ENCUESTAS SOCIALES 2010 y 2018. PANEL DE EDUCACIÓN Y TRANSICIONES AL MERCADO LABORAL EN ANDALUCÍA:

GUÍA DE DESCARGA, EXPLOTACIÓN E INFORME TÉCNICO DE LOS MICRODATOS DEL PANEL

1. Introducción[1]

Las encuestas longitudinales ofrecen posibilidades de análisis que no se pueden conseguir con encuestas transversales o que no se pueden conseguir de una manera suficientemente fiable y precisa. De forma genérica estas encuestas facilitan:

a) La construcción de historias continuadas, mucho más extensas y completas que las construidas retrospectivamente en una sola entrevista

b) Recabar datos más precisos que los que se recogerían en una sola entrevista con recuerdos retrospectivos, los cuales podrían estar sujetos a importantes errores de memoria.

c) Obtener información sobre expectativas y alternativas no afectadas por sucesos y resultados posteriores, así como los resultados o sucesos efectivamente acaecidos posteriormente.

Las encuestas longitudinales también tienen algunas limitaciones en relación con otro tipo de encuestas. La recogida de datos en las encuestas longitudinales, relacionado con el hecho de que la muestra objeto de investigación es finalmente un panel de individuos fijo, puede sufrir del condicionamiento de los paneles y el desgaste de los paneles. El condicionamiento hace referencia a la posible alteración de respuestas para modificar o acortar el recorrido del cuestionario, cuando es similar al de la última edición y se mantiene el recuerdo. El segundo caso hace referencia al agotamiento del panel o desistimiento en la participación en la encuesta.

Los análisis a partir de las encuestas longitudinales también tienen limitaciones, en este sentido no son tan adecuadas como las encuestas transversales para aportar estimaciones transversales. En comparación con las estimaciones de una encuesta transversal, las estimaciones transversales de una encuesta longitudinal tienen más probabilidades de padecer un error de cobertura (porque la muestra se eligió hace mucho tiempo y puede no haber incluido las últimas modificaciones con respecto al interés de la población). Por otra parte, una muestra de una encuesta longitudinal puede adolecer de una baja tasa de respuesta en comparación con la de una encuesta transversal.

Es frecuente que instituciones que ofrecen educación o formación, como son las universidades u organismos gubernamentales responsables de la política en este ámbito, quieran evaluar los resultados de dicha educación o formación a un micro nivel (estudiantes). Estos resultados, de medio o largo plazo, requieren un seguimiento o recontacto con los estudiantes algún tiempo después de haber finalizado sus estudios.

[1]Esta introducción se fundameno en el texto Longitudinal surveys Methodology 2005, Peter Lynn

Los estudios longitudinales, a través de datos de panel, permiten realizar este seguimiento de los individuos que conforman el panel, en diversas oleadas.

En estos casos se suele conformar y contactar por primera vez con el panel en la época en que todavía son estudiantes. Las siguientes oleadas se realizan durante los años posteriores a la finalización de sus cursos. La información que se recoge en cada oleada pretende reconstruir las trayectorias o itinerarios, educativos y profesionales, seguidos por estos jóvenes entre oleadas. Así como las motivaciones y decisiones que conducen a estas trayectorias.

El IECA realizó en 2010 la "Encuesta Social 2010: Educación y hogares en Andalucía" con el fin de arrojar luz sobre los factores que influían en el rendimiento escolar del alumnado en Andalucía. Para poder realizar un análisis detallado, se seleccionaron dos poblaciones objeto de estudio, por un lado, la cohorte de niños/as nacidos/as en 1994 y, por otro, la cohorte de niños/as nacidos/as en 1998. La selección de estas dos cohortes de edad respondió al objetivo de obtener dos momentos temporales privilegiados desde los que observar las trayectorias escolares pasadas y las expectativas para el futuro, ya que si los niños/as seleccionados/as en la muestra no habían repetido ningún curso en su trayectoria escolar, estaban en el último curso de primaria (cohorte de 1998) y en el último curso de secundaria (cohorte de 1994), siendo éstos últimos especialmente relevantes ya que terminaban la etapa de enseñanza obligatoria.

La integración de fuentes diversas, como encuestas y registros administrativos, facilita una información más rica que ayuda a conocer y entender la realidad desde una perspectiva multidimensional, incorporando aspectos que no aumentan la carga de respuesta por parte del informante.

