XVIII JIICAS Jornadas de Estadística de las Comunidades Autónomas
Cáceres del 20 al 23 de Octubre de 2010

Metodología de la Encuesta Social 2010: un caso de subpoblaciones no disjuntas

Muñoz Conde, María

mmacarena.munoz@juntadeandalucia.es

Hernández Moreno, Antonio

antonio.hernandez.moreno@juntadeandalucia.es

Pérez Morales, Germán

german.perez.morales.ext@juntadeandalucia.es

Trabado Lara, Inmaculada

inmaculada.trabado.ext@juntadeandalucia.es

Instituto de Estadística de Andalucía

Resumen:

*Dentro del marco metodológico que proporciona la **Encuesta Social 2010: Educación y Hogares en Andalucía**, cuyo objetivo principal es obtener información de contexto que contribuya a mejorar la calidad de la educación en la Comunidad Autónoma de Andalucía, un punto de verdadera reflexión de la Unidad de Muestreo del IEA se produjo a la hora de definir un diseño muestral que permitiera a la vez una cobertura apropiada de las subpoblaciones sobre las que se trabaja, una gestión eficaz de la muestra, unos tiempos de entrevista razonables y un uso de una plataforma tecnológica óptimo.*

Este trabajo versa sobre el análisis de posibles soluciones de una de las problemáticas particulares de esta encuesta, ya que se dispone de varias subpoblaciones objetivo no disjuntas. En concreto, la Encuesta Social 2010 tiene como subpoblaciones objetivo aquellas viviendas principales con hijos/as nacidos/as en 1994 y aquellas otras viviendas principales con hijos/as nacidos/as en 1998, de forma que existe un núcleo de viviendas en las que podemos encontrar casos de ambas subpoblaciones.

Una vez determinado el diseño muestral óptimo que se ajusta a las condiciones planteadas al inicio, se presentan justificadamente las estrategias que se han adoptado tanto para la resolución de problemas teóricos como prácticos.

Palabras clave: *muestreo, subpoblaciones no disjuntas, factores de elevación, educación, encuesta social.*

INTRODUCCIÓN

Las importantes transformaciones que la sociedad andaluza ha experimentado en los últimos años hacen necesario contar con instrumentos que permitan comparar esta realidad con las de otros espacios y realizar un seguimiento de los cambios.

El Instituto de Estadística de Andalucía ha impulsado para tal fin una serie de actividades estadísticas, entre las que se encuentra la Encuesta Social, incardinada en uno de los objetivos generales del Plan Estadístico de Andalucía 2007-2010: incidir en el conocimiento de los cambios sociales producidos en Andalucía.

Este modelo de encuesta está diseñado con el propósito de recoger información social de carácter específico en distintas ediciones de la misma, así como un grupo de variables sociales básicas que se han desarrollado en el seno de un grupo de trabajo de EUROSTAT (Oficina de Estadística de la Comunidad Europea).

La **"Encuesta Social 2010: Educación y hogares en Andalucía"** es una encuesta que se ciñe a este modelo y se encuadra dentro del objetivo estadístico específico de suministrar información sobre la educación en Andalucía. Este nuevo proyecto, que comienza a forjarse a mediados del año 2009, se enmarca en un contexto social en el que, tanto a nivel nacional como a nivel andaluz, se ha manifestado la intención desde los distintos poderes públicos de dar un paso firme hacia un cambio en el modelo productivo. Esta transformación supone dar un giro en el modelo económico que permita a sociedades como la andaluza pasar a convertirse en "sociedades del conocimiento". Pero el cambio en el modelo productivo pasa a su vez ineludiblemente por un cambio en el tejido social y más en particular en la mejora de la cualificación y del nivel de formación de los ciudadanos. En resumen, este cambio estructural pasa por una mejora en la calidad de la educación, para lo cual es necesario conocer las causas que inciden en factores como el absentismo, el fracaso escolar o los bajos resultados académicos.

El objetivo general de esta encuesta es encontrar los factores que influyen en el rendimiento escolar de los alumnos y alumnas que asisten a centros educativos del territorio andaluz. Para poder realizar un análisis más detallado de este fenómeno, se han seleccionado 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 responde 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 en la muestra no han repetido ningún curso en su trayectoria escolar, estarán en el último curso de primaria (los nacidos en 1998) y en el último curso de secundaria (los nacidos en 1994), siendo éstos últimos especialmente relevantes ya que terminarían la etapa de enseñanza obligatoria.

En el desarrollo de este nuevo proyecto se ha contado con la colaboración de la Consejería de Educación de la Junta de Andalucía, la cual nos ha proporcionado el acceso a una fuente auxiliar complementaria a la encuesta, que es el sistema de información SÉNECA. En él, se recoge la información incorporada por los distintos agentes educativos de los centros públicos y concertados (profesores, personal administrativo y directivo) relativa al seguimiento de los alumnos en distintos aspectos de su trayectoria educativa (rendimiento, comportamiento, competencias, etc.). El aprovechamiento de toda esta información se ha llevado a cabo a través del enlace entre la muestra de la encuesta y el sistema de información SÉNECA. Ésto nos ha permitido una reducción en la información solicitada en la encuesta, una mejora en la

calidad de la muestra y un análisis longitudinal de los datos, ya que dispondremos de la trayectoria escolar de los alumnos a través de los distintos cursos.

