

El trabajo de campo de la Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006

Institut d'Estudis Regionals i Metropolitans
de Barcelona



METODOLOGIES I RECERQUES és una publicació que edita i coordina l'Institut d'Estudis Regionals i Metropolitans de Barcelona (direcció: Carme Miralles-Guasch).

Les opinions expressades en els articles que conté la publicació són d'exclusiva responsabilitat de les persones que n'assumeixen l'autoria.

Coordinació de l'edició

Isabel Clos (IERMB)
José Luis Flores (IERMB)
Sira Ponsa

Correcció de textos

Paloma Calvo

Distribució

Publicacions IERMB
93 581 44 30 - iermb@uab.cat
93 223 42 14 - iermb@amb.cat

Per a més informació, consulteu www.iermb.uab.es

Disseny

Oficina de disseny de l'AMB

Maquetació i preimpresió

Estudi Gràfic Pedregosa (EGP)

Impressió

Gradisa

Barcelona, març de 2008

DL: B-28094-2008

ISSN: 2013-0309

ISBN: 978-84-88068-92-7

© Institut d'Estudis Regionals i Metropolitans de Barcelona

Pròleg

Dels dels seus inicis, l'IERMB ha dedicat esforços a donar a conèixer la seva producció científica entre la comunitat universitària i entre els professionals de l'Administració i de la gestió públiques. També des del començament, els investigadors que han col·laborat amb l'IERMB han desenvolupat metodologies i tècniques d'anàlisi vinculades a la realització de grans enquestes a la població.

La publicació que teniu a les mans és el primer volum de la nova col·lecció **METODOLOGIES I RECERQUES**, que volem dedicar principalment a divulgar, a donar a conèixer les metodologies que s'apliquen en les diferents recerques que l'IERMB desenvolupa. En algunes ocasions són metodologies innovadores –quasi diria que experimentals– que s'han anat perfeccionant al llarg de diferents edicions d'un mateix estudi, d'una mateixa enquesta. Altres vegades són el resultat de l'adaptació de metodologies ja testades i experimentades, però aplicades a nous conceptes o noves variables.

La col·lecció, en la qual s'utilitzaran indistintament el català i el castellà, vol acollir també els resultats de petites recerques que, per la seva temàtica, tenen més difícil la difusió en altres tipus de publicacions. Serà doncs una porta més per conèixer la producció científica que es genera en el si de l'IERMB.

Esperem que aquesta iniciativa que tot just comença tingui un grau elevat d'acceptació i contribueixi al coneixement generalitzat de les qüestions de metodologia, una part del procés de recerca que sovint queda en segon terme, però que és fonamental perquè els estudis obtinguin resultats òptims.

Carme Miralles-Guasch
Directora IERMB

Sumario

01

ORGANIZACIÓN GENERAL DE LA ENCUESTA

Elena Sintes (Instituto de Estudios Regionales y Metropolitanos de Barcelona)

9

02

LA CONSTRUCCIÓN DE LA MUESTRA

Pedro López y Carlos Lozares (Centro de estudios sociológicos sobre la vida cotidiana y el trabajo, Departamento de Sociología, Universidad Autónoma de Barcelona)

17

03

ORGANIZACIÓN DEL TRABAJO DE CAMPO

Meritxell Perramón (Instituto de Estudios Regionales y Metropolitanos de Barcelona)

39

04

ESTADÍSTICAS DE SEGUIMIENTO COMO DISPOSITIVO DE CALIDAD DEL TRABAJO DE CAMPO EN ESTUDIOS POR ENCUESTA

Sergio Porcel (Instituto de Estudios Regionales y Metropolitanos de Barcelona)

51

05

CREACIÓN Y DEPURACIÓN DE LA BASE DE DATOS

José Luis Flores (Instituto de Estudios Regionales y Metropolitanos de Barcelona)

67

Presentación

El trabajo de campo de la *Encuesta de Condiciones de Vida y Hábitos de la Población* ha crecido en complejidad en cada una de las cinco ediciones que se han desarrollado hasta la fecha, tanto por el aumento del número de variables que se han incluido como por la ampliación del ámbito territorial de aplicación. Una de las preocupaciones de los gestores de la *Encuesta* al abordar cada una de las sucesivas ediciones ha sido introducir los mínimos cambios metodológicos necesarios para garantizar la homogeneidad de la serie histórica.

En la publicación que tenéis entre manos se aborda todo el proceso de confección de la *Encuesta*, desde la organización general que se plantean las instituciones promotoras de la operación estadística hasta la obtención de una base de datos lista para ser utilizada en los sucesivos análisis de resultados. El contenido de la revista se estructura en cinco artículos que responden a las cinco etapas principales de un proceso de trabajo de campo: planteamiento general, confección de la muestra, organización de equipos de campo, control de calidad y construcción de la base de datos.

En su artículo, Elena Sintés esboza las características más generales de la *Encuesta de Condiciones de Vida y Hábitos de la Población* —evolución histórica, metodología y fases del trabajo de campo— e insiste en la búsqueda de la calidad como estandarte de todo el proceso de recogida de información mediante el trabajo de campo.

El segundo artículo, de Pedro López y Carlos Lozares, se centra en las distintas etapas que son necesarias para construir la muestra de la edición 2006 de la *Encuesta* y señala diferencias y similitudes con las cuatro ediciones anteriores. Destaca todos los pasos dados para garantizar la comparabilidad de resultados entre los diferentes años.

A continuación, Meritxell Perramón describe en su texto cómo se afrontó el reto de realizar más de 10.000 entrevistas personales repartidas por toda Cataluña y cuáles fueron los mecanismos introducidos en el proceso de encuestación para poder garantizar la calidad y fiabilidad de los resultados.

De forma más teórica, pero partiendo de la realización del trabajo de campo de la *Encuesta*, Sergio Porcel subraya el papel y la necesidad de realizar estadísticas de seguimiento del proceso para poder controlar en todo momento los índices de producción de encuestas y el nivel de calidad de la muestra que se va obteniendo.

En el quinto artículo, José Luis Flores expone las fases de codificación, grabación y depuración de datos, que permiten pasar de la recogida de información en cuestionarios en soporte papel a la construcción de la base de datos definitiva.

01 ORGANIZACIÓN GENERAL DE LA ENCUESTA

Elena Sintés (Instituto de Estudios Regionales y Metropolitanos de Barcelona)

Introducción

1. Inicios e historia de la *Encuesta*

2. Metodología

3. Fases y desarrollo del trabajo de campo

Introducción

La *Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006*, constituye la última edición de un proyecto de análisis de la realidad social que se realiza quinquenalmente desde el año 1985. Se trata de una encuesta concebida como instrumento de recogida de información periódica sobre diversos ámbitos de la vida cotidiana de la población. En este sentido, su finalidad es aportar información sobre las dinámicas sociales y territoriales y, gracias a la serie de datos disponible en la actualidad, también sobre las transformaciones que tienen lugar en estos ámbitos. En sus veinte años de historia, la *Encuesta de condiciones de vida y hábitos de la población* se ha consolidado como una de las principales fuentes de información de la región de Barcelona. La quinta edición ofrece, por primera vez, información precisa sobre el conjunto de la sociedad catalana.

El Instituto de Estudios Regionales y Metropolitanos de Barcelona (IERMB) es quien lleva a cabo, desde la primera edición, las tareas relacionadas con la dirección metodológica, así como el trabajo de campo, el tratamiento de datos y el análisis de la *Encuesta*, lo que ha contribuido al mantenimiento de los criterios metodológicos y a la comparabilidad de resultados de las distintas ediciones.

El período transcurrido entre 1985 y 2006 ha consolidado una metodología de trabajo de campo propia, aunque en continua transformación. Los investigadores y equipos que han desarrollado el proyecto a lo largo de estos veinte años se han enfrentado a los cambios, problemas y retos derivados de las transformaciones sociales y, también, a la evolución en los métodos de investigación.

Desde 1985, la investigación social se ha transformado enormemente. La utilización de la técnica de encuesta como instrumento de conocimiento se ha multiplicado y expandido, tanto en número de encuestas realizadas en la actualidad como en la diversidad de disciplinas que la utilizan habitualmente, e incluso ha crecido su repercusión en los medios de comunicación.

¿Qué tiene de especial, pues, la *Encuesta de condiciones de vida y hábitos de la población* para dar pie a esta publicación? Básicamente, la búsqueda permanente de la calidad en la información obtenida y la experiencia acumulada en veinte años. Los investigadores sociales conocen bien la importancia del trabajo de campo en la obtención de datos fiables: como cualquier otra encuesta tiene limitaciones y condicionantes, pero hemos intentado minimizarlos para conseguir así garantizar la máxima calidad final de los datos y que éstos reflejen de forma fidedigna la realidad social en Cataluña.

El presente artículo se divide en tres partes. En primer lugar se hace un breve repaso de la historia de la *Encuesta*. El segundo apartado está dedicado a los aspectos metodológicos de la edición de 2006 y, finalmente, el tercer apartado presenta, a modo de resumen, las fases y el desarrollo del trabajo de campo.

1. Inicios e historia de la Encuesta

La *Encuesta de condiciones de vida y hábitos de la población* es una encuesta generalista sobre las condiciones de vida y hábitos de la población, cuyo principal objetivo es el de suministrar datos y elementos de análisis que permitan el conocimiento y el seguimiento de las dinámicas sociales y territoriales que se generan en el territorio, en un primer momento en el entorno metropolitano de Barcelona para después, como veremos, estudiar también estas dinámicas en el conjunto de Cataluña. Los datos, análisis y conclusiones que se obtienen han sido y son utilizados para el diseño de políticas públicas, y sirven también de base como fuente de información para los científicos sociales y los investigadores.

En 1985, desde la Corporación Metropolitana de Barcelona (CMB¹) se planteó la necesidad de disponer de datos precisos sobre la realidad metropolitana; y es que la información estadística disponible hasta ese momento no proporcionaba el nivel de conocimiento que en aquel momento se requería para la acción política local. En este sentido, la *Encuesta* nace con una doble voluntad: en primer lugar, producir la información que la gestión del área metropolitana necesitaba y, en segundo lugar, construir una serie temporal que permita conocer la evolución de la realidad social de dicho territorio.

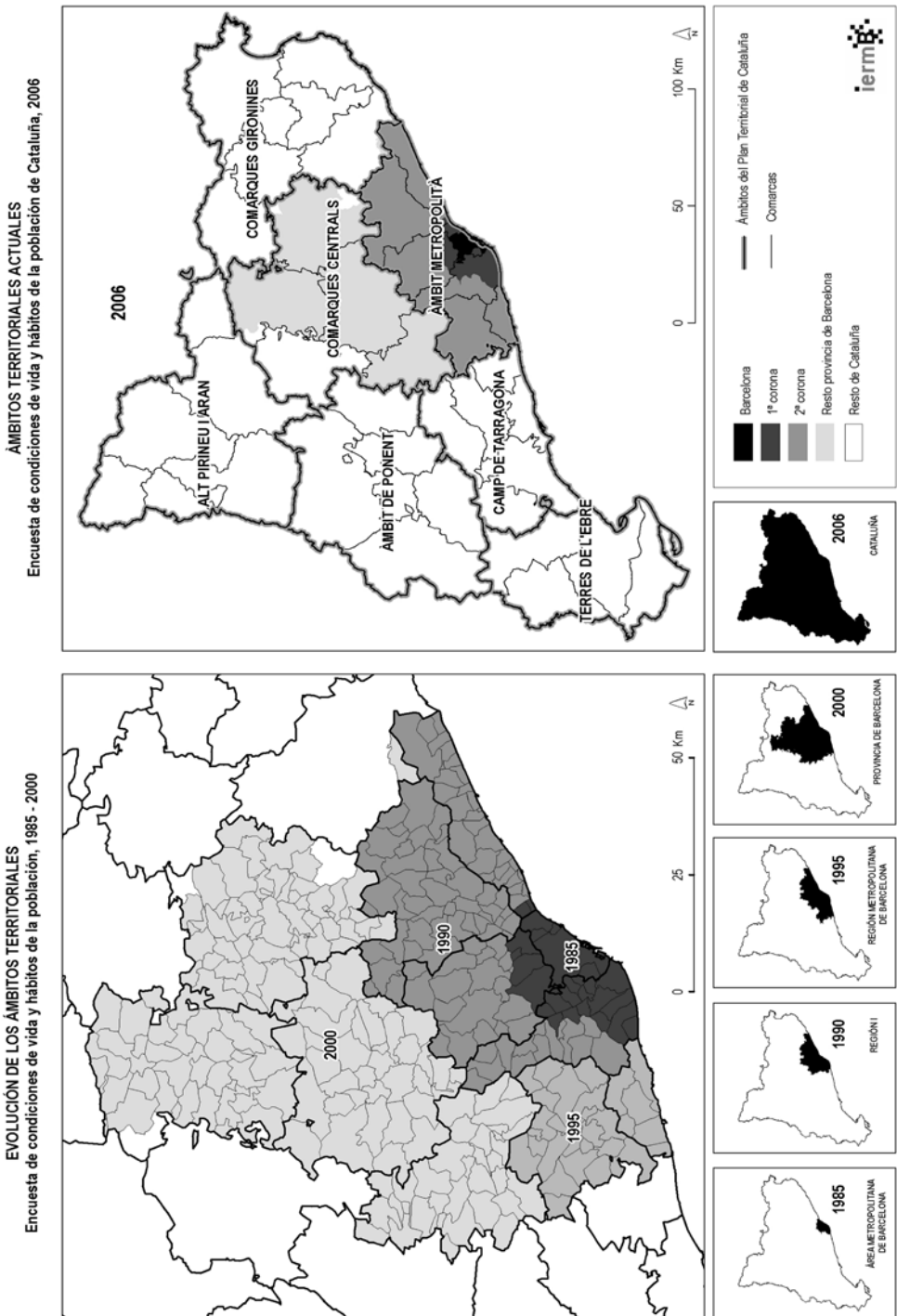
Uno de los aspectos más relevantes de la evolución de la encuesta es el mantenimiento de los objetivos y los criterios de investigación. En el decurso de las cinco ediciones se han incorporado algunos cambios técnicos y metodológicos que no han afectado a la comparabilidad de los resultados. Aun así, el cambio más importante de la encuesta ha sido la ampliación progresiva de su ámbito territorial de estudio. El territorio de la primera edición de 1985 fue la ciudad de Barcelona y los 26 municipios que integraban la antigua Corporación Metropolitana de Barcelona. En 1990 la muestra se expande para alcanzar el conjunto de comarcas que conforman la Región I (Baix Llobregat, Barcelonès, Maresme, Vallès Occidental y Vallès Oriental). En la edición de 1995 se añaden las comarcas del Alt Penedès y el Garraf, con las que se llega a toda la región metropolitana de Barcelona. En el año 2000 se cubre toda la provincia de Barcelona. Y, finalmente, con la presente edición, la *Encuesta* comprende, por primera vez, todo el conjunto del territorio catalán, por lo que puede considerarse la ampliación territorial más ambiciosa del proyecto.

Todas estas modificaciones territoriales mantienen una estrecha relación con el ámbito territorial competencial de las instituciones que a lo largo de estos años han financiado la *Encuesta*. El proyecto fue impulsado y financiado primeramente por la antigua Corporación Metropolitana de Barcelona (edición de 1985) y, posteriormente (ediciones 1990, 1995 y 2000), por la Mancomunidad de Municipios del Área Metropolitana de Barcelona (MMAMB) y la Diputación de Barcelona. La última ampliación territorial obedece a la incorporación en la presente edición (año 2006) de la Generalitat de Cataluña, a través del Instituto de Estadística de Cataluña (Idescat).

Estos cambios, que no han alterado la comparabilidad de los resultados ni las características técnicas generales de la *Encuesta*, se han producido siempre, como hemos mencionado, por agregación de territorios a la muestra, lo que ha supuesto también un incremento considerable de las personas entrevistadas. Así, de las 4.912 entrevistas que se realizaron en la primera edición de 1985, se ha pasado a una muestra de 10.398 individuos para la actual edición del proyecto, lo que la convierte en una de las mayores encuestas presenciales realizadas en el territorio catalán.

¹ La CMB era el organismo administrativo que agrupaba Barcelona y 26 municipios de su entorno. Nació en 1974 y fue suprimida por las leyes territoriales de 1987.

Mapa 1. Ámbitos territoriales de las diferentes ediciones de la ECVHP



La *Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006* es una operación estadística que se incluye en el Programa anual de actuaciones estadísticas 2005 de la Generalitat de Cataluña, a la vez que se inserta en el Plan Estadístico de Cataluña 2005–2008.

2. Metodología

Como ya hemos comentado, las actuales características técnicas y metodológicas de la *Encuesta* prácticamente no han variado respecto a ediciones anteriores. El proyecto ha mantenido los criterios de investigación y una gran similitud en el diseño muestral y los cuestionarios administrados a la población encuestada. El aspecto que ha variado de un modo más significativo ha sido el alcance territorial del estudio, que se ha ido ampliando en cada edición.