Tal enfoque es el que recoge la Comunicación de la Comisión al Parlamento Europeo y al Consejo sobre el método de elaboración de las estadísticas de la Unión Europea: una visión para la próxima década[2]. En esta comunicación se reconoce que las necesidades de información de los usuarios relacionadas con fenómenos multidimensionales complejos e interdependientes aumentan continuamente, por lo que para dar respuestas adecuadas a los retos actuales es preciso que los sistemas estadísticos evolucionen hacia un modelo de producción integrada. Es decir, es necesario transformar la arquitectura operativa estadística para acceder y combinar eficientemente la información procedente de distintas fuentes, encuestas y registros administrativos, con el fin de desarrollar, elaborar y difundir conjuntos de datos estadísticos integrados, coherentes y más ricos en términos de información.

Así pues, los datos de panel y la posibilidad de su integración con registros administrativos desempeñan un papel determinante para superar los análisis en los

[2]https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52009DC0404&from=ES

que solo se tienen en cuenta a las personas desde un punto de vista unidimensional (estudio de un sólo fenómeno) y efectuar la transición a una perspectiva multidimensional (personas que viven, trabajan, estudian, etc.), así como facilitar análisis logitudinales en algunas de estas dimensiones.

Con este enfoque en 2018 el IECA realiza la “Encuesta social 2018. Educación y transiciones al mercado laboral en Andalucía”, diseñada desde una perspectiva longitudinal como continuación de la Encuesta Social 2010 e incorporando algunos de los principios básicos de la producción estadística anteriormente sugeridos:

- Aprovechamiento y reutilización de la información, articulando el diseño de la encuesta de 2018 sobre la información de la encuesta realizada en 2010, teniendo como perspectiva una metodología longitudinal de análisis de cohortes.

- Producir información estadística que permita analizar la evolución de los fenómenos sociales complejos: La metodología que se plantea pretende hacer seguimiento en el tiempo de los fenómenos estudiados con el objetivo de realizar estudios longitudinales.

- Aprovechamiento de la información de registros administrativos o sistemas de información gestionados por otros organismos de la administración pública. Para el desarrollo de la operación ha sido fundamental complementar la información de la encuesta con registros administrativos.

2. Finalidad y objetivos del proyecto

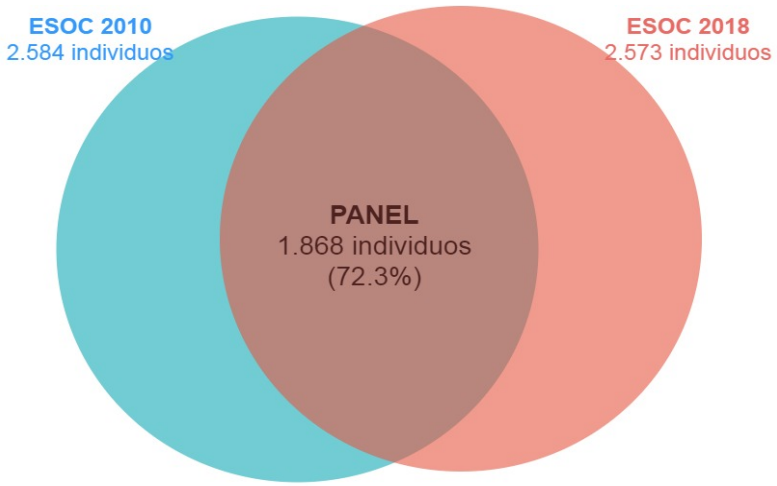
En el contexto descrito se enmarca el conjunto de microdatos “Panel de educación y transiciones al mercado laboral en Andalucía”. Este panel es el resultado de la integración de los datos procedentes de la “Encuesta Social 2010: Educación y hogares en Andalucía", la “Encuesta Social 2018: Educación y transiciones al mercado laboral en Andalucía” e información procedente de registros administrativos, en concreto la información contenida en Séneca (Plataforma para la gestión del sistema educativo andaluz)

Para ello, la “Encuesta Social 2018: Educación y transiciones al mercado laboral en Andalucía” (ESOC-2018) toma como punto de partida el diseño muestral de la “Encuesta Social 2010: Educación y hogares en Andalucía" (ESOC-2010), realizando un seguimiento a jóvenes de la cohorte de 1994. Es por tanto el segundo episodio de recogida de información en el seguimiento que se realiza a esta cohorte.