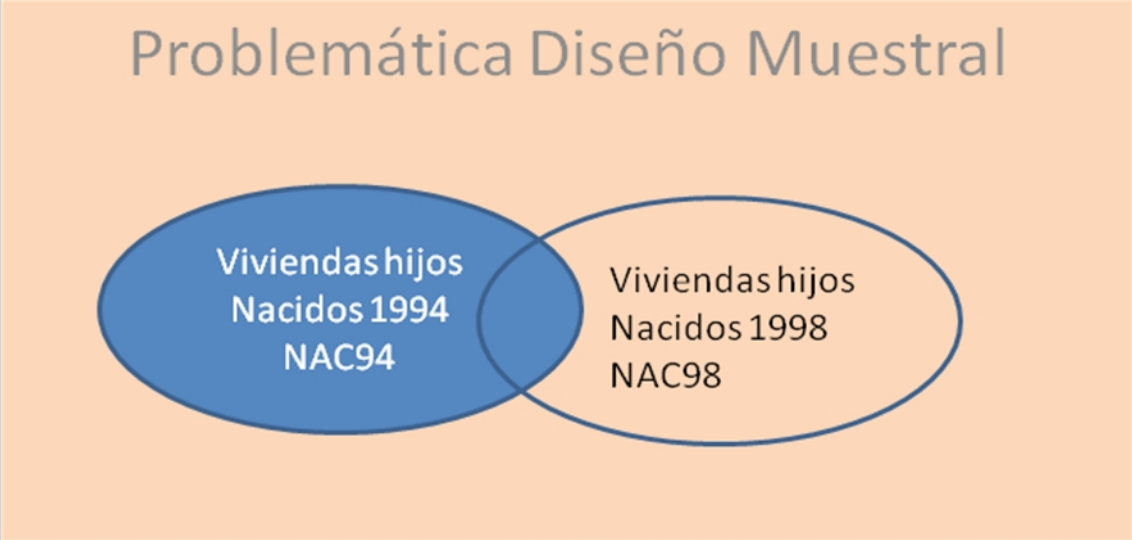
METODOLOGÍA

1. Población objeto de estudio. Intersección. Análisis de casos

Una vez que se fijó el objetivo general y la población objeto de estudio, nos enfrentamos a la problemática de la determinación del diseño muestral que mejor cubriese esta particular población de interés. A continuación se presentan las diversas soluciones debatidas por la Unidad de Encuestas, con sus ventajas e inconvenientes.

En la siguiente imagen se muestra la situación de la población objetivo a la que se enfrenta el diseño muestral:

Figura1.



En ella, se puede observar claramente que hay tres poblaciones objetivo, pues la intersección de ambas poblaciones surge como una subpoblación inherente en el estudio. El tratamiento de esta intersección, que no es más que la población formada por las viviendas de Andalucía que tienen algún menor nacido en 1994 y algún menor nacido en 1998, ha sido un campo a debate dentro de la Unidad Central de Encuestas del IEA, ya que existen diversas formas de análisis, cada una con sus ventajas e inconvenientes.

Estas viviendas que pertenecen a la intersección de nuestras dos poblaciones básicas, generan un gran número de preguntas que el diseño muestral debe responder: ¿a cuál de los dos menores se entrevista?, ¿es asumible o viable la entrevista a ambos menores?, ¿cómo se calculan los factores de elevación y los errores de muestreo en los diferentes casos?, ¿es muy dispersa la muestra de la intersección?, ¿cómo sería el tratamiento de una vivienda en la cual un niño/a ha sido seleccionado/a como titular en la primera población y otro como suplente en la segunda?, ¿cómo afecta al trabajo de campo la existencia de estas viviendas?, ¿es de interés el análisis individual de esta subpoblación para los objetivos de la encuesta?,....

Vamos a hacer una breve descripción de los distintos enfoques propuestos, viendo los puntos fuertes y débiles de cada uno de ellos e intentado responder a dichas preguntas.

Enfoques de muestreo en poblaciones con intersección

a) **Muestreo independiente en las 3 subpoblaciones**.

La primera posibilidad sería considerar las tres poblaciones de forma independiente, es decir, establecer un diseño muestral con tres subpoblaciones, NAC94 (viviendas con sólo hijos nacidos en 1994), NAC98 (viviendas con sólo hijos nacidos en 1998) y NAC9498 (viviendas con hijos nacidos en 1994 y 1998), por provincia, estrato y sección (8 viviendas por sección) en los dos primeros casos y sólo eliminando la etapa de las secciones para NAC9498, ya que representa un porcentaje pequeño de la población y muchas de las secciones estaban totalmente vacías.

En la intersección se entrevistaría a ambos hijos con sus distintos itinerarios según sea su subpoblación de procedencia.

Ventajas: Al ser tres subpoblaciones cerradas y tratadas de forma independiente las unas de las otras, el cálculo de los factores de elevación y el cálculo de los errores serían exactos y sencillos.

Desde el punto de vista del trabajo de campo, la gestión de las sustituciones se simplificarían, ya que cada población se toma de forma independiente a las demás.

Inconvenientes: Desde el punto de vista del trabajo de campo, la carga de información al informante se duplicaría, ya que en una misma vivienda habría dos unidades a entrevistar de la población objetivo lo que conllevaría una duplicidad de los cuestionarios con un itinerario distinto según la población de referencia. Este hecho puede implicar el riesgo de tener una elevada falta de respuesta, ya que unimos el incremento de la carga de información con la elevada duración de la entrevista.

Además, al tratar la intersección por separado y no poder utilizar en ella la sección como elemento del muestreo, provocaría un aumento en la dispersión de la muestra, ya que se incrementarían el número total de secciones a visitar, las distancias entre las viviendas, los desplazamientos y en definitiva, los costes de la encuesta.

b) **Eliminación de la intersección del muestreo**

Esta sería la opción ideal para lograr un diseño muestral sencillo y facilitar las tareas del trabajo de campo, sin embargo la eliminación de este subconjunto sería difícil de justificar metodológicamente, aunque sólo represente en cada subpoblación algo más de un 3 por ciento.

c) **Muestreo diferente en la intersección. Selección aleatoria del niño**

Una perspectiva distinta a las anteriores es realizar un muestreo en la intersección de forma que sólo se entreviste a un menor de la misma. Es decir, se debería elegir de la intersección cuál de los dos menores va a responder a la entrevista.

En concreto, en las poblaciones NAC94 y NAC98, se muestrearía por Provincia, Estrato, y Sección, mientras que en NAC9498 por Provincia, Estrato y seleccionando aleatoriamente al menor a entrevistar.

Ventajas: Evidentemente se presenta una reducción de carga de trabajo al entrevistado respecto a 1a). De nuevo los cálculos de los factores de elevación y de los errores de muestreo son exactos.

