Metodología de Encuestas
Volumen 13, 2011, 55-70
ISSN: 1575-7803

PROCESOS DE VALIDACIÓN, DEPURACIÓN E IMPUTACIÓN DE LAS ENCUESTAS A HOGARES DEL IEA

Hernández Moreno, Antonio
Muñoz Conde, María
Bohórquez Baña, Rosa

Instituto de Estadística de Andalucía.

Correo electrónico: mmacarena.munoz@juntadeandalucia.es

Dirección: Servicio de Estudios, Síntesis y Métodos Estadísticos. Instituto de Estadística de Andalucía. Pabellón de Nueva Zelanda. Leonardo Da Vinci, nº 21. Isla de La Cartuja. 41071-Sevilla, Telf.: 955 033 949

RESUMEN: Toda encuesta ha de estar sometida a determinados requisitos de calidad. Ésta puede evaluarse valorando los dos tipos de errores que se pueden producir en toda investigación a través de encuestas, el llamado error de muestreo, que puede medirse con los correspondientes cálculos probabilísticos, siempre que la muestra se haya obtenido bajo determinadas condiciones formales, y el denominado error ajeno al muestreo, el cual es difícil de cuantificar de manera global, pero sí podemos apreciarlo analizando los métodos utilizados para adaptar y corregir toda la posible casuística que puede presentarse en la definición del problema, en la construcción de los elementos de medición y en los detalles del trabajo de campo. Este artículo se centra en la descripción de los procedimientos que se han utilizado para validar la información recogida, disminuyendo los errores ajenos al muestreo, y por consiguiente incrementando la calidad de la Encuesta Social 2008: Hogares y Medio Ambiente en Andalucía realizada por el Instituto de Estadística de Andalucía.

PALABRAS CLAVE: encuesta, error ajeno al muestreo, calidad, validación

ABSTRACT: All surveys must be subject to certain quality requirements. This can be evaluated assessing the two types of errors that can occur in any research through surveys, called the sampling error, which can be measured with appropriate probabilistic calculations, when the sample has been obtained under certain formal conditions, and the so-called non-sampling error, which is difficult to quantify in a formal way, but we can appreciate looking at the methods used to adjust and correct any possible casuistry that can occur in the problem definition, in the construction of the measuring element and the details of the fieldwork. This article focuses on the description of the procedures that were used to validate the information collection and thus reduce nonsampling errors, and thereby enhance the quality of the Social Survey 2008: Households and Environment in Andalusia by the Institute of Statistics of Andalusia.

.

KEYWORDS: survey, nonsampling errors, quality, validation

Recibido: 23 de febrero 2011
Revisado: 2 de abril 2011
Aceptado: 25 de mayo 2011

Met. de Encuestas 13, 2011, 57-70

1.-Introducción

Uno de los principios fundamentales de cualquier institución que se dedique a realizar encuestas debe ser producir estadísticas de calidad, precisas y fiables. Por ello, simultáneamente a la recogida de información en los trabajos de campo, se deben desarrollar estrictos procesos de validación y depuración de los datos recogidos, que permitan detectar y corregir los errores que puedan surgir, sean éstos de la naturaleza que sean, y además valorar la posibilidad de eliminar encuestas que no cumplan con los criterios de calidad establecidos.

Como se verá a continuación, la validación, depuración e imputación no se deben entender como procesos independientes sino como acciones interrelacionadas que forman un proceso conjunto con un objetivo central: la obtención de información de calidad.

En este artículo se realiza un breve resumen de las etapas que suelen seguirse en cualquier encuesta por muestreo, para a continuación centrarnos en la recogida de información y analizar la necesidad de desarrollar una serie de procedimientos que nos permitan validar la información recogida por los encuestadores. Para ello se presentan los procedimientos de validación que se han utilizado por el Instituto de Estadística de Andalucía en sus encuestas a hogares, y más concretamente en la Encuesta Social 2008.

2.- Producción de operaciones estadísticas

El objetivo al que se enfrenta cualquier institución pública o privada encargada de realizar estadísticas en el ámbito público es la de obtener datos que satisfagan las necesidades de información que la sociedad demanda y la toma de decisiones en distintos ámbitos. Para obtener dichos datos, diseñan y realizan una serie de estudios a partir de datos muestrales, censales o en el caso que existiesen, aprovechando la información presente en los distintos registros administrativos.

Por lo tanto, todo producto estadístico surge como respuesta a una necesidad de información, que una vez detectada, debe analizarse, para determinar si debe llevarse a cabo a través de una operación estadística nueva o puede abordarse desde alguna otra operación estadística existente.

Tras detectar que efectivamente es necesario planificar una operación de recogida de información mediante una encuesta hay que poner en marcha las siguientes etapas:

1) Planificación de la operación estadística
2) Diseño
3) Ejecución de la operación.
4) Validación, depuración e imputación, de forma conjunta.
5) Difusión de los resultados.

Vamos a desarrollar brevemente cada una de las etapas enumeradas.