De manera específica, este proyecto pretende aportar información con perspectiva longitudinal que permita conocer, ocho años después, cómo se ha desarrollado la vida de los jóvenes componentes del panel, en distintos ámbitos, pudiendo contextualizar su situación presente con las características de sus familias de origen y con sus trayectorias escolares anteriores.

3. Fuentes de información

La información de referencia está constituida, por lo tanto, por los nacidos en 1994 que en 2010 residían en Andalucía, conformando el panel los individuos para los que se obtuvo un cuestionario completo tanto en la "Encuesta Social 2018: Educación y transiciones al mercado laboral en Andalucía" (ESOC-2018) como en la "Encuesta Social 2010: Educación y hogares en Andalucía" (ESOC-2010).



A la información individual obtenida mediante encuestación se le han incorporado una serie de características descriptivas de la trayectoria educativo-administrativa de los individuos, procedentes de la información registrada en Séneca.

De esta manera cada uno de los registros que conforma el fichero de microdatos del Panel proporciona un conjunto de variables referidas a la persona seleccionada, procedentes de las siguientes fuentes:

"Encuesta Social 2010: Educación y hogares en Andalucía"

Cuestionario del HOGAR: información procedente del cuestionario de HOGAR de la "Encuesta Social 2010: Educación y Hogares en Andalucía". Esta información fue facilitada por la tutora/tutor del menor seleccionado. De estas variables hay algunas que son sintéticas y que no aparecen en el cuestionario como tales: se construyen a partir de otras variables cuya información sí está recogida en el cuestionario de HOGAR. Han sido incluidas para facilitar los análisis a los usuarios. Estas variables pueden estar referidas:

- A la persona de referencia, que es la persona que aporta más ingresos en el hogar.

- A la persona informante, que es el tutor/a que proporciona la información de los cuestionarios de HOGAR y PADRES
- Al menor seleccionado, en el caso de la variable SEXO_EGO (Sexo del EGO)
- Al hogar, en variables como THOGAR (Tipo de hogar), STUDIOSC (Máximo nivel de formación de los tutores) o STUDIOSR (Tutor con mayor nivel de formación).

Cuestionario de PADRES: información procedente del cuestionario de PADRES de la "Encuesta Social 2010: Educación y Hogares en Andalucía". Esta información fue facilitada por la tutora/tutor del menor seleccionado.

Cuestionario de HIJOS: información procedente del cuestionario de HIJOS de la "Encuesta Social 2010: Educación y Hogares en Andalucía". Esta información fue facilitada por los menores seleccionados.

"Encuesta Social 2018: Educación y transiciones al mercado laboral en Andalucía"

Información procedente del cuestionario de la "Encuesta Social 2018: Educación y transiciones al mercado laboral en Andalucía". Esta información fue facilitada por los individuos seleccionados, ocho años después de haber respondido a la Encuesta Social 2010.

La información metodológica de ambas encuestas se puede consultar de forma más exhaustiva en los siguientes enlaces:

https://www.ieca.junta-andalucia.es/encsocial/2010/index.htm

https://www.ieca.junta-andalucia.es/encsocial/2018/index.htm

Séneca

Los registros administrativos constituyen el instrumento mediante el que la Administración deja constancia de los hechos e interacciones que, en el ejercicio de sus competencias, realiza con sus administrados. La tramitación de cualquier acto administrativo genera información que queda almacenada en el correspondiente registro y esta es susceptible de ser aprovechada para usos estadísticos.

Pero la explotación con fines estadísticos de la información contenida en los registros administrativos adolece de limitaciones (cambios en la normativa que regula el registro, su origen nunca es estadístico, las definiciones o las unidades empleadas pueden ser distintas de las estadísticas, etc.) que deben ser conocidas para aprovechar al máximo su potencial como fuente estadística.

La importancia del uso de los registros administrativos en el ámbito estadístico queda reflejada en el Código de Buenas Prácticas de las Estadísticas Europeas[3], promovido por Eurostat, que cita el uso de fuentes administrativas en varios de sus principios con el fin de limitar la carga de información solicitada a los informantes y mejorar el uso eficiente de los recursos disponibles.