Objetivos

Los objetivos de la *Encuesta* se pueden resumir en los siguientes enunciados:

- Conocer y analizar las condiciones de vida y los hábitos de la población, las tendencias evolutivas de las formas de vida y los fenómenos sociales y la identificación de los factores que intervienen en los cambios sociales.
- Conocer y analizar las diferencias en las condiciones de vida y los hábitos en el territorio y la identificación de los rasgos territoriales distintivos.
- Conocer y analizar los grupos sociales y sus características, identificando los factores que condicionan la formación de estos grupos y analizando los tipos de desigualdades existentes entre ellos, así como las tendencias de convergencia o divergencia que se observan en sus formas de vida.

Unidades de análisis

El universo estadístico lo conforman la totalidad de individuos de 16 y más años que residen en una vivienda familiar en el territorio de Cataluña.

A efectos de representatividad estadística, la unidad de recogida de información de la *Encuesta* es el individuo, aunque la información que recoge la estructura del cuestionario permite conocer algunas características de cada uno de los miembros que residen en el hogar.

Ámbito territorial

El ámbito territorial está constituido por el territorio de Cataluña. La muestra recogida tiene representatividad estadística para un nivel de error aceptable y para distintos niveles territoriales, que pueden ser tratados como submuestras independientes.

El diseño muestral que se utiliza en la *Encuesta* admite la desagregación de la muestra en unidades territoriales más reducidas, lo cual permite la realización de un análisis más ajustado de estas áreas, así como la comparación con el resto de ediciones. Estos territorios son: Cataluña, Ámbitos del Plan Territorial de Cataluña (ámbito metropolitano, Comarques Gironines, Camp de Tarragona, Terres de l'Ebre, Àmbit de Ponent, Comarques Centrals y Alt Pirineu i Aran), Comarcas de la provincia de Barcelona, coronas metropolitanas y Barcelona ciudad.

Ámbitos temáticos: el cuestionario

La *Encuesta de condiciones de vida y hábitos de la población* se basa en la realización de consultas quinquenales a las personas residentes en el territorio objeto de estudio mediante un extenso y completo cuestionario que aborda casi todos los elementos que constituyen las formas de vida de la población. La duración media por entrevista está calculada en unos 50 minutos aproximadamente.

La estructura interna del cuestionario hace que convivan preguntas referidas al contexto del hogar, algunas de las cuales recogen información de cada una de las personas que lo integran, con

preguntas referidas concretamente al individuo seleccionado. En total se trata de 196 preguntas, mayoritariamente precodificadas, que podemos agrupar en los siguientes apartados o ámbitos temáticos:

- Estructura del hogar (información de todos los miembros del hogar).
- Formas de convivencia (información de todos los miembros del hogar).
- Procedencia geográfica (información de todos los miembros del hogar).
- Cambios residenciales (información de la persona entrevistada).
- Adscripción y transmisión lingüística (información del hogar y de la persona entrevistada).
- Niveles educativos (información de todos los miembros del hogar).
- Situación en el mercado de trabajo (información de todos los miembros del hogar).
- Condiciones laborales y profesionales (información de la persona entrevistada).
- Trabajo doméstico (información del hogar y de la persona entrevistada).
- Salud (información de la persona entrevistada).
- Nivel de renta y de recursos (información del hogar y de la persona entrevistada).
- Hábitos de consumo (información del hogar y de la persona entrevistada).
- Características y condiciones de la vivienda (información del hogar).
- Características del equipamiento doméstico (información del hogar).
- Uso del territorio (información de la persona entrevistada).
- Percepción e imagen del territorio cercano (información de la persona entrevistada).
- Hábitos culturales y actividades de ocio (información de la persona entrevistada).
- Relaciones sociales y nivel de asociacionismo (información de la persona entrevistada).
- Adscripción ideológica (información de la persona entrevistada).
- Movilidad (información de la persona entrevistada).

Los datos resultantes son mayoritariamente de tipo objetivo y tienen un marcado carácter estructural. Por otro lado, se trata de una información que tiene un valor esencial, puesto que complementa las fuentes de información existentes sobre el territorio de análisis y, en muchos casos, constituye incluso la única información disponible.

En cada edición se ha revisado el cuestionario con la intención de incorporar a éste nuevos aspectos sociales que no habían estado presentes en la *Encuesta* anterior, valorando en cada caso la relevancia que tiene en el proyecto la comparabilidad de los datos. Así, en la última edición destaca la introducción de preguntas relativas a la inmigración, al uso de nuevas tecnologías y al estado de salud de los entrevistados.

Muestra

Se han realizado un total de 10.398 entrevistas en todo el territorio de Cataluña. Con este tamaño de muestra, si se considera un nivel de confianza del 95,5%, el error muestral es de 0,7% para datos globales.

La muestra, realizada desde la primera edición por el mismo equipo del Departamento de Sociología de la Universidad Autónoma de Barcelona, ha sido diseñada con la técnica de muestreo aleatorio estratificado², a partir de los datos del Censo de Población de 2001. La extracción de la muestra se ha realizado a partir del padrón continuo de población del año 2005.

3. Fases y desarrollo del trabajo de campo

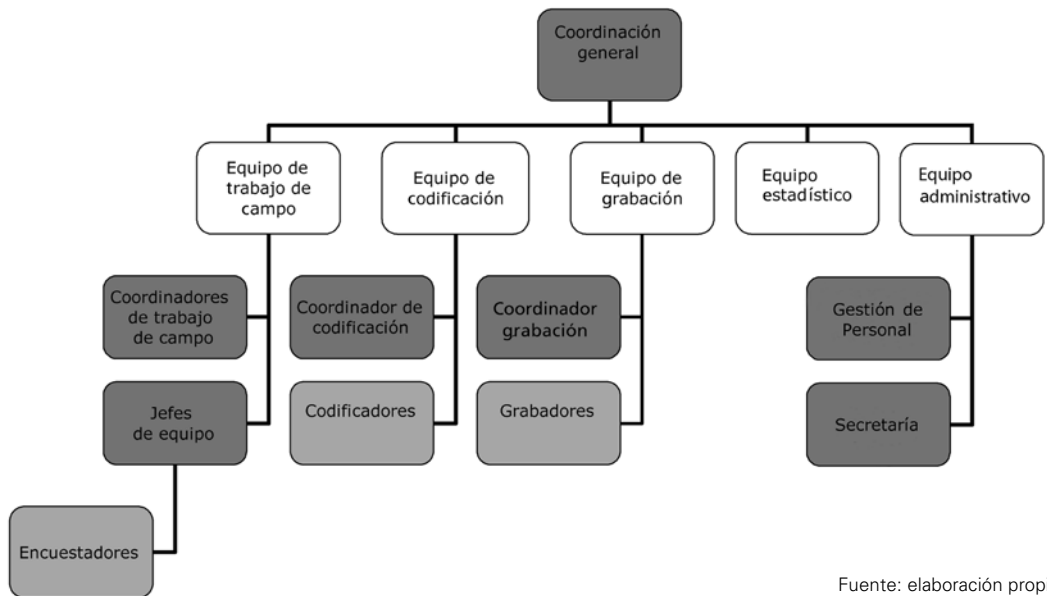
La *Encuesta* consta de diversas fases claramente diferenciadas, aunque perfectamente interrelacionadas. Después de la fase de diseño del proyecto —con la selección de la muestra y la

² En el segundo artículo de la presente publicación, los autores de la muestra realizan una detallada explicación sobre las características y el proceso de construcción de la misma.

definición del cuestionario— se inicia la fase de recogida de información y tratamiento de datos. Dicho período, que abarca el trabajo de campo propiamente dicho, empieza con la realización de las entrevistas personales en los domicilios de los encuestados y acaba con la creación de la base de datos. El equipo que trabaja en esta fase comprende un número elevado de técnicos, entre coordinadores del trabajo de campo, jefes de equipo, de codificación y de depuración, estadísticos y secretaría técnica, además de los encuestadores (alrededor de 70 personas permanentemente distribuidas por el territorio catalán), los codificadores y los grabadores de datos.

En total, cerca de un centenar de personas hacen posible que a lo largo de esta fase, de aproximadamente un año de duración, se puedan desarrollar las tareas de formación de los encuestadores, realización de las entrevistas, revisión de los cuestionarios, tratamiento de la información (codificación, grabación de los datos y depuración de los mismos), así como la estructuración de la base de datos, de manera paralela y con una calidad óptima gracias a que a lo largo de este proceso se establecen diversos controles de calidad. Para poder llevar a acabo todo este proceso se constituyen cinco equipos de trabajo, en dependencia directa de la coordinación general:

Figura 1. Esquema de organización de la recogida de información y tratamiento de datos



Fuente: elaboración propia

- Equipo de trabajo de campo, que se encarga de la realización de las entrevistas y está formado por encuestadores, jefes de equipo y coordinadores de trabajo de campo. Su función principal es la formación de los encuestadores, la organización del trabajo de campo y el control de la producción, así como la revisión de todos los cuestionarios y el control telefónico de los mismos. Asimismo, es el equipo encargado del control de la muestra, las suplencias y los motivos de las mismas.
- Equipo de codificación, que revisa nuevamente la coherencia interna de los datos de cada cuestionario. Su función es resolver las posibles incidencias e incongruencias que puedan hallarse en los cuestionarios, así como transformar en códigos todos los datos para que puedan ser introducidos en una base de datos.
- Equipo de grabación, encargado de introducir la información codificada de los cuestionarios en una base de datos, utilizando el programa SPSS.

- Equipo estadístico, que controla el funcionamiento de todo el proceso y la calidad de la muestra y de la información recogida a partir de un programa de depuración/inconsistencias, para poder construir una base de datos definitiva.
- Equipo administrativo, encargado de todas las tareas de soporte para garantizar el correcto funcionamiento del resto de equipos (tareas de contratación, de gestión de personal, secretaría).

La tercera fase de la encuesta se inicia en el momento en que se dispone de una base de datos definitiva que permitirá llevar a cabo un amplio abanico de tratamiento y explotación de la información recogida.

02 LA CONSTRUCCIÓN DE LA MUESTRA

Pedro López y Carlos Lozares (Centro de estudios sociológicos sobre la vida cotidiana y el trabajo, Departamento de Sociología, Universidad Autónoma de Barcelona)

Introducción

1. Construcción de los estratos

- 1.1. El análisis factorial y la identificación de las variables—criterio de los estratos
- 1.2. El análisis de clasificación o la construcción de los estratos de la muestra estratificada

2. Afijación de la muestra y cuotas muestrales

- 2.1. Tamaño de la muestra y afijación
- 2.2. Cuotas muestrales
- 2.3. Distribución final de la muestra

3. Ponderación a posteriori de la muestra y elevación

4. Aplicación de los resultados de la muestra

Introducción

Este artículo tiene por objeto exponer las principales características del diseño y del proceso de construcción de la muestra estratificada de la *Encuesta de condiciones de vida y hábitos de la población de Cataluña* en su edición de 2006, haciendo hincapié en las novedades que se han incorporado. La metodología seguida en el diseño de la muestra a lo largo de las cinco ediciones de la *Encuesta* conserva un mismo planteamiento básico, que se ha mantenido en el tiempo, si bien en cada nueva edición se han ido introduciendo algunos cambios como resultado de las exigencias de los análisis territoriales que se han ido planteando en cada momento¹. La presente edición es la que incorpora mayores novedades, sobre todo por la ampliación del ámbito territorial al conjunto de Cataluña y por la introducción de cambios en el procedimiento de determinación de las cuotas de muestra.

¹ Una exposición detallada del diseño y del proceso de construcción de la muestra, para la edición de la Encuesta de 1990, se puede consultar en Lozares y López (1990). Una versión reducida de este documento fue publicada en Lozares y López (1991b). Posteriormente, tomando como base esta misma muestra, se consideró el tratamiento de subpoblaciones con la finalidad de construir zonas sociales (Lozares, y Domínguez, 1996). El proceso de construcción de la edición de 1995 se presentó en forma de artículo de revista (López, Domínguez, y Lozares, 2000). Finalmente, con respecto a la edición del año 2000, se puede consultar el apartado metodológico de los textos publicados en dicha edición de la encuesta (Giner *et al.*, 2002), así como el análisis de las zonas sociales y la cartografía social que se obtiene como producto de la propia construcción de la muestra (Lozares, López, y Flores, 2003).

En primer lugar se analizan los rasgos generales de la muestra, para entrar, en los apartados siguientes, en las cuestiones específicas relativas a la construcción de los estratos, la afijación de la muestra, la asignación de cuotas y la ponderación y elevación de la muestra, para acabar con algunas consideraciones que van más allá de la simple aplicación de la muestra.

En el diseño de la muestra se establece como objetivo general la extracción de una muestra aleatoria estratificada representativa de la población de Cataluña. La elección del procedimiento de muestreo estratificado se justifica básicamente por criterios de precisión frente a otros métodos y por la heterogeneidad social que caracteriza a la población objeto de estudio.

Por otra parte, los contenidos del estudio de la encuesta recogen una gran diversidad de temáticas sociales de interés. El objetivo consiste precisamente en poder expresar mejor la variedad de la realidad social de Cataluña respecto a las condiciones de vida y hábitos de la población. Dichos contenidos y objetivos obligan, en la construcción de la muestra estratificada, a tomar en consideración como variables-criterio de estratificación aquellas que den mejor cuenta de dicha multiplicidad social. Esta información poblacional primaria y diversa proviene del censo de población, a través de sus indicadores. Una vez elegidas las variables más pertinentes del Censo que representan la máxima diversidad social, se lleva a cabo un ejercicio de reducción y estructuración de las mismas a fin de identificar y elegir las variables-criterio para construir los estratos de la muestra. Así se garantiza la presencia en la muestra de estratos que expresan las diferencias poblacionales y con unos contenidos que guardan gran correlación con los de la *Encuesta*.

Así pues, el punto de partida es la información que proviene del censo de población del año 2001. Para proceder a la construcción de los estratos se agrupan los individuos en función de las secciones censales donde residen. De esta forma se dispone, para el conjunto de indicadores o variables seleccionados, del número de personas que, en cada sección, poseen la característica dada por el indicador y, en consecuencia, se dispone también del porcentaje que representa dicha variable sobre el total de la población de la sección censal. A partir de esta información se identifican, primero, las variables-criterio y, en segundo lugar, se construyen los estratos mediante métodos de análisis factorial de componentes principales y de análisis de clasificación automática.

Con los estratos construidos se procede a la afijación de un tamaño de muestra predeterminado a cada estrato según el criterio óptimo de Neyman. Realizada dicha atribución en los estratos, se asignan las cuotas de muestra a cada una de las secciones censales de cada estrato según su población. Se lleva a cabo, *a posteriori*, una subestratificación según el territorio y el hábitat, con el objetivo de conseguir una concentración territorial que facilite el trabajo de campo de la encuesta. Finalmente, las unidades muestrales se extraen de forma aleatoria en cada sección censal a partir del marco de la muestra que, en este caso, ha sido el conjunto de personas registradas en el Padrón de Habitantes actualizado a enero de 2005.

En comparación con ediciones anteriores de la *Encuesta* (Tabla 1), la construcción de la muestra presenta algunos cambios que se resumen en los siguientes aspectos:

- a) El ámbito territorial se extiende al conjunto de Cataluña. El ámbito poblacional contemplado se amplía también para considerar el de la población catalana de 16 y más años. Así, el universo poblacional se define como el conjunto de personas de 16 y más años que residen en Cataluña: un total de 5.408.264 personas según el censo de población de 2001, siendo la población total catalana registrada en aquel momento de 6.343.110².

2 Hay que tener presente la infravaloración de este recuento, ya que se han producido importantes movimientos inmigratorios en los últimos años que motivan que, en la actualidad, la población catalana total supere los siete millones de personas. En el momento de extraer la muestra se considera el padrón de habitantes actualizado. *A posteriori*, con la ponderación de la muestra, se tienen en cuenta estas variaciones y distribuciones diferenciadas de la población de las secciones censales en el territorio.

- b) El proceso de construcción y de análisis se realiza a partir de la información contenida en el censo de población del año 2001, modificando y ampliando el conjunto de indicadores–variables utilizados para la muestra.
- c) Se ha procedido a una subestratificación de la muestra en función del territorio, con una selección aleatoria de secciones censales y una redistribución de las cuotas finales de muestra para satisfacer el requisito de concentración territorial del trabajo de campo.
- d) El marco muestral de extracción deja de ser el censo electoral para considerar la actualización del padrón de habitantes.
- e) El tamaño de muestra se amplía para facilitar los análisis independientes de dominios de estudio o subpoblaciones específicas, territoriales en particular. Para garantizar la representatividad de algunos territorios, además, se ha añadido un número de encuestas adicional que supera el tamaño de muestra fijado para datos globales.
- f) La ponderación de la muestra se ha ajustado a los nuevos requisitos del diseño y se ha aplicado por primera vez el factor de elevación para expresar los datos muestrales en términos de valores absolutos poblacionales.