1) Planificación

Ésta es una de las etapas principales en las encuestas ya que antes de comenzar cualquier proyecto es necesario definir los objetivos principales del estudio, determinar quiénes serán sus usuarios potenciales y analizar la información disponible del objeto de estudio. Como resultado de esta etapa se definen:

- La población objeto del estudio al que se quiere dirigir la operación.
- Método de recogida de la información.
- Ámbito geográfico y temporal de la operación, ya sea muestral, censal u obtenida de un registro administrativo.
- Cronograma aproximado del proyecto.
- Organización de los trabajos.
- Estimación de costes asociados a los trabajos y obtención de la financiación necesaria para llevarlos a cabo.

2) Diseño

Una vez decididos en la etapa anterior los rasgos principales de la operación estadística, se deben establecer los métodos adecuados para cumplir con todos los objetivos deseados, definiendo todos los procedimientos que nos lleven a tal fin. Estos procedimientos varían notablemente según nos encontremos con una operación censal o muestral. Podemos destacar como procedimientos asociados a esta etapa

- Diseño del marco poblacional y plan de muestreo en el caso de que se trate de una encuesta por muestreo.
- Diseño del cuestionario.
- Método de recogida.
- Definición de los métodos de validación, depuración, imputación y estimación.
- Definición de los sistemas informáticos que soportarán los métodos diseñados en las actividades anteriores.
- Documentación de procedimientos, tanto para las tareas manuales como informatizadas.

3) Ejecución

En esta etapa es en la que se obtienen los datos de la operación estadística, que posteriormente habrá que tratar, siguiendo los pasos descritos en la etapa de diseño.

4) Validación, depuración e imputación

Ante la publicación de los resultados de una encuesta no es infrecuente escuchar preguntas como cual ha sido el nivel de validación al que ha sido sometida. El término *validación* se maneja en numerosas ocasiones de manera confusa. Es importante puntualizar que, en principio, lo que puede validarse es un instrumento o procedimiento concebido para medir una dimensión correcta, sea esta física o abstracta. Validar tal instrumento equivale a corroborar que realiza de manera efectiva el procedimiento de medición que le corresponde. Es muy común que para medir cualquier magnitud se empleen variables sintéticas construidas a partir de las

respuestas que se obtienen del entrevistado a través de una encuesta. De forma que, esta encuesta y procedimiento tendrán que ser validados.

Como fases del procedimiento de validación distinguimos, por un lado, la validación del cuestionario, a través generalmente de la realización de una encuesta piloto, y por otro, tratar e investigar los trabajos que desarrollan los encuestadores mediante la validación de los procedimientos de encuestación y detección de encuestas fraudulentas a través de distintas estrategias. Pero, como se ha comentado anteriormente, el procedimiento de validación no se debe ver como un procedimiento aislado e independiente sino que va unido a las fases de depuración e imputación de los datos, tal y como veremos.

Uno de los puntos destacados de esta validación será la de minimizar y corregir los errores detectados, por lo que pasamos a exponer brevemente una clasificación de los distintos tipos de errores que podemos encontrarnos durante la fase de recogida de información. No nos referimos a los errores que provienen del muestreo, sino a los que se pueden ver reflejados en los propios datos de la encuesta, es decir, los denominados errores ajenos al muestreo. Estos errores podemos dividirlos en dos modalidades, errores en las identificaciones de cada encuesta –el encuestador, de forma intencionada o no, realiza la encuesta a otra vivienda distinta a la vivienda objetivo, o en el peor de los casos comete fraude al responder él mismo a las preguntas de la encuesta - y los errores en los propios datos - Granquist (1984) los divide a su vez en dos tipos, los errores aleatorios y los errores sistemáticos cometidos por los entrevistadores. Para intentar disminuir estos errores, se proponen los métodos de validación para los errores del primer tipo y los métodos de imputación y depuración para los segundos. De ahí, la relevancia de prestar la máxima atención en el proceso de recogida de información desarrollando procedimientos que controlen dicho proceso

5) Difusión

Por último, toda operación estadística pública se debe ajustar a los planes diseñados para su posterior difusión. La calidad de la investigación estadística se logra extremando el cuidado en la realización de las etapas anteriores.

Como existen tantos tipos de investigaciones como necesidades de información, no hay reglas exactas para cada una de las etapas, pero si existen guías de actuación que nos pueden ayudar a desarrollarlas según sea el caso.

3.- Origen y Objetivos de la Unidad de Encuesta del IEA

Desde sus comienzos, la unidad de encuesta del Instituto de Estadística de Andalucía (IEA), encargada de realizar las encuestas sociales, tiene como uno de sus objetivos principales asegurar la obtención de unos datos de calidad. Para ello ha ido diseñando y modificando sus planes de validación, depuración e imputación de las encuestas hasta llegar al modelo actual, obteniendo resultados cada vez más satisfactorios.