En el desarrollo de este proyecto se ha contado con la colaboración de la Consejería de Educación de la Junta de Andalucía, la cual ha proporcionado el acceso a una fuente auxiliar complementaria a la encuesta: la plataforma para la gestión del sistema educativo andaluz, Séneca. En ella, entre otros datos, se recoge la información incorporada por los distintos agentes educativos de los centros públicos y concertados (profesores, personal administrativo y personal directivo). Contiene información relativa al seguimiento de los alumnos en distintos aspectos de su trayectoria educativa (rendimiento, comportamiento, pruebas competencias, etc.).

De esta fuente se han realizado extracciones en dos momentos temporales:

a) Extracción realizada en 2011. Estos datos se publican junto con los microdatos del cuestionario de hijos de la Encuesta Social 2010. La descripción de las variables incluidas, así como los aspectos a tener en cuenta para la explotación de estas variables se encuentran en el documento “Encuesta Social 2010: Educación y hogares en Andalucía. Guía de descarga y explotación de los microdatos”.

b) Extracción realizada en 2018. Estos datos se publican a través de variables resumen (sintéticas) a partir de la información correspondiente al último año en el que el individuo se ha matriculado en el último curso de estudios de PCPI, ESO, ESA, Bachillerato, FP de Grado Medio y FP de Grado Superior, así como los resultados obtenidos en dichos estudios.

El aprovechamiento de esta información se ha llevado a cabo a través del enlace entre los individuos de la muestra efectiva solapada de ambas encuestas y el sistema de información Séneca. Al tratarse de fuentes distintas, información administrativa y procedente de encuestas, se pueden registrar discrepancias, como se ha mencionado anteriormente, en este caso entre la información aportada por los informantes y la extraída a partir de los registros de Séneca. Cabe recordar que Séneca es un sistema de información complejo en el que participan numerosos agentes del sistema educativo y cuyo objetivo es el sistema educativo andaluz. A modo de ejemplo, aquellos estudios realizados fuera de Andalucía no quedarán registrados en Séneca, igualmente la información relativa a centros privados en la actualidad es casi completa pero su cobertura ha sido desigual en años precedentes. Las encuestas en estos casos recogerán la trayectoria formativa completa del informante, mientras que el microdato del Panel, incluyendo la información administrativa extraída de Séneca,

[3]https://ec.europa.eu/eurostat/documents/3859598/5922097/10425-ES-ES.PDF

presentará la información referida al ámbito de gestión de Séneca. Análogamente, la información procedente de encuestas presenta limitaciones y errores ajenos al muestreo como son la falta de veracidad de las respuestas de un número pequeño de informantes, errores de codificación u otros errores de procesamiento.

4. Elevación, calibración y falta de respuesta

El desgaste de la muestra (también denominado desgaste del panel) es una de las desventajas de este tipo de operaciones longitudinales basadas en muestras panel. La tasa de respuesta de cada oleada de una encuesta longitudinal puede ser tan válida como la de cualquier otra encuesta pero la proporción de unidades de muestra que han respondido todas oleadas puede ser bastante baja.

Por este motivo, la tasa de respuesta efectiva de un análisis longitudinal (para el cual se requieren datos de cada oleada) puede ser más baja que la tasa de respuesta obtenida normalmente en encuestas transversales.

Casi todas las encuestas longitudinales sufren de falta de respuesta o desgaste. Los casos que se pierden de un estudio (debido a que no haya sido posible localizar su nuevo teléfono y/o dirección, a que no se les pueda localizar aún disponiendo de un teléfono y/o dirección correctos o bien que no deseen continuar cooperando), tienen características diferentes de los que permanecen en el panel, es decir existe un sesgo en la falta de respuesta en el panel, y por lo tanto las inferencias sobre la población que se basan en la muestra observada no serán las mismas que las basadas en la muestra objetivo.