Tabla 1. Comparación de las muestras de la ECVHP

Edición de la ECVHP					
	1985	1990	1995	2000	2006
Ámbito territorial	Área Metropolitana de Barcelona	Región I	Región Metropolitana	Provincia de Barcelona	Cataluña
Municipios	27	129	162	311	946
Superficie	476 km ²	2.459 km ²	3.235 km ²	7.718 km ²	32.106 km ²
Población	3.096.748	4.090.406	4.264.422	4.628.277	6.343.110
% Cataluña	49,2%	67,9%	70,4%	76,0%	100%
Universo estadístico	> 18 años 2.174.363	> 18 años 2.991.131	> 18 años 3.275.458	> 18 años 3.762.462	> 16 años 6.049.414
Muestra	4.912	5.061	5.263	6.830	10.398
Error relativo (n.s. 2σ)	± 1,9	± 1,9	± 1,8	± 1,5	± 0,7
Marco de la muestra	Censo electoral 1984	Censo electoral 1989	Censo electoral 1994	Censo electoral 1999	Padrón 2005

1. Construcción de los estratos

El censo de población del año 2001 cumple la condición indispensable de suministrar los datos poblacionales para la construcción de la muestra estratificada. De la información que aparece en el cuestionario del censo se han seleccionado una serie de indicadores (o variables), que son los que serán utilizados para la construcción de las variables-criterio empleadas en la estratificación. Dichas variables-criterio habrán de recoger la máxima variedad o heterogeneidad social de la población de Cataluña. Dado que se han tomado las secciones censales como soporte de dichas variables, serán las secciones censales las unidades elementales de estratificación y no los individuos censados.

Esta manera de proceder se justifica por un triple motivo: por las dificultades que se derivan del tratamiento de una matriz de datos de individuos de esta magnitud; porque la agregación en secciones censales nos permite un tratamiento métrico en base a datos porcentuales derivados de

las variables seleccionadas, y porque la sección censal constituye una unidad de división territorial suficientemente pequeña para garantizar la homogeneidad de las características sociales de la población.

La información censal que se ha empleado en esta edición de la *Encuesta* se ha ampliado considerablemente respecto a las ediciones anteriores, puesto que un mayor número de indicadores sociales proporciona una mayor riqueza de los estratos y de la heterogeneidad poblacional, facilitando la interpretación, sobre todo teniendo en cuenta que se ha ampliado la población objeto de estudio al conjunto de Cataluña. La relación de indicadores-variables es inicialmente muy amplia, aunque un análisis posterior de los mismos nos lleva a realizar una selección final de los que, expresando la mayor heterogeneidad social posible de Cataluña, no proporcionen información redundante.

Esta selección final de las variables-criterio obedece, en primer término, a su disponibilidad en el Censo³; responde también a criterios de pertinencia conceptual de acuerdo con los objetivos del estudio, y, finalmente, a criterios de tipo estadístico empleados en el proceso de análisis, como son la manifiesta combinación lineal entre algunas variables o el escaso valor o dispersión de otras.

En consecuencia, se dispone de una matriz de datos inicial de 234 variables-indicadores con información sobre las 5.222 secciones censales de Cataluña en términos del número de personas (o también de hogares) que poseen una determinada característica (o contenido del indicador) en el año 2001. Estas variables se expresan en términos relativos, calculando el porcentaje de la población de la sección censal que posee una determinada característica sobre el total de la población de la sección censal correspondiente.

Tabla 2. Variables-criterio utilizadas en la muestra estratificada

1. % Sección/Municipio	20. % personas no entienden catalán
2. % personas de 0–18 años	21. % personas entienden catalán
3. % personas de 20–29 años	22. % personas entienden y leen catalán
4. % personas de 30–39 años	23. % personas hablan, leen y escriben catalán
5. % personas de 40–49 años	24. % personas solteras
6. % personas de +65 años	25. % personas casadas
7. % personas nacidas en Cataluña	26. % personas separadas
8. % personas nacidas en España	27. % personas divorciadas
9. % personas nacidas en la UE	28. % personas de +10 años analfabetas
10. % personas nacidas en América	29. % personas de 1–3 años escolarizadas
11. % personas nacidas en África	30. % personas de 17–24 años escolarizadas
12. % personas nacionalidad española	31. % personas de +24 años universitarias
13. % personas residencia España –2 años	32. % personas con EGB
14. % personas residencia España 2–5 años	33. % personas con BUP
15. % personas residencia España 6–10 años	34. % personas con FP I
16. % personas residencia España 11–20 años	35. % personas con FP II
17. % personas residencia municipio –2 años	36. % personas estudiantes
18. % personas residencia municipio 2–5 años	37. % personas ocupadas hombres
19. % personas residencia municipio +10 años	38. % personas ocupadas mujeres

3 A partir del asesoramiento del Instituto de Estadística de Cataluña (Idescat) sobre la disponibilidad y fiabilidad de los datos del censo del 2001 descartamos una parte de las variables individuales o de hogar.

39. % personas activas hombres	61. % personas operario cualificado
40. % personas activas mujeres	62. % personas operario no cualificado
41. Tasa de paro	63. % personas trabajo en su domicilio
42. % personas con invalidez	64. % personas trabajo en su municipio
43. % personas viudas	65. % personas trabajo en otro municipio
44. % personas jubiladas	66. % personas transporte privado al trabajo
45. % personas trabajo doméstico	67. % personas transporte público al trabajo
46. % personas Agricultura	68. % personas con hipotecas
47. % personas Industria	69. % personas vivienda heredada
48. % personas Servicios	70. % personas vivienda de alquiler
49. % personas jornada -35 h	71. % personas vivienda hasta 50 m ²
50. % personas jornada 35-39 h	72. % personas vivienda 51-70 m ²
51. % personas jornada 40-50 h	73. % personas vivienda +100 m ²
52. % personas empresarias con asalariados	74. % personas con segunda residencia
53. % personas empresarias sin asalariados	75. % personas con vehículos a motor
54. % personas asalariadas fijas	76. % hogares unipersonales
55. % personas asalariadas eventuales	77. % hogares madre con hijos
56. % personas profesionales liberales	78. % hogares pareja con hijos
57. % personas director	79. % hogares 1 núcleo con parientes
58. % personas técnico	80. % hogares más de 1 núcleo sin parientes
59. % personas administración	81. % hogares más de 1 núcleo con parientes
60. % personas contraemaestre	82. % hogares sin núcleo

Se trata de variables de naturaleza socio-demográfica, cultural-educativa, de actividad laboral y profesional e inactividad, de capital económico, de composición de los hogares, de movilidad territorial y de tamaño del hábitat. De las 234 variables, y como resultado del análisis de sus correlaciones mutuas para examinar su redundancia y significatividad, se eligen 82 indicadores o variables, que son los que aparecen en la Tabla 2.

A partir de esta selección se plantea como paso siguiente la construcción de estratos de población que sean internamente lo más homogéneos posible. Se trata de garantizar que los estratos de la muestra representen, cuanto más fidedignamente mejor, la variedad de la población. El proceso de construcción de los estratos se realiza mediante dos técnicas de análisis multivariable independientes y complementarias: el análisis factorial de componentes principales, para reducir y expresar los principales factores de diferenciación de la población y que, a su vez, se corresponden con las variables-criterio usadas para la construcción de los estratos, y el análisis de clasificación automática, para obtener grupos de población (de secciones censales) o tipos sociales que constituirán los estratos de la muestra.

1.1. El análisis factorial y la identificación de las variables-criterio de los estratos

Con el análisis factorial de componentes principales (ACP) se pretende reducir la información original para obtener un subespacio vectorial de menos dimensiones o factores, donde éstos sean base y, por tanto, linealmente independientes, y que, ordenados de manera jerárquica, conserven una buena parte estipulada de la varianza total. Se obtienen así las dimensiones fundamentales y reducidas de diferenciación social de la población de Cataluña. El ACP se concibe como una etapa previa y un complemento necesario para la obtención de los estratos.

Tabla 3. Varianza explicada por los componentes retenidos

Factor	Valores propios sin rotación			Valores propios de la rotación			
	Valor	% de la varianza	% acumulado	Valor	% de la varianza	% acumulado	% recalculado
1	15,418	18,8	18,8	13,149	16,0	16,0	23,3
2	14,087	17,2	36,0	9,038	11,0	27,1	16,0
3	12,809	15,6	51,6	8,709	10,6	37,8	15,4
4	6,075	7,4	59,0	7,321	8,9	46,6	13,0
5	3,400	4,1	63,2	7,073	8,6	55,2	12,5
6	2,373	2,9	66,0	5,767	7,0	62,3	10,2
7	2,260	2,8	68,8	5,364	6,5	68,8	9,5
					68,8		100,0

El ACP realizado⁴ parte, pues, de las 82 variables de la Tabla 2. La estructura de la información permite reducir estas 82 variables a sólo 7 dimensiones, ejes o variables factoriales, que acumulan el 68,8% de la varianza total⁵. En la Tabla 3 se presentan los valores propios retenidos y la importancia relativa de cada componente antes y después de la rotación *varimax* aplicada.

A continuación se describe la identidad de los ejes factoriales que se han extraído y que estructuran y resumen la información original utilizada.

Primer eje: la categoría socio-profesional

Este primer eje o componente acumula, en el conjunto de las variables, el 16,0% de la varianza total, una vez efectuada la rotación, y el 23,3% de la varianza sobre el total de los ejes retenidos. Es, pues, un eje decisivo y determinante de la estructura de relaciones de las variables originales. Por lo que se refiere a su descripción, constituye el eje clásico de estructuración social que se identifica por variables vinculadas a las diferentes categorías y sectores profesionales y a los niveles educativos.

En este sentido, aparece una polaridad que opone, por un lado, la presencia de categorías profesionales y ocupaciones como directivos, profesionales y técnicos altos del sector terciario en la administración y en servicios, así como niveles educativos medios o altos (BUP, FP II, universitarios de más de 24 años y escolarización entre los 17 y 24 años) a, por otro lado, la presencia de profesiones bajas, cualificadas y no cualificadas, del sector industrial en particular, y un bajo nivel formativo (EGB o analfabetos).

Tiene un interés singular que en este eje aparezcan otras características, que acompañan a las descritas, que se extienden a lo que se puede denominar como formas de vida; de hecho, tan sólo se muestran en el primer polo enunciado, el de las categorías altas, suponiendo entonces que su negación o ausencia está vinculada al otro polo. Así, por ejemplo, los tipos de horarios de trabajo en los sectores referidos son de entre 35 y 39 horas semanales, pero también de menos de 35; se da una mayor presencia de mujeres ocupadas y activas, de situaciones de separación y divorcio y de posesión de una segunda residencia.

⁴ Todos los análisis se han obtenido a través del programa estadístico SPSS.

⁵ El examen de la matriz de correlaciones y de otras medidas nos confirma la adecuación y la validez del análisis: examen de la matriz de correlaciones, comportamiento de la matriz antiimagen de correlaciones y un índice de Kaiser-Meyer-Olkin de 0,872.

Por tanto, es importante señalar que, si bien se trata de un factor vinculado a la posición de clase social, expresada en términos de categorías profesionales y niveles educativos, está marcado también por el tipo de profesión, la presencia activa de la mujer en el mercado de trabajo y otras características que introducen componentes vinculados a las formas de vida de la estratificación social.

Segundo eje: el origen geográfico, autóctonos *versus* antigua inmigración

Este segundo eje representa el 11,0 % de la varianza original, lo que corresponde al 16,0% de la varianza conservada. Es un eje con dos vertientes que, teniendo en cuenta que se superponen sobre el mismo eje, guardan una intercorrelación que define la misma realidad de diferenciación de la población.

La primera vertiente, que es la que presenta una mayor proyección sobre este segundo eje, separa, en uno de sus polos, la población nacida en Cataluña y que manifiesta el mayor grado de conocimiento de la lengua catalana de, en el otro polo, los nacidos fuera de Cataluña que no dominan el catalán en sus diferentes niveles.

La segunda vertiente tiene un contenido más socio-profesional y socioeconómico, ligado, en uno de los extremos, a la condición de ser empresario con o sin asalariados, más la de disponer de una vivienda espaciosa de más de 100 m². En el otro extremo, de forma opuesta, aparece la condición de ser asalariado con contrato eventual y encontrarse en situación de desempleo, con un mayor predominio de la población que se dedica a las tareas del hogar y residencia en viviendas de menor tamaño, entre 51 y 70 m² o más pequeñas.

Es preciso hacer notar igualmente la presencia de otras características que se arropan a lo largo del eje, como son el indicador del pequeño tamaño del municipio, el uso del transporte privado o la contraposición de edades de la población. Por tanto, este eje es revelador de una diferencia, que se traduce en características sociales, entre una población de origen catalán en los ámbitos más rurales y la población de origen inmigrante de primera generación llegados a Cataluña del resto de España, que se localiza en grandes núcleos urbanos.

Tercer eje: el ciclo vital

Con un 10,6 % de la varianza total y el 15,4% de la varianza retenida, este tercer factor se identifica con el ciclo vital, vinculado fundamentalmente a la edad, pero también a algunas otras características concomitantes que identifican su contexto y condiciones.

Se trata, en uno de los polos del eje, de una población de menos de 49 años pero con mayor incidencia de la más joven, formada, además, por parejas con hijos y también por personas solteras, con un peso destacado de la población estudiantil. Otras dos características connotan, de alguna manera, este primer polo: la hipoteca y el uso de vehículos a motor. También merece destacarse, como rasgo distintivo, que sea en las secciones censales que corresponden a este polo donde se dé una mayor proporción de personas no autóctonas, con residencia en España desde hace un cierto tiempo.

Si las características precedentes definen uno de los polos del eje, el otro lado aparece claramente marcado por una población envejecida y por sus consecuencias con respecto al ciclo vital: mayores de 65 años, residentes en el municipio desde hace mucho tiempo, que están jubilados, viven solos en hogares sin núcleo familiar y/o en situaciones de viudedad, con un peso importante de viviendas de alquiler. Todas ellas características que, si bien denotan un perfil asociado a la edad, comportan otros indicadores familiares y formas de hogar y de vivienda.

Cuarto eje: el factor rural-urbano

La parte de la varianza explicada por cada uno del resto de ejes, del cuarto al séptimo, experimenta un salto descendente muy considerable. En concreto, este cuarto eje representa el 8,9% en relación al total, y un 13,0% de la varianza retenida por los siete primeros ejes elegidos.

En este eje, sólo uno de los polos aparece claramente identificado y con unas características muy coherentes; consecuentemente, en el otro polo se sitúan los rasgos opuestos a los que describimos como definitorios del primer polo. Se trata de una población que reside en municipios pequeños, sobre todo dedicada al sector agropecuario, con una empresa familiar, con la condición de empresarios sin asalariados, y que trabajan en el mismo domicilio. Se relacionan con estos rasgos de la población otros que tienen que ver con el tipo de hogar y de vivienda. Así, se da siempre la presencia de un núcleo familiar, siendo también numerosos los hogares extensos, y con un predominio de viviendas grandes que se heredan. Se caracteriza además por una mayor proporción de personas nacidas en Cataluña.

En el otro extremo del eje se acumulan el resto de características, las de una población que vive en secciones de núcleos urbanos, de diferentes orígenes, del resto de sectores productivos, con otras formas de composición del hogar y otras viviendas.

Quinto eje: la nueva inmigración

Este eje acumula, en términos de varianza explicada, un 8,6% de la varianza total, lo que supone un 12,5% de la dispersión total atribuida a los siete ejes elegidos. Su identidad está perfectamente definida, especialmente en uno de los polos y por contraposición al otro, en el que no aparece ningún indicador posicionado de forma explícita; es decir, que dicho polo está definido por la negación de (o bien la oposición a) las características que señalamos a continuación para el primer polo.

Dicho primer polo representa la nueva inmigración, si bien se trata de personas que ya llevan un cierto tiempo en el país. Estamos hablando de secciones con un peso importante de personas que tienen la residencia en España desde hace al menos dos años, si bien no sobrepasan los diez años en el país (datos censales de 2001), y que no tienen la nacionalidad española. Su origen es sobre todo africano o latinoamericano, viven de alquiler y en viviendas pequeñas, están solteras y viven en hogares sin núcleo. Este perfil de recién llegado se contrapone a la población autóctona, enraizada y que vive en hogares con núcleos familiares.

Sexto eje: el factor de actividad laboral

Es un eje que representa el 7,0% de la varianza explicada y un 10,2% de la conservada en los siete ejes escogidos.

Los extremos de este factor contraponen la ocupación a la inactividad laboral y al paro. Un polo está claramente definido por la población activa y ocupada, tanto de hombres como de mujeres, asalariados fijos (y en menor medida eventuales), con horarios semanales exhaustivos de 40 a 50 horas, que trabaja sobre todo en la industria y en categorías laborales de la población obrera tradicional cualificada. En el polo contrapuesto se puede encontrar una mezcla de población inactiva, compuesta sobre todo de personas jubiladas, junto a perfiles de secciones con una mayor tasa de desempleo y de población estudiantil.