A lo largo de su trayectoria, la unidad de encuesta del IEA ha acometido las siguientes operaciones estadísticas dirigidas a hogares:

1. Encuesta de Redes Familiares en Andalucía: 10.000 encuestas a individuos (2005)
2. Encuesta Mundial de Valores (EMV): 2.000 encuestas a viviendas (2006-2007)
3. Encuesta Social 2007. Una Visión de Andalucía (ESOC-07): 2.000 encuestas a viviendas
4. Encuesta Social 2008: Hogares y Medio Ambiente en Andalucía (ESOC-08): 6.000 encuestas a viviendas

El esfuerzo humano, económico y de recursos en cada proyecto pone de manifiesto este interés en asegurar dicha calidad. Asimismo, el hecho de mantener el equipo de validación de forma constante a lo largo de estos años ha permitido que se aumente la calidad de los datos, lo que además, se ha realizado con una carga de trabajo mucho menor gracias a la sistematización de los procedimientos. En la siguiente tabla podemos ver el número de validaciones realizadas en alguna una de las operaciones anteriormente enunciadas:

Tabla 1.
Histórico de validaciones en las operaciones de la Unidad de Encuesta del IEA.

Encuesta	**Validación telefónica (%)**
EMV	37,20
ESOC-07	45,55
ESOC-08	34,87

Una de las apuestas iniciales que realizó el IEA, que supuso una mejora en la calidad de los resultados, a la hora de crear su unidad de encuesta, fue hacer uso del sistema CAPI (*Computer Assisted Personal Interviewing*), utilizando dispositivos PDA como herramienta de recogida de la información. Este sistema cuenta con numerosas ventajas respecto a la validación de encuestas. Una de las principales, es que se dispone de la fecha y hora de realización de la entrevista y de la duración de la misma. De esta forma, se pueden marcar como sospechosas aquellas encuestas realizadas a horas intempestivas y aquellas cuya duración resulta mucho más corta del resto. Esto permite establecer patrones de duración de las encuestas por cada encuestador. Otra de las ventajas de trabajar con PDA a la hora de la recogida de información en una encuesta, es que la fase de depuración comienza en el mismo diseño del cuestionario en la PDA, ya que se van definiendo los filtros directamente en el diseño del cuestionario, evitando tener que depurarlos una vez que se han recibido los datos. Por último, en la recogida de información a través PDA, se tiene protocolarizado que el encuestador no pueda editar una encuesta una vez que la ha dado por definitiva, esté finalizada o incompleta, pudiendo realizar esta edición sólo los miembros de la Unidad Central de Encuestas, tanto directamente en la PDA como en las aplicaciones informáticas posteriores.

Otro aspecto que cada vez ha ido teniendo mayor importancia en la validación de las encuestas del IEA es la minimización del tiempo que transcurre entre la realización de la encuesta y los posibles re contactos con el encuestado, por ejemplo, a través de la realización de un cuestionario de validación, para que así el

suministrador de los datos no tenga dudas y recuerde en gran parte la encuesta a la que se refiere dicha validación. Este proceso de validación se realiza a través de dos vías: la validación telefónica y la validación "in situ". En los dos siguientes epígrafes (4 y 5) detallaremos cada uno de ellos:

4.- Mecanismo de validación telefónica del IEA

El procedimiento para la validación telefónica, está basado en la aplicación de una serie de controles que permiten la clasificación de las encuestas en estados, teniendo cada uno ellos, posteriormente, un tratamiento específico. Aunque para todas las operaciones realizadas el procedimiento ha sido similar en este texto nos vamos a centrar en el mecanismo utilizado en la "Encuesta Social 2008: Hogares y Medio Ambiente en Andalucía" (ESOC-08).

Este estudio fue promovido por el Instituto Nacional de Estadística (INE), y en él colaboraron las oficinas de estadística de varias Comunidades Autónomas. En Andalucía el organismo que la ha llevado a cabo es el Instituto de Estadística de Andalucía (IEA), que como se expuso anteriormente, cuenta desde hace cuatro años con una Unidad de Encuesta, encargada de planificar y coordinar este tipo de estudios. Esquemáticamente el papel específico de cada una de las instituciones involucradas en el proceso fue:

- El INE coordinó el proyecto y asumió la realización del trabajo de campo en todo el territorio nacional, salvo en las comunidades autónomas participantes en el proyecto.
- Los institutos de estadística autonómicos colaboraron en la elaboración del cuestionario y se hicieron cargo de la realización y coordinación regional de los trabajos de campo en sus respectivos territorios.

En el IEA, nuestra unidad trabaja en cooperación con equipos de las universidades andaluzas, encargados de la recogida de información. La operación se basa en una **muestra de viviendas**, en las que se selecciona a determinados **informantes**. El cuestionario se encuentra dividido en 9 módulos que recogen diferentes aspectos relacionados con el Medio Ambiente.

4.1.- Procedimiento de validación telefónica para la ESOC-08

El sistema de validación aplicado a las encuestas del IEA es principalmente telefónico, para lo cual es necesaria la obtención de los números de teléfono por parte de los entrevistadores durante la recogida de información, a través de cualquier registro administrativo del que se disponga o a través de las distintas bases de datos telefónicas existentes en el mercado.

Para la validación telefónica se ha diseñado un cuestionario general y uno para cada uno de los bloques que formaban parte del cuestionario de la encuesta. ¿Cuándo utilizamos uno u otro? Para poder decidir sobre este particular se utiliza la información que nos proporciona el Registro de Población de Andalucía (RPA)

gestionado por el IEA. En él se recoge el Nombre, Apellidos, Domicilio, Sexo y Fecha de Nacimiento de los individuos empadronados en Andalucía.