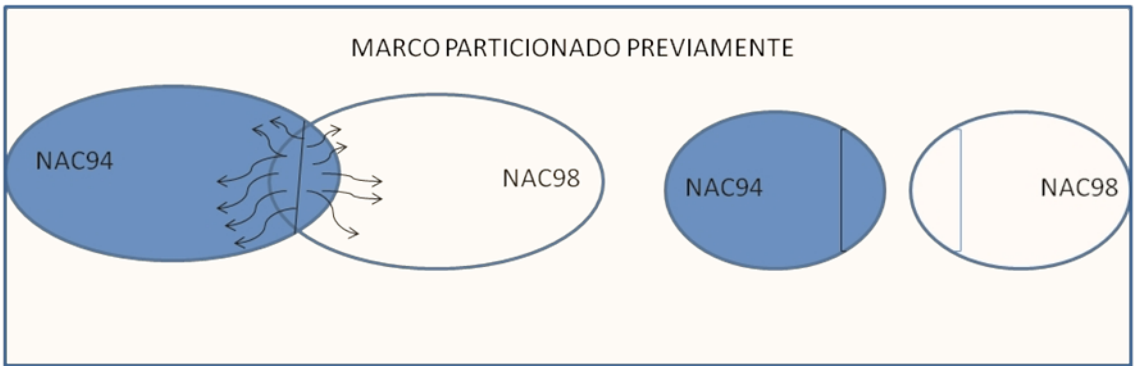
Inconvenientes: Se debería demostrar comportamiento semejante de las dos cohortes, provocando una pérdida de información de las viviendas de la intersección, ya que pueden tener un comportamiento distinto.

Respecto al trabajo de campo presenta complejidad en la intersección, tal y como vimos en el caso de 1a), debido a la dispersión muestral, lo que equivale a aumentar los costes de desplazamiento.

d) **Marco particionado a priori**

Una solución distinta a las anteriores es realizar una partición de las viviendas de la intersección a priori a nivel de marco, de forma que al muestrear ya sepamos a que subpoblación NAC94 o NAC98 pertenece la vivienda.

Figura 2.



Tal y como podemos ver en el gráfico trabajamos con dos poblaciones y un único plan de muestreo, seleccionando por Provincia, Estrato, Sección.

Ventajas: Con esta solución facilitamos enormemente el procedimiento de selección de viviendas en el muestreo. Además se produce una reducción de la carga de trabajo del informante. Obviamente el tratamiento del marco poblacional se hace más sencillo ya que se ha particionado a priori toda la intersección.

Inconvenientes: Para hacer la partición de la intersección a priori se debe hacer la suposición de independencia en el comportamiento de ambas cohortes 1994-1998.

Desde un punto de vista muestral, este enfoque complicaría el cálculo de los factores de elevación y de los errores de muestreo.

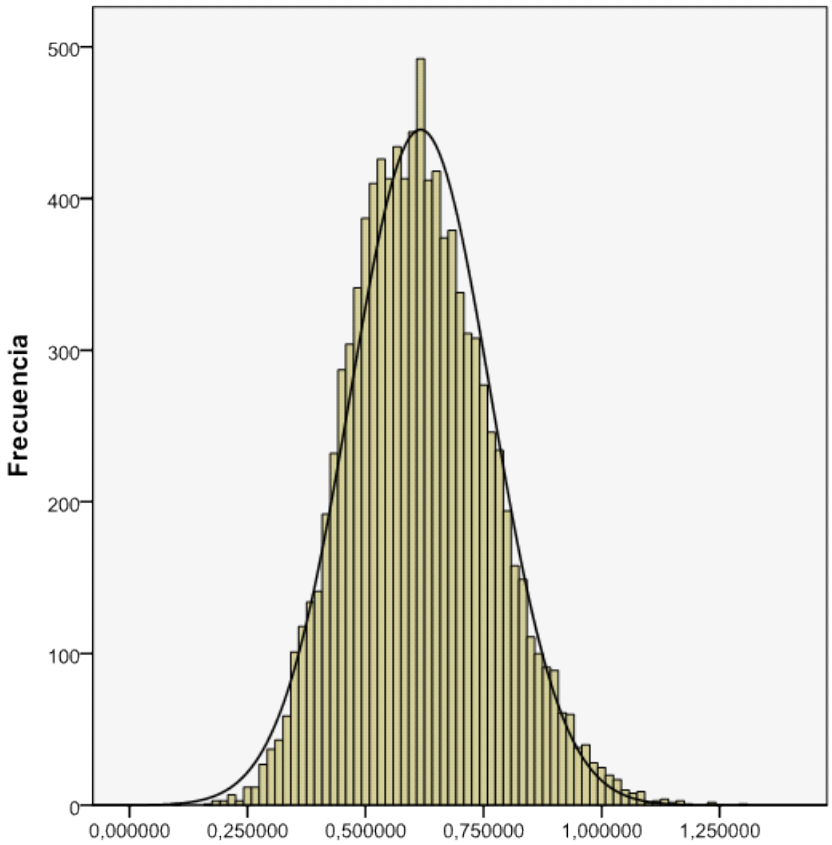
e) **Muestreo en las dos subpoblaciones sin tener en cuenta la intersección a priori**

Como último enfoque nos planteamos realizar un muestreo independiente en cada subpoblación sin tener en cuenta la intersección a priori. Para considerar ésta opción, resultaba imprescindible un análisis previo de la proporción de viviendas de la intersección en la muestra principal, por los inconvenientes que esto supone y que ya hemos comentado. En caso de que una vivienda de este subconjunto de la población sea seleccionada únicamente en una de las muestras, no existiría dificultad alguna, pues se entrevistaría sólo a un menor.

Para justificar esta opción se va a realizar un análisis cuyo objetivo fundamental es determinar si el rango de variación de esta variable está dentro de unos márgenes asumibles dentro de los recursos disponibles y del tiempo previsto para la ejecución de los trabajos de campo (aproximadamente 12 semanas). Esta tarea se resolvió mediante un experimento de simulación

basado en la extracción de 10.000 muestras aplicando un muestreo independiente en cada subpoblación, seleccionando por provincia, estrato y sección. Con este número de repeticiones se garantiza una distribución asintóticamente normal y una estimación por intervalos de confianza más precisa. El primero de los resultados que se presenta es el histograma de la proporción de viviendas de la intersección en la muestra principal. Se ha añadido la curva de la distribución normal estimada a partir de la media (0,616) y la desviación típica (0.149) muestral.

Gráfico 1. Distribución de la proporción de viviendas de la intersección en la muestra principal.



A partir del grafico se observa que la distribución es levemente asimétrica a la derecha (0,313) y que su grado de apuntamiento (0,087) es similar al de la distribución normal. Para determinar su rango de variación vamos a considerar el intervalo Media ± 3 * Desv. Típ., que en distribuciones normales nos aseguran que contienen el 99 por ciento la variabilidad de la distribución. De ahí, podemos decir que el rango o amplitud de esta proporción se encuentra entre el 0.169 y el 1.064 por ciento. Trasladando estos porcentajes a las unidades de muestreo, supondrían un intervalo de entre 10 y 64 viviendas aproximadamente, lo que resultó aceptable incluso para una hipotética extracción de la muestra próxima al extremo derecho del intervalo. A continuación se presentan los estadísticos descriptivos básicos de la simulación llevada a cabo.