Séptimo eje: la movilidad territorial

Este eje representa un peso similar al precedente en la totalidad de la variabilidad, un 6,5%, y un 9,5% en relación con la varianza que representan los siete factores.

A un lado, encontramos a la población de las secciones censales con un importante peso relativo de residentes recientes en el municipio (menos de 5 años). Se trata de municipios de tamaño medio, con población joven o adulta llegada en edad reproductiva, con hipotecas y que trabaja en otro municipio, al cual se desplaza con transporte privado. Este perfil contrasta con el otro polo, donde se localiza sobre todo una población con más de 10 años de residencia en el municipio y que trabaja en él. Vemos, pues, la configuración de un factor ligado a la movilidad territorial de una población que en los últimos años se ha desplazado a ciudades intermedias y a urbanizaciones periféricas, ya sea desde la gran ciudad, ya sea desde otros núcleos.

1.2. El análisis de clasificación o la construcción de los estratos de la muestra estratificada

La estratificación de la muestra descansa sobre el principio según el cual la población en estudio es estadísticamente heterogénea. La homogeneidad estadística se refiere, en el caso de las variables o contenidos sociales, al grado de correlación entre las características elegidas para identificar una población; cuanto más correlacionadas estén dichas características más homogénea será la población considerada. Esta condición de máxima heterogeneidad de la población en cuanto a las variables factoriales elegidas, que pasan a ser las variables-criterio en la construcción de los estratos, se ha llenado satisfactoriamente con el análisis factorial realizado al obtener unos ejes incorrelacionados y con una buena parte de la varianza de la población incorporada.

El objetivo ahora es otro, aunque consecuente y complementario. Se trata de la estratificación, que consiste en clasificar las secciones censales en estratos compuestos por conjuntos de secciones lo más homogéneas posible dentro de cada estrato y lo más heterogéneas posible entre las secciones de los diferentes estratos, todo ello en función de las variables factoriales extraídas del ACP. Con respecto a la muestra aleatoria simple, este procedimiento asegura además, a efectos del muestreo, una mayor ganancia en precisión o disminución de error. Así pues, tomando como variables-criterio los siete ejes factoriales precedentes, o subespacio vectorial correspondiente reducido de siete dimensiones, se lleva a cabo la clasificación de las secciones censales. Este es el punto de partida y en esto consiste el análisis de clasificación automática.

Esta clasificación se realiza sin tener en cuenta ninguna restricción de contigüidad territorial, por lo que se podrá obtener como resultado un mapa de las secciones censales de Cataluña según los estratos de pertenencia y con su diferente caracterización social; es decir, en función de la pertenencia a los estratos observaremos una distribución diferenciada del territorio.

Partimos pues de una matriz reducida con las 7 variables-factoriales que identifican las 5.222 secciones censales de Cataluña. El método de clasificación empleado, tomando como medida de proximidad la distancia cuadrática euclidiana en el espacio vectorial de la base de los siete vectores, se desglosa en dos etapas. En ambas, se ha procedido básicamente a:

- a) En un primer momento se realiza una clasificación jerárquica ascendente, mediante el procedimiento de *Ward* (de mínima pérdida de inercia), para clasificar inicialmente las secciones censales en el número más adecuado de grupos a partir de criterios estadísticos⁶ y del estudio conceptual-descriptivo de los estratos resultantes interpretados a partir de las variables-factoriales y de las mismas variables originales. En esta etapa se efectúa el análisis sistemático de las particiones comprendidas entre 20 y 2 grupos. Finalmente, el número de grupos, estratos a efectos muestrales, se fija en 10.
- b) En un segundo momento, con el número de grupos determinado y los centros iniciales definidos, se opera una clasificación no jerárquica por la agregación alrededor de centros móviles, con el objetivo de optimizar la asignación de las secciones a los estratos. Esta reasignación óptima configura la estratificación final de población con la distribución de secciones y centros finales en los factores que se puede apreciar en la Tabla 4 y en el Gráfico 1.

Presentamos a continuación la caracterización sociológica de los diez estratos, de forma breve y resumida, descrita a partir de las dimensiones de identidad social y aludiendo también a las variables originales. Señalamos los perfiles más destacados con relación al comportamiento medio del conjunto de la población catalana.

⁶ Cambios en las medidas de distancias por la aglomeración de clases y análisis del cambio en la varianza explicada por cada nueva partición.

Tabla 4. Distribución de las secciones censales por estrato y centros

Estrato	Secciones censales		Centros finales						
			Factor 1 Categoría socioprofesional	Factor 2 Autóctonos Antigua inmigración	Factor 3 Ciclo vital	Factor 4 Rural–Urbano	Factor 5 Nueva inmigración	Factor 6 Actividad laboral	Factor 7 Movilidad territorial
1	576	11,0%	-0,780	1,379	-0,233	0,523	-0,148	0,322	-0,302
2	397	7,6%	0,366	0,153	0,387	0,053	-0,406	0,331	2,368
3	723	13,8%	-0,322	0,246	0,759	0,459	-0,282	0,523	0,097
4	504	9,7%	-0,451	0,773	-0,563	-2,290	-0,382	-0,092	-0,053
5	424	8,1%	-1,190	-0,321	0,997	0,134	0,888	-0,667	-0,196
6	425	8,1%	1,679	0,321	1,020	-0,075	-0,103	-0,914	-0,447
7	769	14,7%	-0,108	-1,424	-0,006	-0,107	-0,438	0,636	-0,348
8	407	7,8%	-0,445	-0,716	-1,064	0,628	-0,581	-1,288	0,119
9	771	14,8%	1,091	0,088	-0,701	0,431	0,329	0,027	-0,508
10	226	4,3%	-0,095	-0,459	-0,720	-0,139	2,789	0,166	0,327
Total	5.222	100%	0,000	0,000	0,000	0,000	0,000	0,000	0,000

El primer estrato: población catalana envejecida y de clases trabajadoras en ciudades no metropolitanas

Las secciones censales que corresponden a este estrato son 576, que representan el 11,0% de toda Cataluña. Su presencia es mayor en las Comarques Gironines y Centrals, en municipios de tamaño medio. Se proyectan, sobre todo, en el segundo eje (origen catalán) y, en menor medida, sobre el primero (categorías socio-profesionales bajas).

Su perfil característico es el de personas mayores, en buena medida jubiladas y con una destacada proporción de situaciones de invalidez. Son secciones censales con una población mayoritaria de origen catalán, casada, donde también es característica la ocupación en el sector industrial, en categorías laborales de operario cualificado, y los niveles de estudios bajos.

Comparte con el estrato 4 la mayor presencia de personas de más edad y de origen catalán, si bien en aquél se destaca la importancia del sector agropecuario de los municipios más pequeños.

El segundo estrato: movilidad de familias jóvenes de clase media

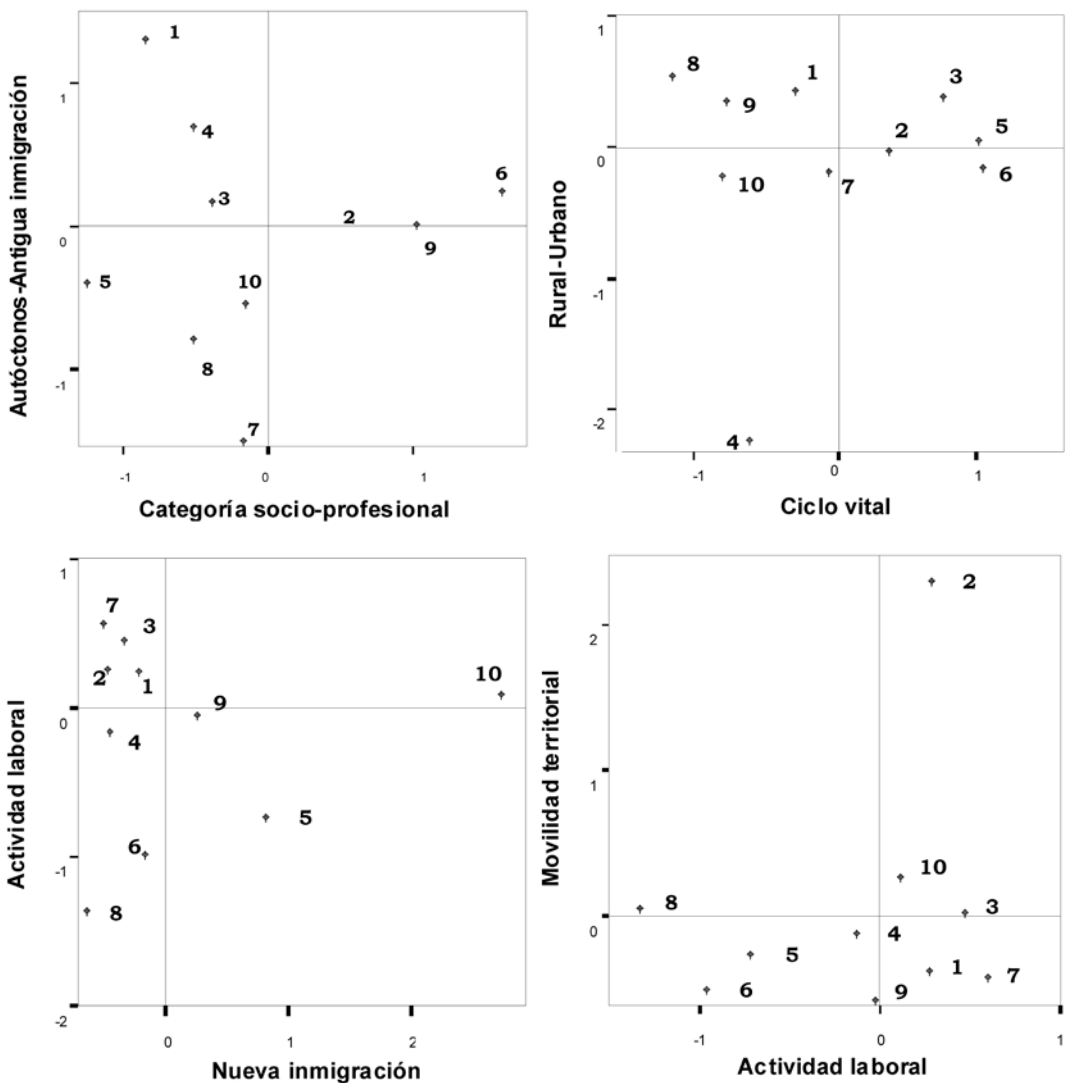
El estrato 2 lo componen 397 secciones censales, el 7,6% del total de secciones de Cataluña. Se distribuyen, sobre todo, en el ámbito metropolitano y en el Camp de Tarragona, en municipios de tamaño medio. Es un estrato con una mayor proyección sobre el factor 7 (movilidad territorial).

Se trata, por tanto, de población joven que ha llegado hace poco al municipio, en edad reproductiva, casada y en familias nucleares, que reside en viviendas grandes que financian con hipotecas. Su lugar de trabajo se localiza mayoritariamente en otros municipios, destacando la proporción de hombres ocupados en sectores industriales o de servicios, en categorías laborales medias o altas, con estabilidad laboral y jornadas extensas. Destaca también, en particular, la mayor presencia de personas procedentes de la Unión Europea con entre 6 y 10 años de residencia en España.

El tercer estrato: población joven de clase trabajadora de municipios urbanos

El estrato 3 comparte algunos perfiles con el anterior, pero es más numeroso. Se compone de 723 secciones censales, que representan el 13,8% de las de toda Cataluña. Es igualmente población

Gráfico 1. Posición de los estratos en los factores



- Estrato 1 Población catalana envejecida y de clases trabajadoras en ciudades no metropolitanas
- Estrato 2 Movilidad de familias jóvenes de clase media
- Estrato 3 Población joven de clase trabajadora de municipios urbanos
- Estrato 4 Los pequeños municipios rurales
- Estrato 5 Población urbana joven de clase trabajadora precaria
- Estrato 6 Clases sociales media y alta de las grandes ciudades
- Estrato 7 Antigua inmigración en el área metropolitana
- Estrato 8 Población envejecida urbana de la antigua inmigración
- Estrato 9 Población de capital de clase media-alta envejecida
- Estrato 10 La nueva inmigración

Fuente: elaboración propia

residente en el ámbito metropolitano y en el Camp de Tarragona, pero con presencia también en el Àmbit de Ponent y en las Terres de l'Ebre, en municipios de mayor tamaño. Sus secciones tienen una mayor proyección sobre el eje 3 (ciclo vital joven).

Así, contemplamos un grupo de población joven, como en el segundo estrato, casada y que convive en hogares también nucleares. Tienen igualmente hipotecas, pero de viviendas más modestas. Precisamente, la principal diferencia en relación al segundo estrato es una posición social de clase inferior; corresponden de forma destacada a hombres activos con jornadas extensas, en categorías profesionales de obreros cualificados del sector industrial, y con una mayor precariedad laboral. Comparten de nuevo con el segundo estrato el hecho de trabajar en otro municipio, si bien no obedecen al factor de movilidad territorial.

El cuarto estrato: los pequeños municipios rurales

Este estrato contiene 504 secciones censales, el 9,7 % del total. Agrupa sobre todo a las secciones censales de los municipios más pequeños de fuera del ámbito metropolitano y alejados de los núcleos urbanos más poblados: se extiende por el Àmbit de Ponent, las Terres de l'Ebre, las Comarques Gironines y Centrals, el Camp de Tarragona y el Alt Pirineu i Aran. Es el estrato que se proyecta fundamentalmente sobre el eje 4 (rural-urbano).

Se trata de una población de origen catalán, bastante envejecida. Destacan los hogares con familias extensas o sin núcleo y con residencia en viviendas grandes, principalmente heredadas. Al tratarse de un ámbito estrictamente rural, predomina la presencia de la ocupación masculina en el sector agropecuario, en las categorías de empresarios, con o sin asalariados, y con una formación básica.

El quinto estrato: población urbana joven de clase trabajadora precaria

Está formado por 424 secciones censales, que corresponden al 8,1 % de las secciones de Cataluña, distribuidas en los municipios de más de 20.000 habitantes del ámbito metropolitano de Barcelona, pero también en las Comarques Gironines, el Camp de Tarragona y las Terres de l'Ebre. Se proyectan sobre todo en los ejes 1 (categoría socio-profesional baja), 3 (ciclo vital joven) y, parcialmente, en el 5 (nueva inmigración).

Estamos ante un estrato marcado por el perfil de población joven de clase trabajadora, como el estrato 3, pero en peores condiciones sociales, en categorías laborales de operarios cualificados y no cualificados, con bajo nivel de estudios y con los más altos niveles de precariedad laboral y de paro. En estas secciones predominan las personas nacidas en el resto de España, con una parte de inmigración más reciente, lo que significa hablar de hijos de una primera generación de inmigrantes que muestran unos bajos conocimientos de la lengua catalana. Son familias jóvenes, nucleares, también hipotecadas.

El sexto estrato: clases sociales media y alta de las grandes ciudades

Las 425 secciones censales de este estrato, que representan el 8,1% del total de Cataluña, se concentran en las grandes ciudades catalanas, especialmente en Barcelona y en su ámbito metropolitano, pero también en Tarragona, Girona y Lleida. Su proyección se da sobre todo en el eje 1 (categoría socio-profesional alta), y también sobre el 3 (ciclo vital joven) y el 6 (actividad laboral).

En este estrato hay una importante presencia de las categorías socio-profesionales y de posición social alta. Se trata de aquellas secciones censales con el valor más elevado en cuanto al primer eje factorial: directivos, profesionales y técnicos altos, empresarios, principalmente en el sector servicios, y con jornadas laborales de menos de 40 horas semanales. Presentan unas tasas más altas de actividad y de ocupación, femenina en particular, así como elevados niveles educativos y de escolarización universitaria. Son catalanes de origen, arraigados en el municipio, en el cual también trabajan, residen en viviendas de más de 100 m² y poseen una segunda residencia.

No sobresale ninguna franja de edad como signo característico, pero se sitúan en el perfil de ciclo vital de familias jóvenes del tercer factor. Sí aparece como propio de este estrato la mayor presencia de personas divorciadas, de hogares compuestos de madres con hijos y de personas que viven en régimen de alquiler.

El séptimo estrato: antigua inmigración en el área metropolitana

Este estrato, con 769 secciones censales, comporta el 14,7% del conjunto de Cataluña, siendo el más numeroso junto con el estrato 9. Sus secciones se concentran en su totalidad en Barcelona y en la primera corona metropolitana. Su proyección máxima es sobre el eje 2 (antigua inmigración).