Pero, ¿cómo usamos el RPA?

En las encuestas donde la unidad muestral es la vivienda, usualmente, se pregunta al principio del cuestionario la estructura del hogar, es decir, se pregunta año de nacimiento y sexo de cada una de las personas que residen en la vivienda –conocido generalmente como tabla de miembros de la vivienda-, ya que es información socio demográfica de gran importancia de cara a la explotación de los resultados y al cruce con otras variables.

Por otro lado, en el RPA, como acabamos de ver, también tenemos esta misma información, en teoría procedente de la misma vivienda, lo cual facilita la comparación de dichas fuentes y la consecuente obtención de un criterio de decisión a la hora de decidir si una encuesta tiene que ser validada. Este criterio es fundamentalmente el siguiente:

- Si coincide de forma estricta la estructura del hogar del RPA con la presentada en la encuesta (se comprueba primero que coincida el número de personas y posteriormente que las personas que estén tengan las mismas características de edad y sexo) se valida sólo un 10% de los casos, ya que tenemos evidencia de haber visitado la vivienda correcta.
- En el caso de que no coincida se valida un 80% de los casos. Visto el alto porcentaje de validación que se produce en estos casos, debemos estar seguros de que la estructura del hogar difiere en gran medida, por lo que se procede a una breve depuración manual en los casos más próximos a la coincidencia. Esta depuración manual la realiza el equipo de validación a través de una aplicación en Visual Express Studio.

Por lo tanto, al recibir los datos de una encuesta, lo primero que se hace es encuadrarla en una de estas dos categorías (una vez que los códigos de las mismas están depurados), que a partir de ahora denominaremos *Correcto_RPA,* cuando ambas fuentes coincidan, e *Incorrecto_RPA* cuando no lo hagan. Este procedimiento está automatizado en Visual Express Studio existiendo una aplicación para ello.

En esta comparación con el RPA, somos conscientes que se pueden producir errores por el desfase temporal entre la fecha de referencia del RPA y el momento de realización de la encuesta. Es decir, en un momento dado el registro de viviendas podría quedar obsoleto, y por tanto no tendríamos a priori la misma estructura en el RPA y en la encuesta. En este caso, esa encuesta se catalogaría en el grupo de *Incorrecto_RPA*. Contamos con este error, pero la mejora sustancial a la que ha sido sometido el RPA en los últimos años a través del trabajo del Servicio de Demográficas y Sociales del IEA, nos hace prever que aún cometiendo errores, las ventajas que obtenemos de esta fuente son cada vez mayores, lo que permite un notable ahorro de esfuerzo a la hora de localizar entrevistas sospechosas.

Siguiendo esta regla de decisión, y los porcentajes fijados anteriormente, se seleccionan aleatoriamente las encuestas que debemos validar de forma telefónica, realizándoles el cuestionario de validación extraído del cuestionario original. Los

resultados de la validación se clasifican en cuatro categorías, que son: Validada, No Validable, Llamada positiva y Sospechosa. Las explicamos más detalladamente:

. Validada: La primera pregunta que se realiza en cualquier cuestionario de validación, tanto telefónico como presencial, es saber sí a la persona en concreto se le ha realizado una encuesta a través de los mecanismos correctos, es decir, en nuestro caso si en la vivienda seleccionada se ha personado un encuestador con PDA para hacerle unas preguntas sobre el tema en cuestión. En el caso que sea afirmativa la respuesta, se procede a preguntarle el breve cuestionario de validación. Si éste es contestado en su totalidad, se codifica en esta categoría.

. Llamada positiva: existirá un amplio grupo de personas que no quieran responder a más preguntas, es decir, que no faciliten las respuestas del cuestionario de validación pero si aseguren que por su vivienda se ha personado un encuestador del IEA con su PDA para realizarle la encuesta. En este caso se codifica la encuesta como llamada positiva.

. No Validable: se trata del caso en el que el equipo de validación no ha sido capaz de contactar de ninguna de las formas posibles con la vivienda seleccionada, bien sea por no poder tener contacto telefónico o porque una vez iniciado el contacto se niegue a dar información alguna sobre la encuesta. En este caso, además de clasificar la encuesta como *no validable*, se sustituye dentro del listado de encuestas a validar por otra del mismo encuestador.

. Sospechosa: el objetivo de la validación telefónica que se realiza en el IEA es poder detectar las encuestas donde hay indicios de fraude. Una vez que se marca como sospechosa se inicia una investigación que nos lleve a discernir si la encuesta es fraude o no, recabando información del coordinador de provincia, del entrevistador y analizando la coherencia interna de la encuesta. Estos casos también se proponen para la validación in situ, como veremos posteriormente, en la que un equipo externo al IEA, vuelve a dirigirse a la vivienda para verificar si se ha realizado la encuesta. En el caso de que se determine que es fraudulenta se procede a su repetición.