Tabla 1. Estadísticos descriptivos de la proporción de viviendas de la intersección en la muestra principal.

	Estadístico	**Error típico**
Media	0,6164	0,00149
IC media 95% Límite inf.	0,6134	
IC media 95% Límite inf.	0,6193	
IC rango 99% Límite inf.	0,1686	
IC rango 99% Límite inf.	1,0641	
Media recortada al 5%	0,6134	
Mediana	0,6167	
Varianza	0,0223	
Desv. típ.	0,1492	
Mínimo	0,1000	
Máximo	1,3000	
Rango	1,2000	
Amplitud intercuartil	0,2000	
Asimetría	0,3137	0,02449
Curtosis	0,0869	0,04898

Ventajas: En cada subpoblación se aplica el mismo diseño muestral con la estratificación y etapas usuales. Tanto los factores de elevación (exactos) como los errores de muestreo se obtienen de un modo asequible. Se reduce considerablemente el número de entrevistas a dos menores en viviendas de la intersección.

Inconvenientes: Existe un grupo reducido de viviendas que tienen una mayor carga en la petición de información. La gestión de la muestra de viviendas principales y reservas es más compleja, porque una vivienda puede ser principal en una subpoblación y reserva en la otra.

Definitivamente, se decidió que esta última era la mejor de las opciones analizadas ponderando, por un lado, la aplicación de un diseño muestral clásico sin fisuras y, por otro, las dificultades que planteaba la entrevista doble en la intersección y su gestión durante el campo. En relación a este aspecto y teniendo en cuenta ese volumen de entrevistas, se decidió que para todas la viviendas principales se realizara la entrevista a ambos menores, aún cuando uno de ellos pudiera ser reserva en la muestra de su correspondiente subpoblación.

2. Marco de la encuesta

El marco de población de la ***Encuesta Social 2010: Educación y Hogares en Andalucía*** es el Registro de Población de Andalucía (RPA), que ya sirvió como marco para las anteriores Encuestas Sociales realizadas por el Instituto de Estadística de Andalucía. El RPA se obtiene de las tareas de revisión y depuración continuas del Padrón Municipal de Habitantes que realiza el Servicio de Estadísticas Demográficas y Sociales. Para esta ocasión, se contó con la información transversal que aporta el Registro para el segundo semestre de 2009. Este marco tiene la ventaja de facilitarnos una imagen bastante precisa de las subpoblaciones objetivo

mediante la selección de aquellos hogares con hijos nacidos en 1994 ó 1998. El concepto de hogar o vivienda, que usamos de forma indistinta, se construye a partir de la agrupación de los individuos que comparten el descriptor postal de su inscripción. Con el hogar como referencia, se facilita el acceso a la población objetivo definida independientemente del tipo de centro en el que el alumno esté matriculado (público o privado). Desde la perspectiva de la metodología de muestreo, la elección de este marco permite la aplicación del diseño usual en encuestas a hogares: muestreo multietápico, con estratificación de las unidades de primera etapa, secciones censales, y selección aleatoria de las viviendas en una segunda etapa.

También se analizó la oportunidad de emplear el Sistema de Información Séneca como marco de la encuesta, pero su uso se desaconsejo por varios motivos. El fundamental es que el registro no cuenta con los datos de los alumnos que estudian en centros privados sin concierto con la Consejería de Educación. En total se estima que un 5% de los alumnos andaluces están matriculados en estos centros. A esto hay que añadir que no contiene un detalle de ámbito territorial que nos permita plantear un diseño muestral por zonas para, entre otras cuestiones, evitar la dispersión de la muestra y reducir los costes de desplazamiento. Por último, la opción de considerar el centro como unidad de muestreo requería de un diseño muestral más complejo y dificultaba el reparto óptimo de la muestra en el territorio.

Por esta razón, se optó por la utilización del sistema de información SÉNECA como una fuente de información auxiliar complementaria al marco de población. La fusión de los registros de RPA y Séneca nos ha facilitado, de un elevado porcentaje de viviendas seleccionadas, un número de teléfono de contacto, los nombres de los tutores del menor seleccionado y una dirección postal para el contraste de la que proporciona RPA. Parte de esta información recabada también se empleará para la propia depuración del registro de población.

3. Objetivos de la encuesta

El objetivo principal de la “Encuesta Social 2010: Educación y Hogares en Andalucía” es encontrar los factores que influyen en el rendimiento escolar de los alumnos y alumnas que asisten a centros educativos del territorio andaluz. Principalmente buscaremos estos factores en el seno de las familias andaluzas partiendo de una hipótesis central: existen conexiones causales entre los modelos educativos de las familias y los resultados escolares de los niños/as.

Además de este objetivo principal, la encuesta trata otra serie de cuestiones para encontrar que otras causas, además de las familiares, influyen en el rendimiento escolar y en la actitud de los padres/tutores y los niños/as hacia la escuela. Cabría destacar las siguientes:

- Conocer la composición y la estructura social de los hogares de estudio de manera exhaustiva. Para ello se recoge información sobre las características sociales, económicas y demográficas de los hogares. Lo que permitirá establecer análisis posteriores teniendo en cuenta la influencia de aspectos como la clase social en los modelos familiares y en el rendimiento escolar.

- Estudiar la incidencia en la educación de los niños/as de otros agentes como los profesores, los compañeros (pares), etc.

- Se investigará la percepción y la opinión de los padres/tutores sobre el sistema educativo en general, con el objetivo de conocer sus valoraciones y opiniones al respecto.

- De igual modo se analizará la relación de los alumnos con la escuela. En este sentido, se obtendrá información sobre variables de contexto (la clase, el centro escolar, la relación con los otros compañeros, la percepción sobre la violencia escolar en su centro) y sobre variables individuales (expectativas, aspiraciones o rendimientos).