Su rasgo característico es el de congregar a la población metropolitana que proviene de la inmigración del resto de España y que está arraigada en el municipio desde hace tiempo, más de 10 años por lo menos. A pesar de ello, presentan unos niveles bajos de conocimiento de la lengua catalana. Su estructura de edades es similar a la del conjunto de la población, con una presencia más destacada de la franja de 20 a 29 años. Es un grupo de categorías socio-profesionales bajas, sobre todo de operarios cualificados de la industria. La formación que caracteriza el estrato son la FP I y FP II. Este grupo posee un relativamente importante nivel ocupacional en su tasa de ocupación y de actividad, en particular masculina, con una cierta estabilidad contractual, si bien aparece también la contratación eventual. Trabajan en mayor proporción en otros municipios distintos al de residencia. De forma característica, son secciones de hogares nucleares y de viviendas pequeñas, de menos de 70 m².

El octavo estrato: población envejecida urbana de la antigua inmigración

Las 407 secciones censales de este estrato, el 7,8% del total, pertenecen también a los grandes núcleos urbanos del área metropolitana, pero abarcan ahora hasta la segunda corona y a una pequeña parte de los grandes núcleos urbanos del resto de Cataluña, sobre todo de Tarragona. Son secciones con una mayor proyección sobre los ejes 6 (inactividad) y 3 (ciclo vital de vejez), y también en el 2 (antigua inmigración).

Es decir, su rasgo distintivo es la elevada proporción de personas de más de 65 años, la mayor de todos los estratos, juntamente con el hecho de ser una población inmigrada del resto de España, como en el caso anterior. Comparte también el arraigo en el territorio, el desconocimiento de la lengua catalana y habitar en viviendas poco espaciosas. Al tratarse de secciones censales muy envejecidas, comporta un predominio de población inactiva en situaciones de invalidez, viudedad y jubilación.

No obstante, son zonas del territorio donde conviven igualmente personas más jóvenes que sufren en buena medida la desocupación y confluye también en ellas una mayor presencia de personas dedicadas a las tareas del hogar.

El noveno estrato: población de capital de clase media-alta envejecida.

Tiene 771 secciones censales, por lo que configura el estrato más numeroso, con el 14,8% de las secciones de toda Cataluña. Están predominantemente localizadas en Barcelona ciudad; el resto se reparte entre las tres capitales de provincia catalanas. El estrato tiene su mayor proyección sobre el eje 1 (categorías socio-profesionales medias y altas) y también sobre el 3 (ciclo vital de vejez).

La población urbana de estos estratos se distribuye, no obstante, entre las diferentes edades, si bien con un peso más destacado de los que tienen más de 65 años, por lo que sobresale la proporción de personas jubiladas o en situación de viudedad. Pero es un estrato en el que se combinan otras características diversas. Aunque se trate de una población que mayoritariamente lleva mucho tiempo en el municipio, de origen catalán y con un buen conocimiento de la lengua catalana, se encuentran también en este estrato personas nacidas en el resto de España y, con una proporción destacada, de origen latinoamericano.

Son secciones con personas también activas, en particular mujeres, con niveles medios o superiores de estudios y cualificadas, con estabilidad laboral, que trabajan fundamentalmente en el sector servicios en el mismo municipio y con jornadas laborales cortas.

Este perfil se completa con un peso importante de los hogares unipersonales, sin núcleo o monoparentales. Sus viviendas son de tamaño medio, con una buena proporción de tenencia en alquiler y una proporción de segundas residencias por encima de la media global.

El décimo estrato: la nueva inmigración

Finalmente, el estrato décimo, con 226 secciones censales que representan el 4,3% del total, es el de menos peso del conjunto de los estratos en número de secciones censales, si bien es especialmente significativo en cuanto a los cambios sociales más recientes. Se distribuye, sobre todo, en zonas urbanas del ámbito metropolitano, pero también en las ciudades de las Comarques Gironines, el Camp de Tarragona y las Comarques Centrals. Su proyección es mayor sobre el eje 5 (nueva inmigración), así como sobre el 3 (ciclo vital de vejez).

Confluye en este estrato una doble realidad social definida: por una parte, a partir de la edad y, por otra, de la nueva inmigración. De un lado, se trata de una población con una proporción destacada de personas de más de 65 años, circunstancia que confluye con el predominio de las situaciones de jubilación y viudedad. Pero, a la vez, otro de sus rasgos destacables es la concentración de la población recientemente inmigrada y llegada al municipio con orígenes diversos, africanos y latinoamericanos sobre todo, no precisamente mayor, y que convive con autóctonos nacidos en Cataluña o en el resto de España.

En consecuencia, se dan diferentes perfiles característicos: la alta proporción de hogares sin núcleo y unipersonales, de personas solteras, de viviendas pequeñas y de alquiler, o el bajo conocimiento de la lengua. La población activa está ocupada especialmente en el sector servicios, con baja cualificación y con la más alta tasa de eventualidad, realidad que se acompaña con altas tasas de desempleo.

2. Afijación de la muestra y cuotas muestrales

Una vez determinados los estratos, con la caracterización social descrita y como expresión de conjuntos homogéneos de población agregada en secciones censales, corresponde ahora determinar la distribución de los individuos o unidades de la muestra en estos estratos, partiendo de un número total preestablecido. El cálculo del tamaño muestral se realiza a partir de los parámetros habituales: tamaño de la población, estimación de la media y de la variabilidad, nivel de significación y error muestral.

2.1. Tamaño de la muestra y afijación

Como medida de variabilidad se considera la distancia cuadrática euclidiana de cada sección al centro global de la nube de puntos compuesto de secciones censales en un espacio vectorial reducido a una base de siete vectores o variables–factores, componentes del ACP, y, como parámetro, la media de esta distancia. De esta manera se consigue estimar y caracterizar cada unidad o sección censal no con una variable que exprese una sola característica aislada de las secciones censales, sino con una variable densa, compuesta de combinaciones de los ejes factoriales que concentran una gran parte de la varianza. De dicha variable sintética se calcula la distancia a cada sección, su valor medio y su desviación, datos fundamentales para el cálculo de la muestra.

El tamaño de la muestra para el conjunto de Cataluña se fijó inicialmente en 8.348 individuos. Si consideramos un nivel de significación de 2σ , con este tamaño muestral n para datos globales de Cataluña, el error muestral relativo que se obtiene es del 0,78%, aplicando la fórmula siguiente:

$$n = \frac{z^2 \cdot \sigma_y^2}{e^2 \cdot \bar{Y}^2} \quad \text{de donde} \quad e = \sqrt{\frac{z^2 \cdot \sigma_y^2}{n \cdot \bar{Y}^2}}$$

con:

- z el número de sigmas de nivel de significación,
- σ_y^2 la varianza de la distancia cuadrática euclidiana de las secciones censales al centro de la totalidad de la nube (valor obtenido 0,783),
- e el error muestral,
- \bar{Y} la media de las distancias cuadráticas euclidianas (valor obtenido de 2,493).

Este número de individuos se distribuye entre los estratos definidos por la muestra con el criterio de afijación óptima de Neyman. Con este criterio se opera el efecto de la estratificación de la muestra según el cual cuanto más grande y variable es un estrato mayor proporción de muestra se le asigna. Por tanto, no se trata de una distribución estrictamente proporcional a la población de cada estrato, sino que a efectos de optimización, de ganancia en la precisión de las estimaciones, se adopta este doble criterio que se expresa en la fórmula:

$$n_h = \frac{N_h \cdot \sigma_h}{\sum_{h=1}^K N_h \cdot \sigma_h} \cdot n$$

donde:

- n_h es el tamaño muestral del estrato h ($h=1...10$),
- N_h es la población de 16 o más años del estrato h ,
- σ_h es la desviación de la distancia cuadrática euclidiana de las secciones censales del estrato h al centro de su estrato.

Los valores que se obtienen de la afijación se presentan en la Tabla 5.

Tabla 5. Afijación muestral óptima de Neyman para cada estrato

Estrato	N_h	σ_h	$N_h \cdot \sigma_h$	Coficiente de afijación	n_h
1	640.561	0,47324	303.138,8	0,10173	849,27
2	556.508	0,83655	465.545,6	0,15624	1.304,27
3	977.070	0,40956	400.172,8	0,13430	1.121,12
4	248.814	0,78853	196.197,4	0,06584	549,66
5	467.962	0,67811	317.328,4	0,10650	889,02
6	435.320	0,60354	262.731,6	0,08817	736,07
7	845.423	0,47077	397.996,4	0,13357	1.115,02
8	354.574	0,56753	201.231,2	0,06753	563,77
9	686.607	0,42044	288.678,8	0,09688	808,76
10	195.425	0,75077	146.719,3	0,04924	411,05
Total	5.408.264	0,62386	2.979.740,3	1,00000	8.348,00

2.2. Cuotas muestrales

Con el número muestral de cada estrato se procede a la asignación proporcional de las cuotas de muestra en términos de individuos correspondientes a cada sección censal del estrato. Con este reparto se garantiza el cumplimiento de la aleatoriedad de la muestra en la elección de un individuo que pertenece a una sección determinada.

La determinación de las cuotas de muestra por sección sigue un proceso en cuatro etapas:

Etapa 1: distribución proporcional de cuotas

Inicialmente, se asigna a cada sección censal de un estrato la cuota de muestra que le corresponde según una distribución proporcional de acuerdo con su población de 16 o más años mediante la fórmula:

$$n_{sh} = \frac{N_{sh}}{N_h} \cdot n_h$$

donde:

- n_{sh} es la cuota de muestra de la sección s del estrato h ,
- N_{sh} es la población de 16 o más años de la sección s del estrato h ,
- N_h es la población de 16 o más años del estrato h ,
- n_h es el tamaño muestral del estrato h .

Etapa 2: subestratificación

La asignación inicial de cuotas para cada sección censal se distribuye en un segundo momento considerando, en cada estrato, una subestratificación a partir de un doble criterio de división territorial: según siete grandes ámbitos territoriales⁷ y según el tamaño de los municipios (hábitat) a los que pertenecen las secciones (Tabla 6).

Tabla 6. Variables territoriales de subestratificación

Ámbito territorial	Población 16 y más	%	Hábitat	Población 16 y más	%
Àmbit Metropolità	3.749.174	69,3	Hasta 500 hab.	78.739	1,5
Comarques Gironines	468.550	8,7	501–2.000 hab.	253.618	4,7
Camp de Tarragona	380.708	7,0	2.001–5.000 hab.	332.814	6,2
Terres de l'Ebre	135.700	2,5	5.001–20.000 hab.	952.450	17,6
Àmbit de Ponent	255.825	4,7	Más de 20.000 hab.	3.790.643	70,1
Comarques Centrals	364.167	6,7	Total	5.408.264	100,0
Alt Pirineu i Aran	54.140	1,0			
Total	5.408.264	100,0			

De esta forma, se realiza la distribución de las 8.348 encuestas determinadas inicialmente entre 350 substratos h_j teóricos, resultado de combinar los 10 estratos con los 7 ámbitos territoriales y con los 5 tamaños municipales, que generan 187 substratos reales no nulos.

⁷ Ámbitos territoriales definidos por el Plan Territorial de Cataluña.

Esta distribución implica contemplar todas y cada una de las secciones censales de Cataluña⁸, lo que significa, desde el punto de vista del trabajo de campo, unos costes muy elevados, derivados de los obligados desplazamientos por todo el territorio catalán para la realización de las entrevistas personales. En este sentido, y para favorecer la concentración territorial, se ha aplicado un criterio de redistribución de las cuotas de muestra a partir de la diferenciación de las secciones que pertenecen a los municipios más pequeños (hasta 20.000 habitantes) de los más grandes.

En el caso de los subestratos con más de 20.000 habitantes, se consideran todas las secciones censales con la cuota de muestra asignada inicialmente.

Si la sección censal es de un subestrato de municipios con una población inferior o igual a 20.000 habitantes, en total 137 subestratos, se efectúa la concentración territorial mediante la atribución de la cuota total de muestra del subestrato a una selección aleatoria de secciones del mismo de la forma siguiente:

- a) En primer lugar se realiza la extracción aleatoria de secciones. Para ello se divide la cuota de muestra del subestrato entre 3 para fijar el número de secciones del subestrato que serán seleccionadas de forma aleatoria, es decir:

$$ns_{h_j} = \frac{n_{h_j}}{3}$$

donde:

ns_{h_j} es el número de secciones a seleccionar del subestrato h_j (con $j=1\dots 137$) de municipios hasta 20.000 habitantes,

n_{h_j} es la cuota muestral del subestrato h_j .

- b) Se procede a la concentración de la cuota de muestra del subestrato en las secciones censales seleccionadas (\bar{s}) de forma proporcional al peso poblacional de la sección sobre el total de las secciones seleccionadas del subestrato. La asignación de cuotas se expresa en la fórmula:

$$n_{\bar{s}h_j} = \frac{N_{\bar{s}h_j}}{N_{\bar{h}_j}} \cdot n_{h_j}$$

donde:

$n_{\bar{s}h_j}$ es la cuota de muestra de la sección s seleccionada (\bar{s}) del subestrato h_j (con $j= 1..137$) de municipios hasta 20.000 habitantes,

$N_{\bar{s}h_j}$ es la población de 16 y más años de la sección s seleccionada (\bar{s}) del subestrato h_j ,

$N_{\bar{h}_j}$ es la población de 16 y más años total del subestrato h_j con las secciones seleccionadas,

n_{h_j} es la cuota muestral del subestrato h_j .

La asignación que se obtiene da lugar a cuotas de muestra no enteras para cada sección censal, por lo que, a efectos de elección de los individuos, se computa el redondeo generando un tamaño muestral final de 8.352 encuestas. La distribución de frecuencias de las cuotas de muestra finales se presenta en la Tabla 7.

El proceso seguido implica finalmente tomar 4.476 de las 5.222 secciones censales; se dejan de considerar, por tanto, 746, aquellas que tienen un peso poblacional más reducido y que, siendo del mismo subestrato, aleatoriamente se han descartado para favorecer la concentración en el territorio.

⁸ Teniendo en cuenta que algunas secciones censales inicialmente obtienen una cuota de muestra redondeada de 0 (Tabla 7).

Tabla 7. Cuotas de muestra finales

Cuota de muestra	Secciones	%	Cuota de muestra	Secciones	%
0	746	14,3	6	26	0,5
1	2.101	40,2	7	7	0,1
2	1.443	27,6	8	3	0,1
3	562	10,8	9	1	0,0
4	230	4,4	11	1	0,0
5	102	2,0	Total	5.222	100,0

Etapa 3: ampliación de la muestra

La necesidad de garantizar un nivel de representatividad suficiente para determinadas unidades territoriales de interés en el estudio condujo a la ampliación de la muestra, por lo que se añadieron 2.094 encuestas que se repartieron de forma desigual entre algunos ámbitos territoriales y comarcas para garantizar, como submuestra, un tamaño suficiente, según se indica en la tabla siguiente:

Tabla 8. Muestra ampliada

Territorio	Muestra Adicional	Territorio	Muestra Adicional
Camp de Tarragona	164	Bages	111
Terres de l'Ebre	361	Berguedà	15
Àmbit de Ponent	348	Garraf	124
Alt Pirineu i Aran	504	Osona	109
Alt Penedès	173	Solsonès	19
Anoia	166	Total	2.094

Para cada uno de estos 11 territorios, el tamaño de muestra adicional se repartió de forma proporcional al peso de la cuotas de muestra de cada territorio. El reparto se expresa en la fórmula:

$$na_{st} = \frac{n_{st}}{n_t} \cdot na_t$$

donde:

- na_{st} es la cuota de muestra adicional de la sección s seleccionada (\bar{s}) del territorio t (t=1...11),
- n_{st} es la cuota de muestra de la sección s seleccionada (\bar{s}) en el territorio t,
- n_t es la cuota de muestra del territorio t,
- na_t es la muestra adicional del territorio t (Tabla 8).

La cuota final de las secciones censales de estos territorios son el resultado del redondeo al número entero de la suma de la cuota inicial más la adicional.

Etapa 4: concentración de cuotas en Alt Pirineu i Aran

Finalmente, en el caso particular del subestrato de los municipios de hasta 500 habitantes del territorio del Alt Pirineu i Aran, que pertenecen al estrato 4, se operó una segunda extracción aleatoria de secciones con el objetivo de reiterar la concentración territorial de las entrevistas, dadas las características particulares de escasa población de esta zona. Del total de 46 secciones censales de este subestrato se seleccionaron aleatoriamente 13, y se les asignó la cuota de muestra del subestrato, 132 encuestas, de forma proporcional al peso muestral de éstas.

2.3. Distribución final de la muestra

Teniendo en cuenta las redistribuciones de cuotas y la muestra adicional, se obtiene la distribución final de encuestas a realizar en cada una de las secciones censales seleccionadas. En las Tablas 9, 10 y 11 se presenta la distribución de las cuotas de muestra que resulta del proceso seguido: la distribución inicial, la redistribución con el criterio de concentración territorial en los municipios de hasta 20.000 habitantes y la ampliación de la muestra. Estas tres distribuciones se presentan según el estrato, el ámbito territorial y el hábitat.