El proceso de validación no termina aquí, ya que entonces quedaría aislado de los procesos de depuración e imputación. Aprovechando que tenemos el cuestionario separado en bloques, y a través del software TEIDE (Técnicas de Edición e Imputación de Datos Estadísticos)[1], vamos a seleccionar las encuestas y bloques en las que encontramos problemas de inconsistencia interna. Este procedimiento se realiza definiendo en dicho programa una serie de *edits* o reglas de consistencia entre variables que debe cumplir la encuesta (la depuración y el software utilizado al efecto se desarrollará con más detalle en el punto 6 del artículo).

Este procedimiento permite detectar problemas en una parte del cuestionario, de forma que sólo se procede a validar esa parte, y no el cuestionario completo. Es decir, si una encuesta tiene problemas de inconsistencia en 2 bloques, le realizamos una validación telefónica sobre esas partes a través de pequeños cuestionarios diseñados para ese fin. Si de una encuesta, observamos un número elevado de

[1] TEIDE (Técnicas de Edición e Imputación de Datos Estadísticos), es un programa informático desarrollado por la Universidad de La Laguna.

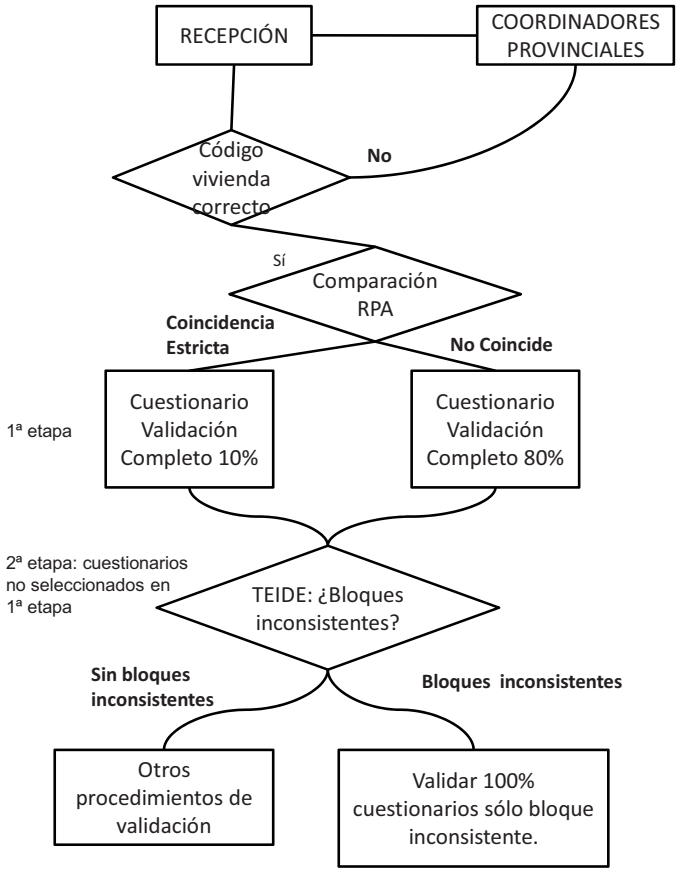
bloques inconsistentes, se seleccionaría para que se le realizase la validación completa y no por bloques sueltos.

Con este procedimiento, estamos conjugando dos procesos, el de validación y el de depuración.

Recapitulando, la validación telefónica se realiza a través de dos mecanismos:

1. Una vez seleccionado el caso, comparado con el RPA, se le realiza el cuestionario de validación completa de la encuesta, siguiendo unos porcentajes predeterminados.
2. Del resto de encuestas se comprueba la coherencia interna de los bloques en los que se divide el cuestionario, de forma que se validen aquellos que sean realmente inconsistentes, si éstos son un número elevado, se validará de forma completa.

Gráficamente, la secuencia de los procedimientos sería la siguiente:



Además de este procedimiento, también utilizamos estrategias clásicas de validación mediante el análisis de:

. Duración de las encuestas
. Horario de realización de las entrevistas.
. Seguimiento individualizado de entrevistadores o grupos de ellos por si actúan irregularmente en equipo.

Como se ha comentado anteriormente, la regla de decisión utilizada fijaba un porcentaje, a validar telefónicamente, de un 80% de los casos en los que no coincide la estructura del hogar del RPA con la estructura del hogar de la encuesta y un 10% en el caso de que sí coincidan. Esta notable diferencia en los porcentajes de ambos casos no es aleatoria. La Unidad de Encuesta del IEA se ha basado en los resultados de sus propias experiencias a la hora de proponer dichos porcentajes, ya que, como se esperaba, el porcentaje de encuestas sospechosas que se encuentran en una operación estadística de este tipo es mucho mayor en el caso de no coincidir la estructura de la vivienda encuestada con la que a priori está registrada en el RPA.