4. Ámbitos de la encuesta

a. Ámbito poblacional

La población objeto de investigación es la del conjunto de viviendas familiares principales con algún hijo nacido en 1994, el conjunto de viviendas familiares principales con algún hijo nacido en 1998, el conjunto de niños/as nacidos/as en 1994 y el conjunto de niños/as nacidos/as en 1998 que residen en dichas viviendas en la Comunidad Autónoma de Andalucía.
La Encuesta social 2010: Educación y Hogares en Andalucía va dirigida a la población con algún niño nacido en los años objeto del estudio, que reside en viviendas familiares principales, es decir, las utilizadas todo el año o la mayor parte de él como vivienda habitual o permanente. No se consideran, pues, ni los llamados hogares colectivos (hospitales, residencias, cuarteles, etc.) ni las viviendas secundarias o de temporada (de veraneo, fines de semana, etc.). Sí se incluyen, sin embargo, las familias que, formando un grupo independiente, residen en dichos establecimientos colectivos (por ejemplo, el director o el conserje del centro).

b. Ámbito territorial

La encuesta se extiende a todo el territorio de la Comunidad Autónoma de Andalucía.

c. Ámbito temporal

El trabajo de campo se ha realizado entre los meses de abril a julio de 2010.

5. Unidades de análisis

Se consideran en la encuesta cuatro unidades básicas de observación y análisis:
• Las viviendas familiares principales con algún hijo nacido en 1994
• Las viviendas familiares principales con algún hijo nacido en 1998
• Las personas nacidas en 1994.
• Las personas nacidas en 1998.

Se entiende por hogar el grupo de personas residentes en la misma vivienda familiar principal que consumen y/o comparten alimentos u otros bienes con cargo a un presupuesto común. En una misma vivienda familiar principal pueden residir uno o más hogares.

6. Diseño muestral

a. Tipo de muestreo. Unidades muestrales

Con el objetivo de dar información de los dominios de interés se ha utilizado para la extracción de la muestra un muestreo multietápico estratificado.
En primer lugar las viviendas familiares principales objeto del estudio se han estratificado en dos subpoblaciones. Una, formada por las viviendas familiares principales con hijos nacidos en 1994 y otra formada por las viviendas familiares principales con hijos nacidos en 1998. Para la extracción de la muestra en cada subpoblación se ha utilizado el mismo tipo de muestreo.

En cada subpoblación se aplicará un muestreo trietápico de conglomerados con estratificación de las unidades de primera etapa. Las unidades de muestreo de primera etapa están constituidas por las secciones censales, las de segunda etapa son las viviendas familiares principales y, por último, en la tercera etapa se selecciona una persona nacida en el año objeto del estudio que resida en esa vivienda. Las unidades de primera etapa se han dividido por provincias, una vez en la provincia se estratifica por tipo de hábitat considerándose los siguientes estratos:

1: Capitales de provincia
2: Municipios mayor de 100.000 hab.
3: 50.000-100.000
4: 20.000-50.000
5: 10.000-20.000
6: Menos de 10.000 hab.

Respecto a la aplicación del diseño muestral, al tratarse de poblaciones reducidas, las unidades de primera etapa, las secciones censales, no alcanzaban en todos los casos un número de viviendas suficiente para el muestreo de viviendas, por lo que fue necesario en algunos casos realizar un agrupamiento de ellas. Esta tarea nos debe permitir, en la medida de lo posible, que el número de viviendas sea suficiente para facilitar la afijación y que la dispersión territorial no se vea incrementada en exceso.

b. Tamaño de la muestra

El tamaño de muestra para esta encuesta se ha fijado en 6.000 viviendas, donde 3098 (51,6 por ciento) corresponden a viviendas con niños nacidos en 1994 y 2902 (48,4 por ciento) a niños nacidos en 1998.

c. Afijación

El reparto o afijación de las viviendas y secciones de la muestra en cada población ha sido el siguiente:

- Entre las subpoblaciones se ha realizado un reparto proporcional según el tamaño de cada una.
- Entre las provincias se ha utilizado una afijación de compromiso entre la uniforme y la proporcional al tamaño de cada provincia.
- Dentro de la provincia, la afijación de secciones por estratos es proporcional al tamaño de los mismos y en las secciones se extrae una muestra de tamaño variable (6 ó 8 viviendas principales) en función del número de viviendas disponibles.

La distribución de la muestra efectiva de viviendas y personas recogida en cada una de las provincias se presenta a continuación:

Tabla 2. Distribución de la muestra teórica y efectiva de viviendas

	Muestra teórica de viviendas	**Muestra efectiva de viviendas**
Almería	604	578
Cádiz	918	858
Córdoba	598	569
Granada	598	510
Huelva	604	588

Jaén	604	553
Málaga	978	878
Sevilla	1096	982
Andalucía	6000	5516

Nota: Datos provisionales

Tabla 3. Distribución de la muestra efectiva de viviendas y niños/as nacidos/as en 1994 y en 1998

	Muestra efectiva Viviendas 94	**Muestra efectiva niños/as 94**	**Muestra efectiva Viviendas 98**	**Muestra efectiva niños/as 98**
Almería	292	282	286	279
Cádiz	454	393	404	347
Córdoba	275	247	294	264
Granada	250	213	260	220
Huelva	301	291	287	277
Jaén	276	265	277	267
Málaga	466	455	412	394
Sevilla	512	471	470	429
Andalucía	2826	2617	2690	2477

Nota: Datos provisionales

d. Selección de la muestra

La selección de la muestra se ha realizado de forma que dentro de cada estrato cualquier vivienda familiar tenga la misma probabilidad de ser seleccionada, es decir, se tengan muestras autoponderadas a nivel de estrato. Este tipo de muestra proporciona pesos de diseño iguales por estrato en los estimadores.

Para ello, dentro de cada estrato, las secciones se han seleccionado por un muestreo proporcional al tamaño de la sección y, en cada sección, las viviendas se han seleccionado por un muestreo aleatorio simple.

Por lo tanto, la probabilidad de selección de la vivienda i, perteneciente a la sección j del estrato h, en la que se han afijado K_h secciones, sería:

$$P(V_{ijh}) = P(S_{jh}) \times P(V_{ijh}/S_{jh}) = K_h \times \frac{V_{jh}}{V_h} \times \frac{8}{V_{jh}} = K_h \times \frac{8}{V_h}$$

donde:

- $P(S_{jh})$ es la probabilidad de selección de la sección j del estrato h
- $P(V_{ijh}/S_{jh})$ es la probabilidad de selección de la vivienda i condicionada a la selección de la sección j
- V_{jh} es el total de viviendas de la sección j
- V_h es el total de viviendas del estrato h.