Una ventaja de los estudios longitudinales y la disponibilidad de información administrativa auxiliar es que podemos predecir la falta de contacto y de cooperación a partir de la información conocida sobre los miembros de la muestra como resultado del diseño longitudinal. De esta forma, se pueden construir estimadores que, basados en el diseño de la muestra, son corregidos a partir de las variables que más inciden en el fenómeno de la falta de respuesta. En este caso, la metodología seguida para la corrección de la falta de respuesta se basa en Ian Plewis (2007)[4] y se describe a continuación:

La probabilidad de respuesta de la vivienda i, en 2018 se puede descomponer de la siguiente forma:

$$P\left(V_{i,2018}\right)=P\left(V_{i,2010}\right)\cdot P\left(V_{i,2018}/V_{i,2010}\right)$$

Donde:

[4] Ian Plewis (2007) Non-Response in a Birth Cohort Study: The Case of the Millennium Cohort Study, International Journal of Social Research Methodology, 10:5, 325-334

- $P\left(V_{i,2010}\right)$ es la probabilidad de respuesta de la vivienda en 2010, aquella empleada en las estimaciones de la Encuesta Social 2010
- $P\left(V_{i,2018}/V_{i,2010}\right)$ es la probabilidad de respuesta de la vivienda en 2018 dado que respondieron en 2010

Para calcular la probabilidad de respuesta de la vivienda en 2018 dado que respondieron en 2010 ($P\left(V_{i,2018}/V_{i,2010}\right)$), se ha construido un modelo logit a partir de las siguientes variables auxiliares:

- Nacionalidad según la Base de Datos Longitudinal de Andalucía (BDLPA)
- Disponibilidad o no de teléfono procedente del Servicio Andaluz de Salud (SAS)
- Modificaciones de dirección desde 2010: Ha cambiado de dirección, no ha cambiado de dirección o no se encuentra en la Base de Datos Longitudinal de Andalucía en la actualidad (BDLPA)
- Nivel de estudios registrado en Séneca (Consejería de Educación)
- Tipología de de hogar, según Encuesta Social 2010
- Tipo de centro en el que se encontraba realizando sus estudios según Encuesta Social 2010

Los pesos utilizados en las estimaciones de la Encuesta Social 2010, corregidos por la falta de respuesta a partir del modelo logit utilizando la fórmula anterior, se han calibrado posteriormente mediante técnicas de reponderación. El objetivo de la utilización de estas técnicas es ajustar las estimaciones de la encuesta a la información procedente de fuentes externas. En esta encuesta se utilizó como fuente externa el Registro de Población de Andalucía y como variable auxiliar la población por provincias y sexo. El calibrado se ha realizado mediante el paquete Sampling del software R (Yves Tillé). Se ha optado por el método Logit truncado con límites 0,1 y 10.

Tras la elevación y la reponderación, la población resultante corresponde con la población de la cohorte objeto de estudio, los jóvenes nacidos en 1994 que en 2010 residían en Andalucía.

4.1 Modelización de la probabilidad de respuesta

Para la corrección de la falta de respuesta, se ha construido un modelo logit con el objeto de estimar la probabilidad de respuesta de la vivienda i en la Encuesta Social 2018 dado que dicha vivienda colaboró en la Encuesta Social 2010.

Dicho modelo se ha construido con los individuos que conforman la muestra efectiva de la Encuesta Social 2010, para los cuáles se observa si en 2018 respondieron (variable dependiente 1) o no respondieron a la encuesta (variable dependiente 0). El total de individuos utilizados para la modelización asciende a 2.578, dado que se excluyen los que fallecieron entre 2010 y 2018.

Las variables utilizadas son las siguientes:

1. **Nacionalidad**. Se observa que los individuos con nacionalidad de un país hispanohablante presentan una tasa de respuesta más elevada en 2018 que los individuos con nacionalidades de países no hispanohablantes
2. **Disponibilidad de teléfono**. Los individuos de los que se dispone un teléfono procedente de la base de datos del Servicio Andaluz de Salud (SAS), son más fácilmente localizables, lo que implica que la tasa de respuesta en 2018 es mayor.

1. **Cambio de dirección respecto a 2010**. Los individuos que siguen registrados en la misma dirección que en 2010 tienen una probabilidad más alta de responder en 2018 que aquéllos que han cambiado de dirección. Asimismo, se observan unas tasas de respuesta muy inferiores de aquellos individuos que no se encuentran actualmente en la Base de Datos Longitudinal de Andalucía (BDLPA)
2. **Estudios en Séneca**. Se observa que los individuos con niveles de estudio más elevados son más propensos a responder en 2018. Esta variable se ha obtenido del sistema Séneca de la Consejería de Educación, agrupando los niveles de estudio en 3 categorías: sin estudios reglados, la ESO o equivalentes o estudios posteriores a la ESO.
3. **Tipo de centro en el que estudiaba en 2010.** Se observa que los individuos que estudiaron en un centro público o concertado en 2010 (variable recogida en la Encuesta Social 2010), presentan unas tasas de respuesta más altas en 2018 que aquéllos que estudiaron en un centro privado no concertado.
4. **Tipo de hogar en 2010.** Se observa que los individuos que en 2010 vivían en un hogar con madre y padre son más propensos a responder en 2018 que los individuos que residían en otro tipo de hogar. Esta variable procede de la Encuesta Social 2010.