Tabla 9. Distribución de las cuotas de muestra según el estrato

Estrato	Cuotas iniciales		Redistribución de cuotas		Muestra adicional	Muestra total	
Estrato 1	849	10,2%	850	10,2%	446	1.296	12,4%
Estrato 2	1.304	15,6%	1.300	15,6%	196	1.496	14,3%
Estrato 3	1.121	13,4%	1.113	13,3%	352	1.465	14,0%
Estrato 4	550	6,6%	549	6,6%	691	1.240	11,9%
Estrato 5	889	10,6%	884	10,6%	182	1.066	10,2%
Estrato 6	736	8,8%	735	8,8%	67	802	7,7%

Tabla 10. Distribución de las cuotas de muestra según el ámbito territorial

Ámbito territorial	Cuotas iniciales		Redistribución de cuotas		Muestra adicional	Muestra total	
Àmbit Metropolità	5.661	67,8%	5.669	67,9%	297	5.966	57,1%
Comarques Gironines	773	9,3%	774	9,3%	0	774	7,4%
Camp de Tarragona	620	7,4%	618	7,4%	164	782	7,5%
Terres de l'Ebre	239	2,9%	237	2,8%	361	598	5,7%
Àmbit de Ponent	435	5,2%	434	5,2%	348	782	7,5%
Comarques Centrals	525	6,3%	530	6,3%	420	950	9,1%
Alt Pirineu i Aran	95	1,1%	90	1,1%	504	594	5,7%
Total	8.348	100,0%	8.352	100,0%	2.094	10.446	100,0%

Tabla 11. Distribución de las cuotas de muestra según el hábitat

Hábitat	Cuotas iniciales		Redistribución de cuotas		Muestra adicional	Muestra total	
Hasta 500 hab.	169	2,0%	162	1,9%	181	343	3,3%
501–2.000 hab.	484	5,8%	487	5,8%	462	949	9,1%
2.001–5.000 hab.	561	6,7%	557	6,7%	306	863	8,3%
5.001–20.000 hab.	1.565	18,7%	1.563	18,7%	619	2.182	20,9%
Más de 20.000 hab.	5.570	66,7%	5.583	66,8%	526	6.109	58,5%
Total	8.348	100,0%	8.352	100,0%	2.094	10.446	100,0%

3. Ponderación *a posteriori* de la muestra y elevación

El diseño muestral estratificado de la *Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006* supone la afijación no proporcional de las cuotas de muestra de cada estrato. Según el tamaño poblacional del estrato y la variabilidad de las características sociales propias de éste, habrá individuos de 16 y más años que tendrán una probabilidad mayor de ser elegidos a partir de la cuota que se asigna al estrato donde se ubica la sección censal a la que pertenecen, es decir, no se garantiza el criterio de equiprobabilidad cuando un individuo es elegido al azar.

La afijación óptima que se ha aplicado tiene un sentido instrumental, ya que nos asegura la presencia en la muestra de aquellas características menos frecuentes en la población, pero, a la vez, sobredimensiona la presencia de los individuos de los estratos más heterogéneos que las poseen. Por otra parte, por el procedimiento seguido en la distribución y posterior redistribución de las cuotas muestrales, la ponderación *a posteriori* de la muestra restituye igualmente la proporcionalidad poblacional de los diferentes substratos definidos por la estratificación inicial de la muestra, el ámbito territorial y el tamaño del hábitat. En consecuencia, una vez obtenida la muestra, es preciso restituir el valor real de las frecuencias ponderando su peso en el conjunto y garantizar así una muestra aleatoria proporcionalmente dimensionada a su población.

La ponderación w_s que se aplica es una magnitud que transforma la probabilidad real de que un individuo, perteneciente a una sección censal, sea elegido en la probabilidad teórica bajo la hipótesis de estricta aleatoriedad, y se expresa con la relación siguiente:

$$w_s = \frac{\text{Probabilidad teórica}}{\text{Probabilidad real}} = \frac{N_s/N}{n_s/n}$$

donde:

- w_s es la ponderación de cada individuo de la muestra de la sección s ,
- N_s es la población de 16 y más años de la sección s , del año 2006,
- n_s es el número de individuos de la muestra de la sección s ,
- N es la población total de 16 y más años, del año 2006,
- n es el número de casos de la muestra.

Es decir, se otorga un menor peso a aquellos individuos que tienen una probabilidad mayor de ser elegidos, y un mayor peso a aquellos otros con una probabilidad menor.

El diseño de la muestra de la *Encuesta* implica la selección de personas a encuestar pertenecientes a unidades agregadas homogéneas como son las secciones censales. Como hemos indicado, este diseño conlleva una selección aleatoria de secciones censales, de forma que el conjunto de encuestas no se ha realizado en todas ellas; la concentración de cuotas de muestra realizada en algunos substratos implica que una parte de las encuestas sean expresión de un número de secciones y de población más amplio. La muestra final obtenida ha sido de 10.398 individuos que pertenecen a 4.415 secciones censales, elegidas aleatoriamente, del total de 5.219⁹. Estas secciones censales están distribuidas en 166 substratos, que abarcan una población de 5.356.760 personas de 16 y más años, según los datos recogidos en el Padrón Continuo a 1 de enero de 2006, con relación a una población catalana total de 16 y más años de 6.049.414, que se distribuye en 187 substratos.

Es decir, en relación con el total poblacional, el diseño muestral, en su objetivo de favorecer la concentración territorial del trabajo de campo, no recoge unidades muestrales con respecto a 692.654 personas de 16 y más años de la población del año 2006: de 21 substratos (cuya cuota muestral

9 Inicialmente las secciones eran 5.222, pero tres de ellas desaparecieron en el Padrón de Habitantes del año 2005. Inicialmente, también el tamaño de la muestra era de 10.446, pero la muestra final se cerró con 10.398 personas encuestadas.

era mínima e inferior a 0,5, y comprendía a 8.855 personas de la población) y de 804 secciones que no son elegidas en la muestra final (y que afectan a una población de 683.799 personas).

La ponderación de la muestra, que actúa en relación con las secciones elegidas y a su peso poblacional, restituye la representación del conjunto de la población de Cataluña de 16 y más años. A efectos de ponderación, y con el objetivo de abarcar la magnitud total de la población, se ha procedido a la redistribución de la población excluida entre las secciones con cuota muestral del mismo subestrato.

$$w_s = \frac{N_{s_{h_j}} + \frac{N_{s_{h_j}}}{\sum N_{s_{h_j}}} \cdot (N_{h_j} - \sum N_{s_{h_j}})}{n_{s_{h_j}}/n} \Big/ N = \frac{N_{s_{h_j}} + \bar{N}_{s_{h_j}}}{n_{s_{h_j}}/n} \Big/ N$$

donde:

- w_s es la ponderación de cada individuo de la muestra de la sección s ,
- $N_{s_{h_j}}$ es la población de 16 y más años de la sección s elegida en la muestra del subestrato h_j ,
- $\sum N_{s_{h_j}}$ es la población de 16 y más años de todas las secciones s elegidas en la muestra del subestrato h_j ,
- $\bar{N}_{s_{h_j}}$ es la población de 16 y más años de todas las secciones s no elegidas en la muestra del subestrato h_j ,
- N es la población total de 16 y más años,
- $n_{s_{h_j}}$ es el número de casos de la muestra pertenecientes a la sección s del subestrato h_j ,
- n es el número de casos de la muestra.

Esta redistribución se ha efectuado de forma proporcional al peso poblacional de las secciones del subestrato elegidas para la muestra. De esta forma, la ponderación w_s adopta la expresión:

Una vez realizada la ponderación de los datos muestrales, éstos se expresaran en magnitudes poblacionales a partir de la elevación de la muestra introduciendo un factor de ponderación f para cada individuo de la muestra, que se obtiene simplemente poniendo en relación el tamaño de la población y el tamaño de la muestra, a partir de la relación:

$$\text{Factor de elevación} = f = \frac{N}{n}$$

Finalmente, pues, se ha obtenido información de 10.398 personas encuestadas, que, una vez ponderadas, corresponden a un tamaño de muestra al que se le asocia un nivel de error global del 0,72%, para un nivel de significación de 2σ .

4. Aplicación de los resultados de la muestra

Para acabar, queremos apuntar brevemente algunos resultados o productos que se derivan del diseño y de la construcción de la muestra estratificada que hemos presentado. Más allá de los beneficios de eficacia estadística que el diseño comporta, el procedimiento seguido proporciona conclusiones de interés para el análisis sociológico y para el conocimiento social del territorio.

- a) En primer lugar, la modalidad del diseño de estratificación propuesta es de por sí un estudio sociológico, basado en datos censales de toda la población, sin error aleatorio salvo el de la medición, que nos proporciona un contenido social de gran valor como imagen de la estructura social de Cataluña.
- b) Este análisis e interpretación de la realidad social que supone la confección de la muestra estratificada tiene, a su vez, una base y un soporte territorial geográfico, la sección censal, que nos proporciona la posibilidad de elaborar una cartografía social de gran interés con una gran flexibilidad en la elección de los territorios y en su comparación.

- c) En este sentido, es posible reutilizar la información de la muestra con el objetivo de construir lo que hemos denominado zonas sociales (Lozares, y Domínguez, 1993, 1996; Lozares, López, y Flores, 2003), que no son otra cosa que los estratos y los resultados del análisis e interpretación de los datos censales en ellos, añadidos a los datos de la encuesta, puesto que la homogeneidad de los estratos permite un análisis autónomo de los mismos a partir de las encuestas realizadas en ellos. Así pues, el contenido social de estas zonas se enriquece con la información obtenida en la propia *Encuesta*, que facilita la validación de los resultados y proporciona una nueva imagen social de la realidad.
- d) La estratificación, a partir de los datos censales, tal y como se ha desarrollado, sirve además para ser aplicada en la realización de otras muestras y estudios que asuman las diferencias sociales expresadas en el territorio.

Bibliografia

- COCHRAN, W. G. *Técnicas de muestreo*. México: CECSA, 1971.
- GINER, S. [et al.]. *Enquesta de la Regió de Barcelona 2000. Informe General*. Barcelona: Mancomunitat de Municipis de l'Àrea Metropolitana de Barcelona i Diputació de Barcelona, 2002.
- GROSBAS, J. M. *Méthodes statistiques des sondages*. París: Ed. Economica, 1987.
- LÉBART, L. [et al.]. *Tratamiento estadístico de datos*. Barcelona: Ed. Marcombo, 1985.
- LÓPEZ ROLDAN, P.; LOZARES COLINA, C. *Anàlisi bivariàble de dades estadístiques*. Bellaterra: Universitat Autònoma de Barcelona, 1999. (Col·lecció Materials; 79)
- LÓPEZ, P. [et al.]. «Disseny i construcció d'una mostra estratificada a partir de dades censals». *Qüestió*, vol. 24, núm. 1 (2000), p. 111–136.
- LOZARES, C.; DOMÍNGUEZ, M. *Enquesta de la Regió Metropolitana de Barcelona 1990. Territori i realitat social: les zones sociodemogràfiques de la Regió Metropolitana de Barcelona*. Barcelona: Mancomunitat de Municipis de l'Àrea Metropolitana de Barcelona y Diputació de Barcelona, 1993.
- LOZARES, C.; DOMÍNGUEZ, M. «Tratamiento multivariado de subpoblaciones en una gran encuesta social: la construcción de zonas sociales». *Papers. Revista de Sociología*, núm. 48 (1996), p. 71–87.
- LOZARES, C.; LÓPEZ, P. *Enquesta Metropolitana de la Regió Metropolitana de Barcelona. Construcció de la mostra estratificada*. Sèrie Documents de Treball 90/1. Barcelona: Institut d'Estudis Metropolitans de Barcelona, 1990.
- LOZARES, C.; LÓPEZ, P. «El análisis de componentes principales. Aplicación al análisis de datos secundarios». *Papers. Revista de Sociología*, 37, (1991a), p. 31–63.
- LOZARES, C.; LÓPEZ, P. «El muestreo estratificado por análisis multivariado». En: LATIESA, M. *El pluralismo metodológico en la investigación social: ensayos típicos*, Granada: Universidad de Granada, 1991b, p. 107–160.
- LOZARES, C.; LÓPEZ, P. *Anàlisi multivariàble de dades estadístiques*. Bellaterra: Universitat Autònoma de Barcelona, 2000. (Col·lecció Materials; 93)
- LOZARES, C. [et al.]. *Zones socials per al territori de la província de Barcelona: una cartografia social*, Barcelona: Institut d'Estudis Regionals i Metropolitans de Barcelona, 2003.
- NEL·LO, O. [et al.]. *Enquesta Metropolitana de Barcelona. Condicions de Vida i Hàbits de la Població. La transformació de la societat metropolitana*. Barcelona: Mancomunitat de Municipis de l'Àrea Metropolitana de Barcelona i Diputació de Barcelona, 1998.
- SÁNCHEZ CARRIÓN, J. J. *Introducción a las técnicas de análisis multivariante aplicadas a las ciencias sociales*. Madrid: Centro de Investigaciones Sociológicas, 1984.

03 ORGANIZACIÓN DEL TRABAJO DE CAMPO

Meritxell Perramón (Instituto de Estudios Regionales y Metropolitanos de Barcelona)

Introducción

1. Etapa previa al trabajo de campo

- 1.1. Determinación de los ámbitos de actuación
- 1.2. Organización y funciones del equipo de trabajo de campo
- 1.3. Selección y formación del equipo de trabajo
 - 1.3.1. Formación de los jefes de equipo
 - 1.3.2. Selección y formación de los encuestadores

2. Etapas en el desarrollo del trabajo de campo

- 2.1. Distribución de la muestra entre los equipos de trabajo
- 2.2. Sistema de encuestación y funcionamiento de los equipos
- 2.3. Control de producción
- 2.4. Control de calidad
- 2.5. Supervisión: control telefónico

Introducción

La organización del trabajo de campo de la *Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006* supuso un reto por su envergadura: más de 10.000 entrevistas personales a una población repartida por un territorio de más de 32.000 km² y un plazo máximo de realización de un año. Así pues, se planeó toda la operación teniendo en cuenta la complejidad que se preveía y se dispusieron todos los mecanismos de control y de ajuste necesarios para superar los tropiezos propios del desarrollo del trabajo de campo.

En este artículo se describen las dos etapas fundamentales del desarrollo del trabajo de campo: la organización previa de los equipos de campo (ámbitos territoriales de actuación, definición de funciones y proceso de selección y formación de los integrantes de los equipos de campo y de coordinación) y la organización de los equipos de campo durante todo el tiempo de encuestación (distribución de la muestra, sistema de encuestación, control de producción y control de calidad).

1. Etapa previa al trabajo de campo

Los trabajos de preparación y organización del trabajo de campo realizados durante los meses anteriores al momento en que los encuestadores salen a la calle son fundamentales para el éxito de la operación. La organización del ámbito territorial de actuación y el proceso de selección y formación de los encuestadores son determinantes, ya que sirven para ubicar a estos últimos en el espacio a analizar y para asegurar una buena cualificación de las personas que llevarán a cabo la recogida de la información.

1.1. Determinación de los ámbitos territoriales de actuación

La ampliación del universo territorial de la *Encuesta de condiciones de vida y hábitos de la población* a toda Cataluña planteaba nuevos escenarios en la organización del trabajo de campo respecto a ediciones anteriores, no sólo porque significaba un importante incremento de la muestra y, por lo tanto, del equipo de trabajo de campo, sino también por el hecho de que la ampliación del territorio objeto de estudio hacía necesario distribuir el trabajo en diferentes equipos que lo abarcaran todo. En este sentido, se dividió Cataluña en tres grandes zonas y éstas a su vez en ámbitos de actuación más pequeños hasta un total de nueve, teniendo en cuenta tanto las divisiones territoriales administrativas como las características geográficas y el volumen de trabajo determinado por la distribución de la muestra. Como veremos más adelante, la organización del trabajo de campo se estableció en función de dichas zonas y ámbitos. En la tabla 1 se muestra la distribución territorial de las encuestas a realizar.