Como se observa, esta probabilidad no depende de i ni de j, es decir, ni de la vivienda ni de la sección, y por lo tanto la muestra es autoponderada.
Por último, hay que indicar que en caso de existir más de un menor en la vivienda de la misma edad requerida, se selecciona aleatoriamente a uno de ellos. Esta última etapa también se tiene en cuenta para calcular los factores de elevación de los menores.

7. Incidencias en la muestra y tratamiento

En los casos en que se presentó alguna incidencia justificada en el trabajo de campo, los elementos de la muestra de viviendas podían ser sustituidos por otros. El número máximo de sustitutos por sección se fijó en ocho viviendas.
Las incidencias que pueden motivar una sustitución y su tratamiento se exponen a continuación.

A) Incidencias al localizar y acceder a la vivienda

- Vivienda inaccesible: es aquella a la que no se puede acceder para realizar la entrevista por causas climatológicas (inundaciones, nevadas, etc.) geográficas (cuando no existen vías transitables para llegar a la misma) o de cualquier otro tipo.
  Las viviendas inaccesibles sólo son objeto de sustitución, si no desaparece la causa de la inaccesibilidad durante el tiempo que dure el trabajo de campo en el municipio.

- Vivienda ilocalizable: se produce esta incidencia cuando no se localiza la vivienda por un error en la dirección de partida. La vivienda no puede ser localizada en la dirección que figura en la relación de viviendas seleccionadas, bien porque la dirección no es correcta o bien porque ya no existe físicamente la vivienda.

- Vivienda vacía: la vivienda seleccionada está deshabitada por cualquier causa, como puede ser el fallecimiento, cambio de residencia de las personas que vivían en la misma, estar en reforma, en ruina, demolida,...

- Vivienda destinada a otros fines: la vivienda seleccionada se dedica en su totalidad a fines diferentes a residencia familiar. Por ejemplo: comercio, garaje, oficina, residencia de ancianos, etc.
  En todos estos casos las viviendas se sustituyen por otras de la misma sección. Para ello el entrevistador dispone de una relación de viviendas reserva para utilizarla cuando haya que sustituir alguna de las viviendas originalmente seleccionadas.

B) Incidencias al contactar con el grupo humano que reside en la Vivienda

Además de las incidencias propias de los procesos de localización, pueden surgir otras relacionadas con el intento de contactar con las personas que residen en la vivienda o bien en la relación que se establece con éstas. Serían las siguientes:

- Ausencia: se produce cuando no se consigue contactar con nadie de la vivienda, bien porque se obtiene información de que están ausentes todos sus ocupantes y van a seguir estándolo durante el periodo de tiempo que dure el trabajo de campo en la sección, o cuando tras las sucesivas visitas estipuladas a la vivienda no se consigue

contactar con los padres/tutores, sin que se tenga información de que la causa sea otra distinta a la ausencia:

Una vez agotadas todas las visitas previstas, si los residentes en la vivienda siguen ausentes o el tutor está ausente, se sustituye la vivienda. En 'observaciones' se deberá anotar cómo se ha sabido que están ausentes y la causa si se conoce.

- Negativa: se considera este caso cuando no se consigue hacer la entrevista por negativa a contestar de las personas que residen en la vivienda o por negativa de los padres/tutores, de forma que no se consigue un cuestionario que cumpla los requisitos de 'Cumplimentado'. Puede darse una negativa rotunda desde el primer momento, o producirse por aplazamientos, evasivas o posteriormente, después de haber empezado a colaborar.

  Cuando se da cualquiera de estas circunstancias se sustituye la vivienda por la primera reserva disponible.

- Incapacidad para contestar: se produce esta incidencia cuando no se consigue hacer la entrevista por incapacidad del tutor para responder a la misma a causa de alguna discapacidad, enfermedad, desconocimiento del idioma o cualquier otra circunstancia. Antes de proceder a la sustitución, se debe intentar (siempre sin forzarle) hacer la encuesta a través de alguien próximo. Si finalmente no es posible, se procede a su sustitución.

- Desconocido: se produce cuando al contactar con el hogar, las personas que allí residen manifiestan no conocer al menor. En estos casos se sustituye la vivienda por la primera reserva disponible.

- Residente en otra vivienda: se considera este caso cuando al contactar con el hogar nos indican que el menor reside habitualmente en otra dirección y no es posible realizar el contacto en la nueva dirección. En estos casos se sustituye la vivienda por la primera reserva disponible.

- Vivienda seleccionada anteriormente: tiene lugar cuando la vivienda seleccionada para esta encuesta lo ha sido anteriormente (hace menos de cinco años) en cualquier otra encuesta de población realizada por un organismo público oficial y colaboró en la misma.

  Cuando esta situación se detecte antes de la salida a campo, la vivienda será sustituida por la primera reserva válida disponible sin necesidad de que sea visitada. En caso de que la anterior colaboración no se detecte previamente a la salida a campo, sino ya en la propia visita a la vivienda, existirán dos posibles tratamientos: a) si el grupo humano que habita la vivienda acepta colaborar en la encuesta se le hará la entrevista normalmente; b) si el grupo humano no acepta colaborar debido a una anterior colaboración, se sustituirá la vivienda por la primera reserva válida disponible.
- Otra causa de no colaboración: esta es una incidencia 'residual'; se incluye por si se da algún caso en que no se consigue un cuestionario que cumpla los requisitos de 'Cumplimentado' y la causa es distinta de las anteriores.

- Vivienda colaboradora: Se da esta situación cuando se ha contestado al cuestionario completo o al menos a las preguntas que se consideran necesarias para que el cuestionario esté 'Cumplimentado'.

C) Incidencias al contactar con el menor seleccionado

Como ya se ha descrito esta encuesta consta de un cuestionario que debe responder un menor seleccionado previamente. A la hora de su cumplimentación pueden darse las incidencias siguientes:

- Negativa: se considera este caso cuando el menor se niega a dar la información que se le solicita o los padres/tutores se niegan a que el niño haga la entrevista.

- Ausencia: se produce cuando tras las visitas a la vivienda estipulada no se consigue hacer la entrevista porque el niño que resulta seleccionado está ausente y no se consigue contactar con él.

- Incapacidad para contestar: se produce esta incidencia cuando el menor seleccionado no es capaz de responder a la entrevista, ya sea por discapacidad, enfermedad, desconocimiento del idioma o cualquier otra circunstancia.