Todo esto se puede observar en las siguientes tablas, donde se muestran dos hechos:

1. Las primeras tablas se centran en calcular cuántas viviendas se clasifican como correctas y cuántas como incorrectas, en cada una de las encuestas realizadas por el IEA. Además en estas tablas se podrá observar la mejora que el propio RPA ha tenido a lo largo de los 4 años de funcionamiento de la Unidad de Encuesta.
2. Por otro lado, se corrobora las notables diferencias que se encuentran en la procedencia de una encuesta sospechosa, según venga de un tipo *Correcto_RPA* o de uno *Incorrecto_RPA*. Las encuestas sospechosas que se contabilizan son sólo las detectadas directamente por la encuesta de validación telefónica. Además se realizan otros análisis, para la detección de encuestas sospechosas, por encuestador u otro tipo que hacen que estos números aumenten.

Tabla 2.
Distribución de encuestas según comparación con el RPA.

Clasificación	EMV (%)	ESOC-07 (%)	ESOC-08 (%)
Correcto_RPA	68,1	66,4	75,39
Incorrecto RPA	31,9	33,6	24,61
Total	100	100	100

Se observa cómo el contraste entre el RPA y la muestra ha ido mejorando a lo largo de las distintas experiencias que se han tenido: en el año 2006-2007 (EMV y ESOC-07), el 67% de los casos de las viviendas encuestadas coincidían con la teórica del RPA, mientras que en el año 2008, ascendía a un 75,39%. Asimismo, otras de las pautas observadas es que prácticamente el total de las encuestas sospechosas son encuestas con Incorrecto_RPA (Por ejemplo en la ESOC-08 el 97,05% correspondía a viviendas con Incorecto_RPA; véase tabla 4), de ahí la importancia que tiene el contraste con el RPA en la validación.

Tabla 3.
Distribución de encuestas validadas por resultado de validación y comparación con RPA

Clasificación	Resultado Validación	EMV	ESOC-07	ESOC-08
Correcto_RPA	Validada/Llamada positiva	209, 27,83%	295, 30,67%	845, 40,43%
	No validable	37, 4 ,93%	49, 5,09%	137, 5 41%
	Sospechosa	2, 0,27%	2, 0,21%	1, 0,05%
Incorrecto RPA	Validada/Llamada positiva	397, 52,86%	475, 49,38%	869, 41,58%
	No validable	99, 13,18%	126, 13,10%	299, 10,96%
	Sospechosa	7, 0,93%	15, 1,56%	33, 1,58%
Total		751, 100%	962, 100%	2090, 100%

Tabla 4.
Distribución de encuestas sospechosas por comparación con RPA

Clasificación	EMV	ESOC-07	ESOC-08
Correcto RPA	2, 22,22%	2, 11,76%	1, 2,94%
Incorrecto_RPA	7, 77,78%	15, 88,24%	33, 97,06%
Total encuestas Sospechosas	9, 100%	17, 100%	34, 100%

4.2.- El cuestionario de validación

El diseño del cuestionario de validación es importante para que todo el procedimiento de validación tenga sentido. La realización de preguntas al azar del cuestionario no nos proporciona el tipo de información que nos permita comparar satisfactoriamente las distintas fuentes de información provenientes de la misma vivienda. Por esta razón dicho cuestionario se debe centrar en el uso conjunto de preguntas filtro y de aquellas en las que durante el período de tiempo entre ambos cuestionarios no vaya a cambiar la respuesta del entrevistado.

El proceso de validación de encuestas no termina aquí pues se tiene una ingente cantidad de información proveniente de los cuestionarios de validación, que debe ser analizada. Un primer análisis se produce en el cruce de la información de los dos cuestionarios que a priori son obtenidos de la misma vivienda, el de *validación* y el *original*. A través de un procedimiento automatizado se procede al análisis de las diferencias entre ambos, prestando especial atención a las diferencias existentes en una serie de preguntas filtro en las cuales el encuestador dispuesto a cometer irregularidades puede *ahorrarse* un tiempo considerable en la realización de la encuesta. Pongamos un ejemplo correspondiente a la ESOC-08. Si el encuestador directamente no pregunta a una vivienda si tiene o no aire acondicionado o calefacción se ahorra uno de los bloques más largos de la encuesta. Si se detecta un

comportamiento sistemático, en este tipo de casos, para ese encuestador concreto, se pide que se repita esa parte del cuestionario, como así ha ocurrido en alguna ocasión.

Otro análisis se produce en el número total de diferencias entre ambos cuestionarios. Si estas diferencias son notables, pasan a investigarse, por los cauces normales, es decir, contacto con coordinadores provinciales, encuestadores o envío de dicha encuesta al proceso de validación in situ.

5.-Validación *in situ*

Los apartados anteriores se han centrado en la validación de encuestas de forma telefónica, pero hay casos que no se pueden validar mediante estos procedimientos, bien porque no se dispone del teléfono, bien porque éste es incorrecto o bien porque se prefiera un contacto directo presencial con los entrevistados. Para estas encuestas se utilizará otro tipo de validación que es la validación in situ.

Además, hay otras situaciones que también pueden ser validadas y evaluadas mediante éste método. Un ejemplo claro es el correcto uso de las sustituciones de muestra por parte de los encuestadores. Es habitual que un encuestador que dispone de un listado de sustitutos muchas veces no haga las visitas obligadas a una vivienda o simplemente, por razones de cercanía y tiempo, acuda a la vivienda más próxima al lugar donde se encuentre, de forma que no se rija por el orden de sustitución previo del que dispone, indicando por tanto un motivo de sustitución ficticio e inexistente.