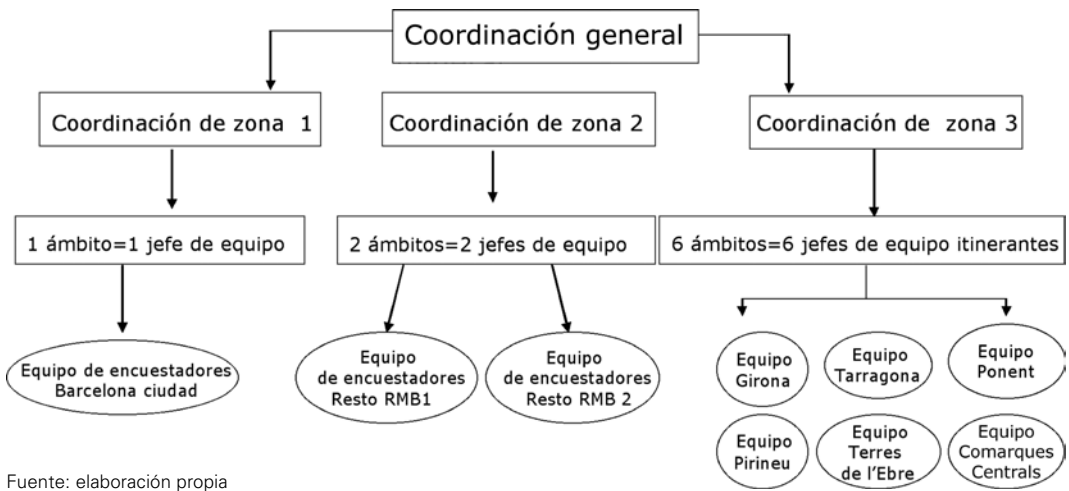
Tabla 1. Zonas y ámbitos territoriales utilizados en el trabajo de campo

Zona	Ámbitos territoriales	Muestra
1	Barcelona ciudad	1.910
2	Resto ámbito metropolitano 1: resto del Barcelonès, Baix Llobregat, Garraf y Alt Penedès	2.208
	Resto ámbito metropolitano 2: Maresme, Vallès Occidental y Vallès Oriental	1.842
3	Comarques Gironines	764
	Camp de Tarragona	772
	Terres de l'Ebre	589
	Àmbit de Ponent	777
	Alt Pirineu/Aran	593
	Comarques Centrals	943
	Total	10.398

1.2. Organización y funciones del equipo de trabajo de campo

En la organización de toda la operación de trabajo de campo pueden definirse cuatro perfiles de puesto de trabajo: coordinador general, coordinador de zona, jefe de equipo y encuestador. En esta organización de tipo piramidal se sitúa en primer lugar la figura del coordinador general, responsable de la dirección y coordinación metodológica, seguido de un coordinador de campo para cada una de las tres grandes zonas en las que se distribuye el ámbito territorial de estudio: la ciudad de Barcelona, el resto de la región metropolitana y el resto de Cataluña. Además, para cada uno de los nueve

Figura 1. Organización del trabajo de campo



ámbitos territoriales en que se subdividen las zonas hay un equipo de campo formado por un jefe de equipo y el correspondiente grupo de encuestadores. El número de encuestadores de cada equipo está en función de la muestra y de la dificultad de encuestación de la zona geográfica en cuestión (características demográficas y territoriales).

Las funciones a desarrollar por las personas que ocupan cada uno de los puestos de trabajo en el organigrama de organización del trabajo de campo son las siguientes:

a) Coordinador general:

- Coordinación y supervisión general de todo el trabajo de campo.
- Establecimiento de los criterios metodológicos.
- Resolución de incidencias y problemas metodológicos.

b) Coordinadores de zona:

- Coordinación, supervisión y resolución de problemas de los equipos de trabajo de su zona.
- Control de la distribución y redistribución de la muestra.
- Control de producción: establecimiento de las pautas de productividad y supervisión de las estadísticas de seguimiento.
- Control de calidad de la información recogida: control y verificación de encuestas y de los criterios metodológicos.
- Control de posibles desviaciones de la muestra y supervisión de las estadísticas de seguimiento.

c) Jefes de equipo:

- Organización del trabajo de los encuestadores: contacto directo.
- Control de producción: elaboración de las estadísticas de producción y seguimiento de las pautas de productividad.
- Control de la información recogida: revisión del 100% de los cuestionarios.
- Control de las sustituciones de muestra para evitar posibles desvíos.

d) Encuestadores:

- Organización de la ruta de trabajo.
- Contacto con las personas seleccionadas.
- Realización de las entrevistas y primera revisión.
- Solicitud de muestra de sustitución.

1.3. Selección y formación del equipo de trabajo

Como ya se ha avanzado, es necesario que las personas que integran el equipo de trabajo de campo reciban una formación adecuada a las funciones que van a desarrollar. Para ello se ha elaborado un plan de formación, tanto para los jefes de equipo como para los encuestadores. Por su parte, los coordinadores forman parte del equipo estable de la *Encuesta* del IERMB, por lo que son buenos conocedores de la metodología de la misma.

1.3.1. Formación de los jefes de equipo

Los jefes de equipo deben incorporarse al equipo de coordinación en la etapa previa al inicio de la operación de campo, con el objetivo de participar en el proceso de organización y recibir una formación específica para su puesto de trabajo.

Los jefes de equipo han de tener un conocimiento exhaustivo tanto del cuestionario (especialmente de las preguntas en que es más complicado obtener respuesta y/o de aquellas en las que se suelen dar más incoherencias) como del funcionamiento organizativo interno del trabajo de campo, así como del sistema de reparto y control de la muestra, del sistema de control de la producción y, en general, del proceso global de encuestación.

En concreto, los contenidos del plan de formación para los jefes de equipo son los siguientes:

- Antecedentes, objetivos y contextualización de la *Encuesta*.
- Organización de los equipos.
- Sistema de distribución, redistribución y control de la muestra.
- Conocimiento del cuestionario y sistema de encuestación.
- Sistemas de control.

La experiencia obtenida en la realización del trabajo de las ediciones anteriores de la *Encuesta* permite realizar una formación orientada a facilitar la tarea de recogida de la información, al control de calidad de las encuestas y a resolver las posibles dudas de los encuestadores en todos los ámbitos de su intervención.

1.3.2. Selección y formación de los encuestadores

Una vez formados los jefes de equipo empieza el proceso de selección de los encuestadores. Aunque no hay requisitos imprescindibles, es positivo tener en cuenta experiencias anteriores como encuestador y haber participado en otras operaciones de campo con cuestionarios de dificultad y longitud similares. También es muy importante que los futuros encuestadores tengan disponibilidad horaria, una buena actitud frente a las tareas a desarrollar y un buen nivel de catalán y de castellano, puesto que, en general, la población desea ser entrevistada en su lengua materna.

En esta operación de campo, donde el territorio de estudio es muy extenso, la selección de los entrevistadores se lleva a cabo por ámbitos territoriales, ya que en la medida de lo posible se intenta que puedan trabajar en su zona de residencia.

Antes de “salir a la calle”, los encuestadores reciben un curso de formación que consta de una parte teórica, que se lleva a cabo con el apoyo de un *Manual de instrucciones para el entrevistador* —donde se especifican punto por punto todos los contenidos de la formación, con un especial énfasis en el cuestionario—, y una parte práctica, en la que se realizan simulaciones de entrevista.

La formación que se da a los encuestadores se centra en los siguientes puntos:

- Antecedentes, objetivos y contextualización de la *Encuesta* e importancia del trabajo de los encuestadores en el conjunto del estudio.
- Sistema de encuestación e instrucciones para establecer contacto con la persona a entrevistar.
- Actitud general ante la persona a entrevistar.
- Conocimiento del cuestionario.

En el proceso de formación hay que incidir especialmente en la importancia de la tarea del encuestador en el conjunto del estudio: si la recogida de la información no es la correcta, la explotación y análisis posteriores de los datos no tendrá ninguna validez.

2. Etapas en el desarrollo del trabajo de campo

En este apartado se describen detalladamente el funcionamiento y las tareas a realizar por los diferentes miembros de los equipos de trabajo de campo, que garantizan la fiabilidad y la calidad de la información obtenida.

2.1. Distribución de la muestra entre los equipos de trabajo

Como se ha explicado en el subapartado 1.1., a nivel organizativo se distribuye el territorio de análisis y la muestra en nueve ámbitos de actuación. Las encuestas de cada ámbito serán llevadas a cabo por un equipo de trabajo de campo formado por un jefe de equipo y los correspondientes encuestadores. El primer paso es la distribución de la muestra entre estos últimos.

La *Encuesta de condiciones de vida y hábitos de la población de Cataluña* se realiza a una muestra de personas seleccionadas a partir del Censo de población. Así, se dispone de una lista de personas —que serán los Titulares de la muestra— facilitada por el Instituto de Estadística de Cataluña (Idescat), en la que además de los nombres y apellidos de los individuos seleccionados figura también la dirección, el sexo, la edad y el origen geográfico (por un lado, los nacidos en los 15 estados europeos que formaban la UE-15 y, por otro, los nacidos en el resto de países del mundo) de los mismos. Se dispone también de un listado de suplentes (siete por cada individuo) con las mismas características que el titular (misma sección censal de residencia, sexo, grupo de edad y origen), por si éste no pudiera o no quisiera contestar. En los casos en que en la sección censal no residan más individuos con las mismas características del titular, los suplentes serán personas residentes en secciones cercanas del mismo estrato social. Sólo en casos extremos, en los que no hay secciones del mismo estrato a una distancia razonable, se ha mantenido la sección original cambiando algunas de las características del titular, como el sexo o el grupo de edad.

Con el fin de distribuir la muestra entre los encuestadores y poder tener un control exhaustivo del proceso se realiza una hoja de seguimiento —denominada hoja de muestra— por cada persona seleccionada para ser entrevistada.

En la hoja de muestra consta el individuo (titular) a entrevistar y los datos personales que le caracterizan (dirección, origen geográfico, grupo de edad y sexo). Además, sirve para anotar los intentos realizados por los encuestadores para contactar con las personas seleccionadas y, si se da el caso, los motivos por los cuales no han querido o no han podido contestar la entrevista. En la misma hoja de muestra aparece también el espacio necesario para apuntar los datos personales de los posibles suplentes, que serán dados por el jefe de equipo en caso de no poder realizar la encuesta al titular.

Figura 2. Hoja de muestra



HOJA DE MUESTRA

Número de referencia	
Municipio	
Encuestador	

TITULAR			
Nombre, dirección, edad y origen geográfico	Día	Hora	Situación
Teléfono:			
Fecha entrevista:			
Motivo entrevista no realizada:			

SUSTITUTO 1			
Nombre, dirección, edad y origen geográfico	Día	Hora	Situación
Teléfono:			
Fecha entrevista:			
Motivo entrevista no realizada:			

SUSTITUTO 2			
Nombre, dirección, edad y origen geográfico	Día	Hora	Situación
Teléfono:			
Fecha entrevista:			
Motivo entrevista no realizada:			

Fuente: elaboración propia

Cada jefe de equipo tiene la responsabilidad de distribuir, entre los encuestadores que dirige, la muestra del ámbito territorial que debe controlar. En primer lugar, reparte las hojas de muestra de un territorio, municipio o distrito entre los que forman su equipo. El número de hojas de muestra a repartir irá en función de la experiencia del encuestador y de su disponibilidad temporal para realizar este trabajo. Cuando se trata de entrevistadores noveles es mejor dar un número de hojas de muestra reducido (entre 15 y 20). En función de las encuestas realizadas semanalmente y de la muestra que queda pendiente, el jefe de equipo considerará si tiene que repartir o no muestra nueva al encuestador y la cantidad necesaria para no limitar su producción.

Es muy importante que la muestra que se da a cada encuestador esté lo más concentrada posible en el territorio, ya que así podrá llevar a cabo más fácilmente su trabajo y, por lo tanto, de una manera más correcta. Es por ello que siempre que sea posible, sobre todo al principio del trabajo de campo, es aconsejable considerar las preferencias de los encuestadores al adjudicarles un área de trabajo.

2.2. Sistema de encuestación y funcionamiento de los equipos

Unos 10 o 15 días antes de que los encuestadores contacten con las personas a entrevistar, éstas reciben una carta de comunicación institucional que les informa de la realización del estudio, de la importancia de su colaboración para llevar el estudio a buen puerto, así como del hecho que se trata de una encuesta oficial. El objetivo de la carta es informar a la persona que ha sido seleccionada para la entrevista y facilitar así al encuestador el primer contacto con ella.

Sistema de encuestación

En el proceso de encuestación, el primer objetivo del encuestador es que sea la persona seleccionada como titular la que conteste el cuestionario. Con el fin de conseguir este objetivo se ha establecido el siguiente protocolo a seguir: el encuestador tiene que intentar contactar con la persona seleccionada (titular) en su domicilio, realizando un mínimo de cinco intentos antes de sustituirla por el primer suplente; si en alguno de estos contactos el entrevistador recibe una negativa, o se determina que no se puede realizar la encuesta (por defunción, enfermedad o cambio de domicilio de la persona a entrevistar, por un error censal, etc.), podrá pedir a su jefe de equipo una sustitución de forma automática; si no, tendrá que agotar los cinco intentos de contacto antes de pasar al primer suplente y repetir este proceso hasta conseguir realizar la encuesta.

Funcionamiento de los equipos

Los encuestadores tienen que visitar a su jefe de equipo una vez por semana y entregarle las encuestas que hayan realizado durante los días anteriores. El jefe de equipo anota la producción entregada y revisa por encima los cuestionarios, informando y corrigiendo al encuestador sobre los errores más evidentes observados en dicha primera revisión y sobre los que haya podido anotar de la revisión exhaustiva de las encuestas entregadas la semana anterior. Finalmente, comprueban conjuntamente la muestra pendiente y, si hace falta, se le entrega más muestra al encuestador. Durante el resto de días de la semana, el jefe de equipo está en contacto telefónico con el encuestador para resolver dudas que le puedan surgir y facilitarle suplencias de la muestra.

2.3. Control de producción

El control de la producción es exhaustivo y continuo. Así, se registran de forma detallada, una por una, todas las hojas de muestra entregadas a los encuestadores y todas las encuestas realizadas, siendo de esta manera como se puede conocer en todo momento el estado de la muestra: la parte realizada y la que queda pendiente. A partir de dicha información se realizan estadísticas de resumen para poder hacer balance del estado del trabajo de campo.

Figura 3. Ejemplo de hoja de control semanal

Encuestador:			Día y hora de visita:	
Fecha de entrega	Cuestionarios entregados al encuestador	Fecha de retorno	Cuestionarios realizados	Cuestionarios pendientes
8 septiembre	20	15 septiembre	18	2
15 septiembre	20 (+2) = 22	22 septiembre	14	8
.....				

Fuente: elaboración propia

Control de producción semanal

El jefe de equipo anota en una hoja de control semanal por encuestador la muestra entregada, la ya realizada (encuestas completadas) y la que le queda pendiente. En función de la producción semanal del encuestador y de la muestra pendiente, el jefe de equipo le entrega o no más muestra para que siga trabajando.

Control de producción mensual

El control de producción mensual resulta del sumatorio de los controles semanales. Con este dato se determina la producción mensual por encuestador.

Figura 4. Hoja de resumen de producción en un momento concreto del trabajo de campo

Estadísticas de supervisión: resumen producción							Día dd/mm/aaaa		
Ámbito territorial	1 Total cuestionarios a realizar	2 Total entregados a los encuestadores	3 Cuestionarios realizados ¹	T+ S1 + Os = 3 Cuestionarios realizados a:			4 Cuestionarios revisados ²	5 Cuestionarios supervisados ³	6 (1-3) Cuestionarios pendientes de realizar
				Titulares	Suplente 1	Otros suplentes			
Ámbito X	2000	1630	1504	962	400	142	1210	121	496
Ámbito Y	2000	1410	1196	721	363	112	1082	173	804
Ámbito Z	1000	765	705	583	102	20	607	66	295
TOTAL	5000	3805	3405	2266	865	274	2899	360	1595

¹ Cuestionarios incluidos en el registro de producción

² Cuestionarios leídos completamente por el jefe de equipo

³ Cuestionarios confirmados mediante llamada telefónica a la persona entrevistada (10% por encuestador como mínimo)

Fuente: elaboración propia

Aparte del control de producción mensual por encuestador, al finalizar cada mes también se realiza un cómputo de la muestra entregada y de las encuestas registradas por ámbitos territoriales.

En la hoja de control de producción que se muestra a continuación se recoge, para cada ámbito territorial de un caso ficticio, el total de encuestas a realizar, la muestra entregada a los encuestadores, las encuestas realizadas (tanto a titulares como a suplentes), las encuestas revisadas por los jefes de equipo, las supervisadas (vía telefónica) y las que quedan pendientes de hacer; todo ello en un momento determinado del trabajo de campo (que en el caso ficticio mostrado sería el dd/mm/aaaa).

Control final de producción

Al finalizar el trabajo de campo se realiza un recuento de todos los resúmenes mensuales de producción y se elabora una estadística final que recoge el calendario de producción real y el estado final de la muestra. El objetivo es poder contrastar estos datos con las previsiones realizadas y hacer un balance de situación que será de gran utilidad para próximas ediciones de la *Encuesta*.

2.4. Control de calidad

Para garantizar la validez de los datos obtenidos es necesario llevar a término un control de calidad exhaustivo del trabajo realizado. Dicho control de calidad va orientado tanto a validar la información obtenida a través de las encuestas como a garantizar la distribución de la muestra y el sistema de aleatoriedad.