- Colaboración del niño: se anota este código cuando no se produce ninguna de estas incidencias y el niño seleccionado, contesta al apartado de hijos del cuestionario, cumpliendo los requisitos de 'Cumplimentado'.

En ninguna de las tres situaciones anteriores (negativa, ausencia e incapacidad para contestar) está permitido que otra persona del hogar conteste por el niño seleccionado. En el caso de ausencia se desarrollarán protocolos de contacto para intentar realizar este cuestionario. Este tipo de incidencias no son causa de sustitución.

8. La recogida de la información

El método de entrevista utilizado preferentemente es el de la entrevista personal asistida por ordenador (CAPI), aunque también se ofreció la posibilidad de suministrar la información a través de Internet (CAWI).
En la ESOC-2010, la información de la encuesta ha sido proporcionada por dos informantes. La información relativa al hogar ha sido proporcionada por el padre, madre o tutor/a del niño/a que esté al corriente de la materia sobre la que versa la encuesta (informante de la vivienda). En cambio, la información del niño/a sólo puede darla el/la menor seleccionado/a aleatoriamente, sin admitirse que otra persona del hogar conteste por él/ella.

ENLACE DE REGISTROS. MEJORA EN LA CALIDAD

En los inicios del proyecto se investigó la existencia de alguna fuente de información que pudiese complementar a la encuesta y así evitar una duplicidad en la recogida de información lo que conllevaría un aumento en la carga de información al informante. De esta forma, surgió la colaboración con la Consejería de Educación que nos permitió el acceso al sistema de información SÉNECA, el cual aglutina en sí toda la información referente a los centros educativos andaluces.

El Sistema de Información Séneca se constituye como el instrumento preciso para la gestión telemática integral de los centros docentes, los servicios de apoyo a la educación, los programas y las actividades del sistema educativo andaluz, a través de la utilización de las tecnologías de la información y las comunicaciones en un entorno seguro e integrado de tramitación de documentos en el marco de las infraestructuras de Administración electrónica reguladas y gestionadas por la Junta de Andalucía, favoreciendo un acceso igualitario de la población a los servicios educativos.

Séneca incluye los datos de carácter personal, referidos al profesorado y otro personal que presta servicio en los centros docentes y al alumnado y sus familias, que deban ser facilitados por los centros docentes públicos y privados concertados. Asimismo, los centros docentes privados ceden a la Administración educativa los datos de carácter personal a los que esta debe tener acceso para el ejercicio de las funciones que le son propias en el ámbito de sus competencias.

Por esta razón, SÉNECA se ha convertido en una fuente de información totalmente enriquecedora para la ESOC-2010, ya sea desde el punto de vista del análisis de los datos como para la depuración de los mismos. La utilización de SÉNECA se ha realizado desde varias ópticas:

- Una reducción en la carga de información del informante.

- En la depuración de la muestra, ya que al disponer de los teléfonos de un elevado porcentaje de las viviendas en muestra, nos ha permitido llevar a cabo el proceso de confirmación muestral y establecer controles en el cuestionario introduciendo variables recogidas en SÉNECA.

- Un análisis longitudinal de los datos, ya que dispondremos de la trayectoria escolar de los alumnos a través de los distintos cursos

1. Procedimiento de enlace

El procedimiento de enlace de ambos registros se ha podido llevar a cabo gracias al conjunto de campos o variables que comparten. Por otro lado, hay que indicar que el enlace se ha realizado teniendo en cuenta tanto la información de los alumnos de Séneca como la de sus tutores. Para ello, han resultado imprescindibles los campos nombre y apellidos, fecha de nacimiento y documento de identidad (especialmente en el caso de los tutores).

En una etapa previa a la operación de enlace, se trataron de normalizar las variables de Séneca conforme a los mismos criterios de RPA, con el objetivo de maximizar las posibilidades de éxitos. En el caso de los nombres y apellidos se eliminaron caracteres extraños, acentos, y partículas. Con los documentos de identidad, se comprobó si su longitud era la adecuada, por ejemplo 8 dígitos en el caso de los DNI, completándose con ceros a la izquierda en los casos necesarios.

En cuanto al procedimiento de enlace en sí, se pueden resumir los pasos llevados a cabo del siguiente modo:

- Coincidencia exacta de literales y fecha de nacimiento para alumnos/as y de literales y DNI (no nulo) para tutores/as.
- Coincidencia de literales estandarizados y, además, fecha de nacimiento para alumnos y DNI (no nulo) para tutores
- Coincidencia de DNI del tutor y proximidad de literales.
- Una vez localizado a alguno de los intervinientes y faltando el enlace del resto, se busca entre las personas que alguna vez han convivido (misma hoja padronal o misma clave de vivienda) con dicha persona y se aplican criterios de proximidad entre los pares.

Hay que indicar que la operación de enlace entre estas dos fuentes de información también ha resultado útil para la depuración de algunos registros de RPA. En particular, ha servido para detectar casos duplicados en el registro de población gracias a que durante el proceso de enlace se les asignó la misma clave de educación a dos individuos, en principio, distintos.

2. Resultados del enlace

En primer lugar, tenemos que comentar que el enlace se ha implementado para el conjunto completo de los registros de RPA y Séneca de las subpoblaciones de menores nacidos en 1994 ó 1998. En general, este enlace nos ha proporcionado un porcentaje de éxitos en la localización de alumnos muy satisfactorio, superior al 90 por ciento. En relación con la muestra definitiva, los resultados no son sólo positivos porque el enlace haya alcanzado el 94 por ciento, sino que cobra más importancia porque esto nos ha permitido conocer los nombres de los tutores y contar con un teléfono de contacto, móviles en una amplia mayoría, en algo más de un 88 por ciento de los casos. Y en un 46 por ciento de la muestra total, se contaba con dos números de teléfono.

En la siguiente tabla podemos observar, como era de esperar, que las cifras referidas al enlace son muy similares en la muestra principal y de reserva.

Tabla 4. Estadísticos descriptivos de la proporción de viviendas de la intersección en la muestra principal.