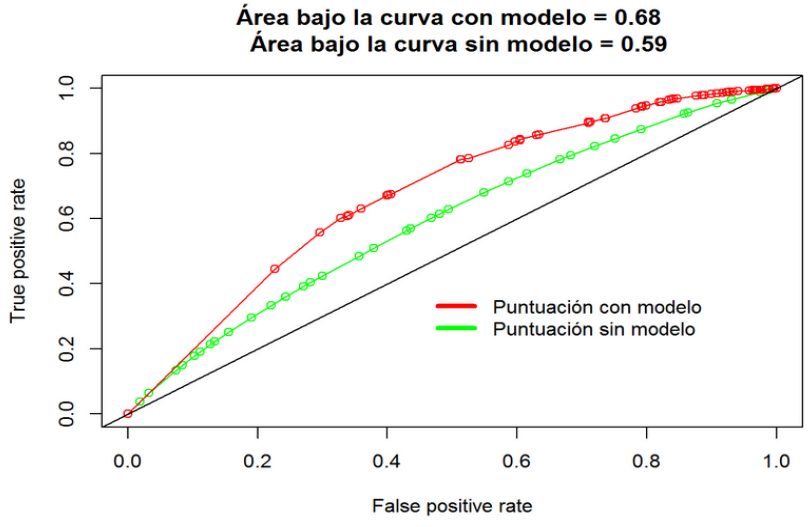
A continuación se muestran los coeficientes del modelo:

Predictor		Response rate (%)	Estimate	Std. Error	z value	Pr (>|z|)	
(Intercept)			0,4276	0,1206	3,545	0,000393	***
Nacionalidad	Hispanohablante	73,2%	0				
	No hispanohablante	48,8%	-0,5159	0,2491	-2,071	0,03837	*
Disponibilidad de teléfono	Con teléfono SAS	73,5%	0				
	Sin teléfono SAS	61,0%	-0,6901	0,1573	-4,386	1,15E-05	***
Cambio de dirección respecto a 2010	No cambia de dirección	76,0%	0				
	Cambia de dirección	67,3%	-0,2247	0,1051	-2,137	0,032571	*
	No está en BDLPA en 2018	42,3%	-1,0325	0,2284	-4,521	6,16E-06	***
Estudios en Séneca	Sin estudios reglados en SÉNECA	50,2%	0				
	ESO o equivalentes en SÉNECA	64,4%	0,419	0,1394	3,006	0,002651	**
	Estudios posteriores a ESO en SÉNECA	80,7%	1,219	0,1252	9,737	< 2E-16	***
Tipo de centro en el que estudiaba en 2010	Público o concertado	73,0%	0				
	Privado No concertado o NA	53,3%	-0,478	0,2531	-1,888	0,058988	.
Tipo de hogar en 2010	Hogar con madre y padre	74,5%	0				
	Otro tipo hogar	61,0%	-0,362	0,1237	-2,926	0,003437	**

--
Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Una alternativa para corregir la falta de respuesta empleada con frecuencia, es utilizar como proxy de la probabilidad de respuesta el cociente entre el tamaño muestral efectivo y el tamaño muestral teórico por estrato h (provincias y sexo en nuestro caso): $n(efectiva)_h / n(teórica)_h$

En el siguiente gráfico se muestra la comparativa del área bajo la curva ROC, que determina el poder predictivo de la falta de respuesta, utilizando el modelo logit anterior "puntuación con modelo" o bien utilizando el proxy $n(efectiva)_h / n(teórica)_h$ "puntuación sin modelo".



5. Contenido de la publicación: fichero de microdatos "Panel de educación y transiciones al mercado laboral en Andalucía"

5.1. Características generales y uso de los datos

A partir del panel de individuos que completaron las encuestas sociales 2010 y 2018 (a modo de oleadas), la incorporación de información procedente de registros administrativos y los factores de elevación correspondientes, se ha generado el fichero de microdatos para su difusión: un fichero de microdatos integrados preparados para suministrar a los usuarios la información necesaria que puedan requerir en sus respectivos proyectos de análisis e investigación.