En gran parte de las encuestas realizadas desde la Unidad de Encuesta del IEA no se les ha proporcionado los listados de sustituciones a los encuestadores, de forma que se gestionaban desde la propia Unidad. Así ocurrió en las encuestas de Redes Familiares, Encuesta Mundial de Valores, Encuesta Social 2007: Una visión de Andalucía y Encuesta sobre las necesidades de formación y Cualificación en Andalucía. No ha ocurrido así en la Encuesta Social 2008: Hogares y Medio Ambiente en Andalucía, donde el listado sí ha sido proporcionado a los encuestadores.

Cada opción tiene sus ventajas e inconvenientes. Si desde el IEA se gestionan las sustituciones de muestra – generalmente las sustituciones se daban *en directo* a través de una línea 900 – el gasto humano por parte del IEA era superior y dificultaba los ritmos de trabajo a los entrevistadores. Por el contrario como ventaja principal, se controlaban en mayor medida las sustituciones.

Teniendo presente estos aspectos, la validación in situ se planteó con un doble objetivo: la detección de encuestas fraudulentas y la comprobación de las causas de sustitución alegadas por el entrevistador.

La validación in situ se realizó a través de una empresa externa a las Universidades y ajena al IEA. A esta empresa se le proporcionó las secciones que debía visitar. Estas secciones fueron seleccionadas atendiendo a varios criterios:

- Número de sustituciones elevado.
- Existencia de encuestas sospechosas detectadas en la validación telefónica.
- Presencia de un encuestador con encuestas sospechosas.

. Un porcentaje elevado de encuestas sin teléfono.

El trabajo de la empresa ha estado coordinado por el IEA de forma que se debía rellenar el siguiente formulario:

Imagen 1.
Formulario de validación In Situ

PARTE DE INSPECCION DE SECCIÓN

Encuesta Social 2008: Hogares y Medio Ambiente.
Provincia:
Municipio:
Sección:

Inspector/a:

ID Vivienda	Datos de recogida			Datos de la inspección				OBSERVACIONES
	Inciden. de Entrevista (1)	Fecha D	Fecha M	Resultado (3)	Fecha D	Fecha M	Inciden. de Inspección (2)	
1	VAC			H	26	6	VAC	PORTERO DICE QUE NO VIVE NADIE
2	AUS			SC	26	6	AUS	NO SE ENCUENTRA NADIE
5	ILO			H	26	6	VAC	SI SE LOCALIZA LA VIVIENDA PERO NO HAY NADIE EN ELLA
6	AUS			H	26	6	INC	MUJER MAYOR PERO DISPUESTA A COLABORAR, NO RECUERDA QUE LA HAYAN VISITADO NI CARTA ALGUNA
7	AUS			H	26	6	NEG	SE NIEGA A CONTESTAR
9	AUS			H	26	6	NEG	NO QUIERE CONTESTAR ESTÁ OCUPADO
11	AUS			SC	26	6	AUS	NO SE HA LOCALIZADO A NADIE EN LA VIVIENDA
13	NEG			H	26	6	DISTINTA	PODRÍA CONTESTAR LA ENCUESTA POR LA MAÑANA,
14	NEG			H	26	6	INC	PERSONA DE EDAD EVANZADA
17	NEG			H	26	6	NEG	NO QUIERE PARTICIPAR EN LA ENCUESTA, CONFIRMA LA VISITA DE ENCUESTADORES
19	NEG			H	26	6	DISTINTA	MUJER COLABORADORA, HORARIO DE TARDE PARA REALIZAR LA ENCUESTA, PERO NO RECUERDA QUE VINIERA NADIE
20	ILO			H	26	6	ILO	NO ENCUENTRO LA VIVIENDA

(1) INCIDENCIAS: **INA** (Vivienda inaccesible). **ILO** (Vivienda ilocalizable). **VAC** (Vivienda vacía). **OFI** (Vivienda destinada a otros fines). **NEG** (Negativa). **AUS** (Ausencia).
INC (Incapacidad para contestar). **SLA** (Vivienda seleccionada anteriormente). **OTR** (Otra causa de no colaboración) **SIN TLF** (vivienda colaboradora sin tlf)
(2) INCIDENCIAS: (1) + distinta+no saben
(3) Resultado de la Inspección: **H**= Hecha, **SC**= Sin contacto, **NEG**= negativa, **O**= Otro caso (Especificar en Observaciones)

En él se puede observar claramente el citado comportamiento en los encuestadores, ya que hay viviendas que han sido sustituidas por alguna causa justificada por el entrevistador, demostrándose después en la validación in situ que era una vivienda potencialmente encuestable.

6.-Depuración e Imputación en la Unidad de Encuesta

Como se ha visto anteriormente, los datos recogidos en las encuestas, se pueden ver afectados por dos tipos de errores, los errores de muestreo y los ajenos al muestreo. La corrección de estos últimos se convierte en una tarea imprescindible para mejorar la calidad de la investigación estadística. Por lo tanto, y debido a la imposibilidad de reanudar los trabajos de campo dado el coste y el tiempo que ello conlleva, la depuración de los datos se convierte en una tarea necesaria antes de comenzar el procesamiento de datos de la encuesta.