	Muestra total		Principales		Reservas	
	Viviendas	%	Viviendas	%	Viviendas	%
Seleccionadas	14.382	100,00	6.000	100,00	8382	100,00
Enlazadas	13.542	94,16	5.681	94,68	7861	93,78
Enlazadas con tutor/a	12.751	88,66	5.342	89,03	7409	88,39
Enlazas con algún teléfono	12.725	88,48	5.332	88,87	7393	88,20
Enlazadas con teléfono 1	12.629	87,81	5.297	88,28	7332	87,47
Enlazadas con teléfono 2	6.742	46,88	2.860	47,67	3882	46,31

Este procedimiento de enlace se va a someter a un doble proceso de evaluación a lo largo de las distintas fases de la Encuesta Social 2010. El primero de ellos con resultados satisfactorios, ya que previamente a la salida de campo, durante los trabajos de confirmación de muestra, se ha constatado que los teléfonos corresponden a los hogares de los menores seleccionados. La segunda evaluación se llevará a cabo cuando en la Consejería de Educación se recuperen los resultados de los alumnos enlazados a través de su clave en Séneca, volviendo a contrastar la información con la recogida en la encuesta.

3. Fases del Enlace. Intercambio de información

Una vez realizado el enlace entre la muestra y SÉNECA se establecieron distintas fases para el intercambio de información:

Fase 1) Previo al trabajo de campo

En la primera fase, en la cual se produce el enlace entre las dos fuentes, sólo vamos a utilizar los datos personales de los tutores y de las viviendas (dirección…) que han sido seleccionadas en muestra, para tener información auxiliar cargada en las herramientas de trabajo que nos sirvan de contraste con la proporcionada por el RPA.

Además, en esta primera fase se obtienen los teléfonos de cada una de las viviendas seleccionadas y cuyo teléfono o teléfonos estuviera cargado en Séneca. Esta información es de vital importancia pues previo a la salida al trabajo de campo se ha realizado por parte de componentes de la Unidad Central de Encuestación del IEA una tarea de confirmación de muestra, es decir, de comprobación de la dirección de las viviendas, a través de llamadas telefónicas a las propias viviendas.

Fase 2) Durante el trabajo de campo

En esta segunda fase, que temporalmente coincide con el final del trabajo de campo, se enriquece la información de la misma obteniendo diversas variables claves en el estudio. Estas variables versan sobre temas como matriculación en el curso 2009-2010, evaluación del curso anterior, bloque absentismo escolar y variables de disciplina. Todos estos datos, se deben anexar a los propios datos obtenidos en la encuesta a través de la entrevista. También se

utiliza como variables de control, viendo si los resultados de la encuesta resultan coherentes respecto a los presentes en Séneca.

Fase 3) Tras el trabajo de campo

Esta última fase se puede enmarcar como caso particular del anterior, pero por temas de temporalización los datos de matriculación en el curso 2010-11, y los resultados de las pruebas de diagnóstico no se pueden obtener hasta esta última fase. La única diferencia es que este enlace sólo se realiza para las viviendas efectivamente encuestadas.

4. Mejora en la calidad de la encuesta

Una vez analizadas las ventajas de la utilización de una fuente de información auxiliar a los datos de la encuesta, veamos como se refleja en los datos obtenidos del trabajo de campo. En primer lugar, el proceso de confirmación muestral realizado antes de la salida a campo permitió primero, una depuración de la muestra, detectando casos de unidades no encuestables y, segundo una reducción de las incidencias de ausencias en las unidades encuestables, al permitirnos los teléfonos establecer una cita con los hogares para realizar las entrevistas.

En la siguiente tabla se observa como a través del proceso de confirmación muestral se confirmó que las direcciones del 70% de la muestra estaban bien recogidas en nuestro fichero de datos.

Tabla 5. Resultados de la confirmación de muestra titular.

Casos	Total
Dirección confirmada con menor encuestable	4217
Dirección no confirmada con menor encuestable	29
Negativa	394
Otro fines	13
Vivienda sin menor encuestable	291
Sustitución	91
Sin contacto	340
Sin teléfono	625
Total	6000

En la Tabla 6 se refleja el pequeño porcentaje de ausencias de los integrantes del hogar

Tabla 6. Distribución de las incidencias en las viviendas titulares encuestables

	Encuestables	Encuestadas	Falta de Respuesta		
			NEG	AUS	Otro tipo de no repuesta
Almería	497	87,1	10,3	1,6	1,0
Cádiz	789	78,1	16,7	3,7	1,5
Córdoba	528	79,5	14,8	4,7	0,9
Granada	490	72,9	20,0	6,1	1,0
Huelva	547	79,7	15,4	4,8	0,2
Jaén	525	81,0	16,6	2,1	0,4
Málaga	744	78,1	16,9	3,8	1,2
Sevilla	946	73,9	21,1	4,2	0,7
Andalucía	5066	78,3	16,9	**3,9**	0,9

Nota: Datos provisionales

La posibilidad de contactar con las familias de los hogares seleccionados por teléfono, para presentar el proyecto, confirmar los datos de muestra y ofrecer la posibilidad de concertar una cita para la entrevista, ha sido decisiva en el sustancial incremento que ha experimentado el volumen de viviendas titulares en la muestra definitiva en relación a otras encuestas dirigidas a hogares. Máxime cuando el hogar se fijaba en muestra a partir de los datos nominales de un menor seleccionado, cuya presencia en el hogar, como ya se ha comentado, era obligatoria para participar en el proyecto.

Tabla 7. Distribución porcentual de viviendas titulares y reservas en la muestra definitiva.

Provincia	Muestra efectiva	% Titular	% Reserva
Almería	578	74,91	25,09
Cádiz	858	71,79	28,21
Córdoba	569	73,81	26,19
Granada	510	70,00	30,00
Huelva	588	74,15	25,85
Jaén	553	76,85	23,15
Málaga	878	66,17	33,83
Sevilla	982	71,18	28,82
Andalucía	5516	71,92	28,08

Nota: Datos provisionales

El hecho de realizar el proceso de confirmación de muestra, ha influido también enormemente en la realización de un gran número de cuestionarios al niño/a seleccionado, ya que el entrevistador solía acudir al hogar cuando se encontraban en él las dos personas informantes de la encuesta:

Tabla 8. Distribución del número de cuestionarios de niños/as por provincia

Provincia	Muestra efectiva	% Muestra al niño/a seleccionado/a
Almería	578	97,06
Cádiz	858	86,25
Córdoba	569	89,81
Granada	510	84,90
Huelva	588	96,60
Jaén	553	96,20
Málaga	878	96,47
Sevilla	982	91,65
Andalucía	5516	92,31

Nota: Datos provisionales