Con la publicación de estos microdatos del panel, se pone a disposición de los usuarios un fichero con los individuos que respondieron tanto a la Encuesta Social 2010 como a la Encuesta Social 2018, junto con una selección de las principales variables recogidas en ambas encuestas. Esta submuestra de 1.868 individuos, permitirá realizar análisis longitudinales de la cohorte objeto de estudio, los jóvenes nacidos en 1994 que en 2010 residían en Andalucía.

De esta manera el fichero de microdatos del panel está conformado por la submuestra de 1.868 individuos que respondieron en ambas encuestas sociales (2010 y 2018), el panel supone un 72.3% de los que respondieron en 2010 (2.584) y un 72.6% de los que respondieron en 2018 (2.573).

Como se ha mencionado anteriormente el análisis a partir de la información longitudinal, o de panel, presenta ciertas limitaciones. En comparación con las estimaciones de una encuesta transversal, las estimaciones transversales de una encuesta longitudinal provienen de una muestra que puede adolecer de una baja tasa de respuesta en comparación con la de una encuesta transversal.

En este sentido, las estimaciones y análisis basados en la submuestra permiten analizar de forma conjunta variables recogidas en la Encuesta Social 2018 y en la Encuesta Social 2010, facilitando de esta forma la realización de análisis longitudinales de la cohorte objeto de estudio, los jóvenes nacidos en 1994 que en 2010 residían en Andalucía. Esta información hace posible la contextualización de las trayectorias educativas y laborales de estos jóvenes, en función de la situación de partida en el año 2010.

Por otra parte, cuando el objetivo sea realizar estimaciones y análisis de variables recogidas exclusivamente en la Encuesta Social 2018 (sin contextualizarlas con otras de 2010), conviene tener presente que, en general, las estimaciones realizadas con la submuestra (1.868 individuos), diferirán de las realizadas con la muestra completa

(2.573 individuos). De este modo, para cumplir con su objetivo, en este tipo de análisis debe utilizarse el archivo de microdatos completo ya publicado de dicha Encuesta Social, que aportará estimaciones más precisas soportadas por un mayor tamaño muestral.

Análogamente, si se desean realizar estimaciones y análisis de variables recogidas exclusivamente en la Encuesta Social 2010, debe utilizarse el archivo de microdatos completo de dicha encuesta (2.584 individuos), pues aportará estimaciones más precisas que los obtenidas a partir de la submuestra.

El fichero de microdatos publicado incluye la variable FEIR que informa sobre el factor de elevación del individuo (designado como EGO y que corresponde con las personas nacidas en la cohorte del 1994) y representa el número de elementos que hay en la población por cada elemento de la muestra.

De cara a la explotación de los microdatos, aquellas estimaciones de menos de 750 individuos de la población (obtenidas como suma de FEIR) están sujetas a una alta variabilidad, con errores de muestreo relativos superiores al 30% en la mayoría de los casos. Las estimaciones inferiores a este umbral, 750 individuos, se consideran por tanto poco fiables debido a su alta variabilidad. Se desaconseja el empleo de estas estimaciones, inferiores al umbral, en estudios y análisis estadísticos de la población objeto de estudio a partir de los microdatos del panel.

5.2. Descarga de microdatos

Los microdatos del panel contienen información a nivel de registro individual debidamente anonimizado. El código identificativo de cada registro es el código anonimizado del individuo ID.

El archivo de microdatos es un fichero de datos ASCII que no contiene cabecera y no contiene delimitadores de campos (comas, etiquetas, etc.), por tanto no puede ser interpretado automáticamente por el software. Por ello, facilitamos el código para importar los datos (ASCII) a R y SPSS. Las instrucciones para la importación se encuentran en los siguientes subapartados.

