}(j)$: dependen del estado final.
- **Transiciones diferenciadas por estado destino.**

Estimación con pseudo-verosimilitud

- Metodología de Feinberg y Stasny (1983) extendida por Gutiérrez (2014).
- Se usa **máxima pseudo-verosimilitud** bajo muestreo complejo.
- Implementado en el paquete `surf` de R Jacob (2020).

¡Gracias!

Email: andres.gutierrez@cepal.org

Referencias

Feinberg, Stephen, y Elizabeth Stasny. 1983. «Estimating monthly gross flows in labour force participation». *Survey Methodology* 9 (1): 77-102.

Gutiérrez, H. A. 2014. «The estimation of gross flows in complex surveys with random nonresponse». *Survey Methodology* 40 (2): 285-321.

Jacob, Guilherme. 2020. *surf: Survey-based Gross-Flow Estimation.*

LaRoche, Silvia. 2003. *Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics*. Statistics Canada.

Lynn, P. 2009. *Methodology of longitudinal surveys*. Wiley series en survey methodology. Wiley.

Naud, Jean-Francois. 2002. *Combined-panel longitudinal weighting - Survey of Labour and Income Dynamics*. Statistics Canada.

Verma, Vijay, Gianni Betti, y Giulio Ghellini. 2006. «Cross-sectional and longitudinal weighting in a rotational household panel: applications to EU-SILC», 36.