La detección de errores requiere que previamente se definan las situaciones erróneas o sospechosas, tarea que corresponde a los estadísticos expertos en el tema objeto de la investigación. La definición de posibles situaciones erróneas se realiza habitualmente por medio de los denominados *edits* o reglas de coherencia, ya citados en otros puntos. Los edits especifican restricciones a los valores individuales de las variables (edits de validación, casi todos implementados en la PDA y por lo tanto de obligado cumplimiento a priori) o de conjunto de variables (edits de consistencia). La detección de errores se realiza enfrentando todos los registros a depurar con el conjunto de edits especificados. Un registro se considera erróneo si no cumple la condición especificada por un edit de conflicto. Mediante los edits se especifican:

. Situaciones imposibles.
. Situaciones improbables.
. Restricciones contables.
. Outliers.
. Control de flujo de respuesta del cuestionario.

Resulta muy útil utilizar herramientas automáticas que ayuden en este proceso tan complejo. En el IEA se utiliza el software TEIDE (Técnicas de Edición e Imputación de Datos Estadísticos), programa informático desarrollado por la Universidad de La Laguna, el cual, mediante un conjunto de relaciones entre variables, decide si un registro es correcto y en caso contrario qué variables hay que modificar para que satisfaga todas las relaciones, de forma que el número de variables modificadas en el registro sea el menor posible.

Una vez terminado el proceso de análisis de los edits, hemos de pasar a la "depuración manual", pues hay casos que sólo pueden arreglarse de esta manera. Para ello se ha utilizado principalmente información de fuentes auxiliares, como es el RPA, e incluso un pequeño porcentaje de respuestas telefónicas de ciertas variables socio-demográficas.

Una vez concluida esta fase, todos aquellos registros que queden con alguna incoherencia tendrán que pasar a la depuración automática o imputación. El método de imputación estadística que emplea TEIDE es el de registro donante de Fellegi-Holt. En él, se tienen un conjunto de registros totalmente correctos que servirán de "donantes" para otros registros con alguna incorrección. Midiendo las distancias en determinadas variables entre los registros correctos e incorrectos, el donante para cada registro a imputar será determinado por aquel registro correcto que minimice dicha distancia. Los valores de los microdatos del registro "donante" serán utilizados para sustituir los valores a imputar en el registro incorrecto.

7.-Conclusiones

Como se sabe y se ha recogido previamente, la realización de una encuesta ha de estar sometida en toda su ejecución a exigencias de calidad que respalden con la misma, los resultados que se obtengan. En este sentido se han descrito, los criterios y exigencias a los que se ha sometido el trabajo de campo realizado para la ejecución de la Encuesta Social 2008, y esa experiencia nos permite concluir, que los criterios a exigir, durante la obtención de la información han de ser fijados antes del inicio de

los trabajos de campo, de forma que la aplicación de los criterios ha de ser evaluada desde el mismo instante en que empiezan a recogerse los datos, a través de procedimientos de control, que validen la información recogida por los entrevistadores. Paralelamente, se tendrán que realizar los oportunos procesos de depuración de los datos, para corregir las inconsistencias que se produzcan.

8.-Bibliografía

Basulto Santos, J. et al. (1999). Imputación Automática de Datos. Instituto de Estadística de Andalucía. Sevilla.

Biemer, P.P. and Lyberg, L. E. (2003). Introduction to Survey Quality: Wiley Series in Survey Methodology, Wiley, Hoboken.

Biemer, P. (2009). Measurement errors in sample surveys. In D. Pfeffermann, & C.R. Rao (Ed.), Handbook of Statistics, Vol. 29A - Sample Surveys: Design, Methods, and Applications. New York, NY: Elsevier.

Bravo, M. S. (1991). Speer y Geis: Dos sistemas para la depuración de datos cuantitativos Comentarios. Revista Estadística Española. 127. 191-242.

Cea D'Ancona M. (2005). La senda tortuosa de la calidad de la encuesta. Revista Española de Investigaciones Sociológicas, 111, 75-103

Fellegi, I.P., Holt, D., (1976). A systematic approach to automatic edit and imputation. Journal of the American Statistical Association 71, 17-35.

Granquist, L. (1984). On the role of editing. Statistisk Tidskrift, 2, 105-118

Kish, L. (1965). Survey Sampling, Wiley, New York.

Malhotra Naresh K. (2004). Investigación de mercados: un enfoque práctico. Prentice-Hall (4ª Ed)

Salazar González, J. and Delgado Quintero, S. (2004). Técnicas de Edición e Imputación de datos estadísticos. http://webpages.ull.es/users/istac/TEIDE/pdf/findecarrerateide.pdf

Villán Criado, I y Bravo Cabria, M. S. (1990). Procedimiento de depuración de datos estadísticos. Instituto Vasco de Estadística. Seminario Internacional de Estadística En Euskadi.