5.2.1 Descarga de los archivos de microdatos ASCII

Guardar y extraer el .zip, del enlace “Fichero de microdatos ASCII”. En la ubicación que indiquemos se generará un archivo de microdatos. Por ejemplo:

```
C:\ Microdatos_PANEL_ASCII\md_PANEL.dat
```

5.2.2 Para el programa SPSS

Paso 1: Situarse en el fichero “Código para importar SPSS” y pulsar el botón derecho del ratón. Pinchar en la opción “Guardar el enlace como”. Extraer del .zip el archivo de sintaxis de SPSS. Por ejemplo:

```
C:\Microdatos_PANEL_ASCII\PANEL_IMPORTAR_DATA.sps
```

Paso 2: En el documento de sintaxis que se ha guardado en el **Paso 1** se introduce en las líneas que comienzan por el comando FILE la ruta en la que se ha guardado el “Fichero de microdatos” y la ruta donde queremos que se guarde el archivo de salida. Es importante mantener las comillas tal y como aparece en el siguiente ejemplo.

Ejemplo para el fichero de importación:

```
FILE='C:\Microdatos_PANEL_ASCII\md_PANEL.dat'
```

Paso 3. Ejecutar las sintaxis.

5.2.3 Para el programa R

Paso 1: Situarse en el fichero “Código para importar R” y pulsar el botón derecho del ratón. Pinchar en la opción “Guardar el enlace como”. Extraer del .zip el programa de R. Por ejemplo:

```
C:\Microdatos_PANEL_ASCII\PANEL_IMPORTAR_DATA.R
```

Paso 2: Abrir el código de la tabla que se desea importar desde RStudio.

Paso 3: En el EDITOR introducir en la línea de código para definir la ruta donde hemos guardado el archivo de microdatos (**Apartado 2.1**). Es importante mantener las comillas tal y como aparece en el ejemplo, así como el uso de la barra “/” en lugar de la barra invertida “\”.

Ejemplo:

```
setwd("C:/Microdatos_PANEL_ASCII")
```

Paso 4: Ejecutar el script con la opción/botón "Source" o bien seleccionando todo el script y con la opción/botón “Run”.

Al finalizar el proceso de lectura del fichero de microdatos, aparecerán los mensajes:

Fin del proceso de lectura

Tiempo transcurrido: xx.xxx

Entonces se dispondrá, dentro de la sesión de RStudio, del fichero salida que contiene el archivo de microdatos en formato data.frame de R:

```
md_PANEL
```

5.2.4 Descarga de los archivos de microdatos CSV

Los microdatos también se encuentran disponibles en formato CSV. El separador es el punto y coma ";" y el carácter utilizado para el decimal es el punto ".". Para descargarlo, guardar y extraer el .zip, del enlace "Fichero de microdatos CSV". En la ubicación que indiquemos se generará un archivo de microdatos por cada tabla. Por ejemplo:

```
C:\Microdatos_PANEL_ASCII\md_PANEL.csv
```

5.3. Control del riesgo de revelación y preservación del secreto estadístico

La difusión de los resultados y microdatos de toda actividad estadística se enfrenta siempre a un conflicto entre el derecho a la información y el derecho a la intimidad. El dilema se concreta entre el nivel de detalle con el que difundir la información, como medida de la utilidad de la misma, y paralelamente la necesidad de preservar el secreto estadístico, o lo que es lo mismo, la identidad de los informantes. Dada la riqueza y detalle de la información contenida en el Panel, incrementada aún más tras su integración a nivel individual con registros administrativos, este conflicto es especialmente manifiesto. En este fichero de microdatos se ha optado por una estrategia de difusión que maximice el nivel de detalle de la información suministrada a los usuarios con objeto de que estos puedan realizar sus propios análisis, adecuados a sus objetivos e intereses. La principal limitación en cuanto al nivel de detalle y desagregación de las variables del fichero de microdatos publicado está motivada por el hecho de que el incremento del detalle amplifica proporcionalmente los riesgos de revelación. Por lo que, en último extremo, el máximo nivel detalle ofrecido es una solución de compromiso entre la utilidad e interés de la información y el riesgo de revelación a él asociado. Se han tomado dos medidas tendentes a garantizar la necesaria preservación del secreto estadístico en el fichero de microdatos del Panel:

– Los registros han sido anomimizados mediante la eliminación de todas las variables que sirven como identificadores directos de las unidades individuales.

– Se han considerado las combinaciones clave de variables que pudieran dar lugar a la identificación indirecta de las unidades individuales y se ha seguido una estrategia de recodificación global de las categorías de las mismas, con objeto de reducir el nivel de detalle informativo de dichas combinaciones y así minorar aceptablemente su potencia identificativa.