Encuestas

Las encuestas entregadas por el encuestador pasan un doble control de calidad:

1. Primera revisión: la que se realiza en el momento de entrega de las encuestas. El jefe de equipo realiza una revisión rápida de los cuestionarios junto con el encuestador con el fin de comprobar que estén bien cumplimentados, que no se haya olvidado preguntas, así como para solucionar posibles errores de comprensión. El tiempo de duración de dicha revisión estará en función de la experiencia del encuestador; así, lógicamente, a medida que va avanzando el trabajo de campo se necesita menos tiempo para realizar esta primera revisión.
2. Revisión a fondo: todas las encuestas pasan una segunda revisión por parte del jefe de equipo. En ella, además de comprobar que no falte ninguna pregunta por contestar, se revisa la coherencia de las respuestas recogidas y los posibles errores o irregularidades. A partir de esta revisión a fondo se rellena una hoja por encuestador, donde se recogen sus principales incidencias y sus errores más comunes, confusiones en determinados términos, etc., para poder comentárselo en su día de visita.

Muestra

Todas las encuestas entregadas por el encuestador al jefe de equipo se registran en una base de datos de control donde, además de controlar la producción, se recoge la información necesaria para conocer en todo momento las posibles desviaciones de la muestra que se vayan produciendo.

En unas hojas de control donde se encuentran anotados los datos (ámbito territorial, comarca, municipio, distrito, sección censal, origen, sexo y edad) de la persona seleccionada inicialmente (Titular), se introducen los datos del individuo finalmente entrevistado (Suplente 1, S2, S3...), en caso que no sea el titular. En estos últimos casos se recogen también los motivos de sustitución (defunción, negativa, cambio de domicilio, etc.). La consulta de dicha base de datos permite tener un control sistemático de las posibles desviaciones muestrales y corregirlas lo antes posible en el mismo proceso de campo.

Figura 5. Hoja de control de la muestra

Número cuestionario	DATOS DEL TITULAR											Titular o Suplente	Motivo suplencia														
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)		S1	S2	S3	S4	S5	S6									
1	DATOS DEL TITULAR (vienen dados por la base de datos de la muestra)																										
2																											
3																											
4																											
5																											
6																											
7																											
8																											
9																											
10																											
11																											
12																											
13																											
14																											
...																											

Número cuestionario	DATOS DE LA PERSONA ENTREVISTADA											Comparación entre los datos del titular y los de la persona entrevistada *										
	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)		
1	Com.	Cód. mun.	Dist.	Sec.	Est. soc.	Orig. Geo.	Sexo	Grup edad	Subg edad	Año nac.	Com.	Cód. mun.	Dist.	Sec.	Est. soc.	Orig. Geo.	Sexo	Grup edad	Subg edad	Año nac.		
2																						
3																						
4																						
5																						
6																						
7																						
8																						
9																						
10																						
11																						
12																						
13																						
14																						
...																						

* Automáticamente, el programa informático indica en cada casilla si hay o no coincidencia entre los datos del titular y los de la persona entrevistada.

Fuente: elaboración propia

2.5. Supervisión: control telefónico

La supervisión de los cuestionarios se realiza mediante control telefónico. Tal como marca la normativa de calidad¹, se supervisa un mínimo del 10% de la producción realizada por cada encuestador. La supervisión realizada en el trabajo de campo de la *Encuesta de condiciones de vida y hábitos de la población* puede ser de dos tipos:

Control sistemático: se escoge un número al azar y, automáticamente, todas las encuestas registradas acabadas en dicho número pasan a ser supervisadas.

Control no sistemático: deriva del control de calidad. Cuando los jefes de equipo detectan algún error o incoherencia en la información de alguna encuesta, ésta pasa a ser supervisada con el fin de solucionar el error y determinar el motivo.

Como puede observarse, hay una parte de comprobación general (si la persona seleccionada ha contestado la encuesta, su duración aproximada, el comportamiento del encuestador...), una parte reservada al control sistemático del cuestionario (en la que se realizan unas preguntas de comprobación iguales para todas las llamadas) y una parte de control no sistemático, en la que no se especifican las preguntas a realizar, ya que dependerán de la incidencia o error detectado en cada caso. Si se detecta alguna irregularidad grave o alguna encuesta que haya sido falseada, se recuperan inmediatamente todos los cuestionarios realizados por el encuestador involucrado y se supervisan detenidamente uno por uno. Las encuestas falseadas se anulan. (Véase cuestionario telefónico en la página siguiente).

¹ Normativa Asomar 2006: UNE-ISO 20252:2006.

Figura 6. Ficha de control telefónico



**INSTITUT D'ESTUDIS
REGIONALS I METROPOLITANS
DE BARCELONA**



**Generalitat de Catalunya
Institut d'Estadística
de Catalunya**

Questionario de control telefónico

Questionario Núm: _____ Núm de teléfono: _____
 Fecha control: _____
 Persona Control: _____
 Encuestador: _____
 Control:

Sistemático No

sistemático

Identificación

Persona de contacto: _____

Parentesco con el entrevistado

1. Entrevistado	2. Pareja	3. Padre/madre
4. Hijo/a	4. Otros miembros del hogar	6. Otros. Especificar:

Comprobación general

¿Ha contestado el/la entrevistado/a el cuestionario? 1. Sí 6. No

Fecha aproximada: _____

Duración de la entrevista: _____

Comportamiento / actitud del encuestador: 1. Muy correcto / 2. Correcto / 3. Incorrecto

Control sistemático

¿Cuántas personas viven habitualmente en la vivienda del entrevistado?

¿Cuál es la situación laboral del entrevistado?

¿Disponen de lavavajillas en la vivienda del entrevistado? 1. Sí 6. No

¿Practica el entrevistado algún deporte de manera habitual? 1. Sí 6. No

¿Hubo alguna pregunta que le molestó al entrevistado? 1. Sí 6. No 0. NP

En caso afirmativo, ¿cuál? _____

Control no sistemático

Ver cuestionario (con problemas)

Seguimiento			
Día	Hora	Persona contacto	Situación

Fuente: elaboración propia

04 ESTADÍSTICAS DE SEGUIMIENTO COMO DISPOSITIVO DE CALIDAD DEL TRABAJO DE CAMPO EN ESTUDIOS POR ENCUESTA

Sergio Porcel (Instituto de Estudios Regionales y Metropolitanos de Barcelona)

Introducción

1. La base de datos relacional

- 1.1. ¿Por qué una base de datos relacional?
- 1.2. Estructura y funcionamiento de la base de datos relacional

2. Estadísticas de seguimiento

- 2.1. Control de la producción del trabajo de campo
 - 2.1.1. Control territorial de la recogida de datos
 - 2.1.2. Control del circuito de producción
- 2.2. Control de la calidad de la muestra

3. Conclusiones

Introducción

Este artículo plantea una reflexión sobre cómo el uso intensivo del seguimiento estadístico en los procesos de trabajo de campo de los estudios por encuesta puede incidir en la calidad de los datos que se obtienen. Para ello, el texto se centra en el papel que han jugado las estadísticas de seguimiento en el trabajo de campo de la *Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006*.

A pesar de que *a priori* pueda parecer que el seguimiento estadístico tenga poco que ver con cuestiones relacionadas con la calidad, lo cierto es que su uso muestra un gran potencial en la mejora del tratamiento de la información, en la optimización de los recursos y en la reducción de la no respuesta, entre otros aspectos.

Se puede decir que los efectos más visibles en términos de calidad que se producen en los datos de una encuesta son los que provienen de los controles de calidad en sentido estricto, que actúan directamente sobre el contenido de los cuestionarios, como la revisión de cuestionarios, el control telefónico, la depuración de datos, etc. Pero, como explica Cea D'Ancona, los errores que merman la calidad de una encuesta pueden aparecer durante todo el proceso de su realización (Cea

D'Ancona, 2005). Según la autora, el diseño muestral, la actitud y capacidades del encuestador, la predisposición del encuestado a colaborar o no, etc. son ejemplos de fuentes de error previas a la obtención del contenido del cuestionario que también se han de considerar. Por lo tanto, siguiendo esta lógica, en la medida en que las estadísticas de seguimiento favorecen la organización de los equipos que intervienen en el trabajo de campo y mejoran y evalúan la producción, éstas acaban aportando calidad a la información que se recoge.

Aún así, las posibles fuentes de error en un estudio por encuesta (como en otras técnicas de investigación) son tan amplias que resulta prácticamente imposible controlarlas todas. Por ello resulta imprescindible controlar y documentar todo el proceso para conocer los posibles sesgos o limitaciones de los datos obtenidos. En este sentido, el seguimiento estadístico también ofrece grandes posibilidades.

Aunque los efectos sobre la calidad que se derivan de las estadísticas de seguimiento son, por lo general, difíciles de medir y pueden ser considerados más sutiles o menos palpables que en el caso de los controles de calidad en sentido estricto, su aportación a la calidad parece innegable.

Así pues, las estadísticas de seguimiento constituyen el eje central de este artículo, en el que se abordarán cuestiones como el tipo de sistema de gestión y control de la información que se ha utilizado en el trabajo de campo de la *Encuesta*, los diferentes tipos de estadísticas que se han elaborado, su función en el contexto de este proyecto y su posible incidencia respecto a la calidad.

1. La base de datos relacional

Un elemento fundamental, gracias al cual es posible llevar a cabo un uso intensivo y eficaz de las estadísticas de seguimiento en trabajos de campo complejos como el de la *Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006*, es la base de datos basada en el modelo relacional. La base de datos relacional es el instrumento que se ha utilizado para recoger y administrar toda la información del trabajo de campo, que, dadas sus características, ofrece grandes posibilidades que merecen ser tratadas.

1.1. ¿Por qué una base de datos relacional?

La decisión de utilizar en este proyecto una base de datos relacional, como instrumento de control y gestión de la información, viene precedida de una reflexión sobre la idoneidad que presenta esta herramienta para superar los retos que plantea un trabajo de campo de esta envergadura. El interés por obtener información de la máxima calidad posible a través de la técnica de encuesta exige que cada una de las fases del proceso de creación de la base de datos se desarrolle bajo altas cuotas de control, lo que, a su vez, plantea la necesidad de disponer de información constante sobre diversos aspectos del trabajo de campo. Información que, por un lado, debe resultar útil en términos organizativos y productivos durante la ejecución de la operación y, por otro lado, ha de proporcionar un conocimiento exhaustivo de las diferentes fases del proceso y de la naturaleza de los datos recogidos.

Si se tienen en cuenta algunos elementos que caracterizan el trabajo de campo de la *Encuesta*, como las condiciones de fragmentación y simultaneidad en que se llevan a cabo las diferentes fases (recogida de datos, revisión, codificación, grabación y depuración), o el complejo y numeroso equipo técnico que conforma la encuesta (25 técnicos y alrededor de 70 encuestadores), es posible intuir la dificultad que encierra el ejercicio de recopilar y gestionar toda esta información. En este sentido, la necesidad de tener acceso constante a información actualizada durante la ejecución del proceso y el hecho de descentralizar el almacenaje de los datos —centralizando al mismo tiempo su gestión— son las principales cuestiones que resuelve el modelo relacional.

Edgar Frank Codd, uno de los teóricos más importantes de este modelo de base de datos, atribuye al modelo relacional grandes cualidades en el control de la producción, precisamente por su capacidad de informar en cualquier momento del estado de una parte o de la totalidad de la actividad que se lleva a cabo en una determinada empresa (Codd, 1990). Según Codd, el modelo relacional es capaz de garantizar consistencia, integridad y precisión en el manejo de los datos, aspectos imprescindibles en el contexto del seguimiento de la producción.

Respecto a la manera de gestionar la información, este modelo considera la base de datos como una colección de relaciones entre diferentes tablas, que al mismo tiempo realizan la función de fuentes de datos. Las tablas vienen determinadas por filas (registros) y columnas (campos), donde cada fila presenta una serie de valores sobre determinados atributos o variables (columna) correspondientes a una entidad o caso. El lugar y la forma en que se almacenan los datos es irrelevante, ya que los datos concernientes a cada caso quedan vinculados, conservando su integridad aunque tengan ubicaciones diferentes, lo que diferencia claramente este modelo de otros, como el modelo jerárquico o el modelo en red. Esta característica facilita que diferentes actores puedan introducir datos independientemente y de manera simultánea en diferentes tablas, sin que esto afecte a la posterior obtención de la información.

Según el modo de funcionamiento de una base de datos relacional, se puede extraer información desde todas las tablas (fuentes de datos) en la medida en que los datos, en sus diferentes ubicaciones, permanecen relacionados a partir de un mismo valor (ID), que se encuentra presente en todas estas fuentes. Este valor permite identificar de manera unívoca los diferentes casos y debe cumplir con las propiedades de unicidad (ha de ser un valor único) y minimalidad (ha de constituir una unidad mínima). De esta manera, los datos que se obtienen de las diversas tablas se basan en valores coincidentes de los campos, que la relación entre las tablas establece como comunes. La información puede ser recuperada a través de "consultas" *ad hoc*, ofreciendo una amplia flexibilidad para administrar los datos. Este mecanismo posibilita también que sea más fácil de utilizar para usuarios esporádicos de la base de datos.

En el caso de la *Encuesta*, el modelo relacional integra los datos de las diferentes fases del trabajo de campo, desde el registro de cuestionarios realizados hasta la grabación de los mismos, pasando por la revisión y la codificación. Las diferentes tablas, donde de manera independiente cada equipo ocupado de una fase del trabajo de campo introduce la información sobre las diferentes actividades que desarrolla, quedan vinculadas a través de los valores (ID) de la variable *nº de cuestionario*, presente en todas ellas. Como se ha explicado anteriormente, las características del modelo relacional permiten que a través de consultas se obtenga la información solicitada, ya sea sobre los cuestionarios o sobre cualquier otro aspecto recogido, independientemente de la fuente de datos donde se encuentre.

Resumiendo, entre las principales ventajas que presenta el modelo relacional se pueden destacar las siguientes (Codd, 1990):

- Garantiza herramientas para evitar la duplicidad de los registros.
- Garantiza la integridad referencial, de manera que al eliminar un registro elimina todos los registros relacionados dependientes.
- Garantiza la integridad de los datos.
- Favorece su uso en equipo, posibilitando el trabajo independiente desde diferentes terminales.
- Aporta flexibilidad en el tratamiento de los datos.

1.2. Estructura y funcionamiento de la base de datos relacional

La base de datos relacional diseñada para gestionar la información del trabajo de campo de la *Encuesta* consta de cuatro tablas o fuentes de datos (en formato *x/s*) vinculadas entre sí. La información que contienen estas tablas se refiere al registro de cuestionarios realizados (registrados),

al de cuestionarios codificados, al de cuestionarios grabados y al índice de cajas donde se guardan los cuestionarios grabados agrupados en microbases de datos (BD)¹. Como se puede observar en la figura 1, la tabla *Registrados* se crea a partir de fusionar las tres tablas diferentes con las que cada coordinador territorial trabaja de manera independiente. También se puede apreciar que, en el caso de las tablas *Registrados*, *Codificados* y *Grabados*, la relación se establece a partir de los valores de la variable *nº de cuestionario*, mientras que entre la tabla *Grabados* y el *índice de cajas* se utiliza el número de la microbase de datos (*nº BD*).

Las *tablas de cuestionarios registrados* son gestionadas por los tres coordinadores de campo, que se encargan del desarrollo del trabajo de campo de los tres ámbitos territoriales en que se ha dividido Cataluña: Barcelona ciudad, resto de la región metropolitana de Barcelona y resto de Cataluña. Estas tablas son las que recogen mayor volumen de información. En ellas se introducen datos sobre los siguientes aspectos:

- Nº de cuestionario
- Encuestador al que se le entrega la muestra
- Variables territoriales teóricas (según la muestra inicial): provincia, ámbito del Plan Territorial de Cataluña, comarca, municipio, distrito, sección censal
- Estrato social teórico (según la muestra inicial)
- Origen geográfico teórico (según la muestra inicial)
- Sexo teórico (según la muestra inicial)
- Grupo de edad teórico (según la muestra inicial)
- Variables territoriales reales: provincia, ámbito del Plan Territorial de Cataluña, comarca, municipio, distrito, sección censal
- Estrato social real
- Origen geográfico real
- Sexo real
- Grupo de edad real
- Sujeto al que se le ha realizado la encuesta: titular o suplente
- Motivos de sustitución (si no se realiza al titular)
- Encuestador que ha realizado la encuesta
- Fecha de recepción
- Revisión
- Control telefónico
- Comentarios (como, por ejemplo, si el cuestionario ha sido anulado en esta fase por algún motivo)

Para facilitar la gestión de las tablas, los coordinadores cuentan con la parte de datos teóricos de la muestra inicial, pertenecientes a los titulares, ya introducidos, incluyendo el *nº de cuestionario* y la información relativa a las variables muestrales teóricas. Así, la dinámica de funcionamiento en el registro de datos de estas tablas se articula en dos fases:

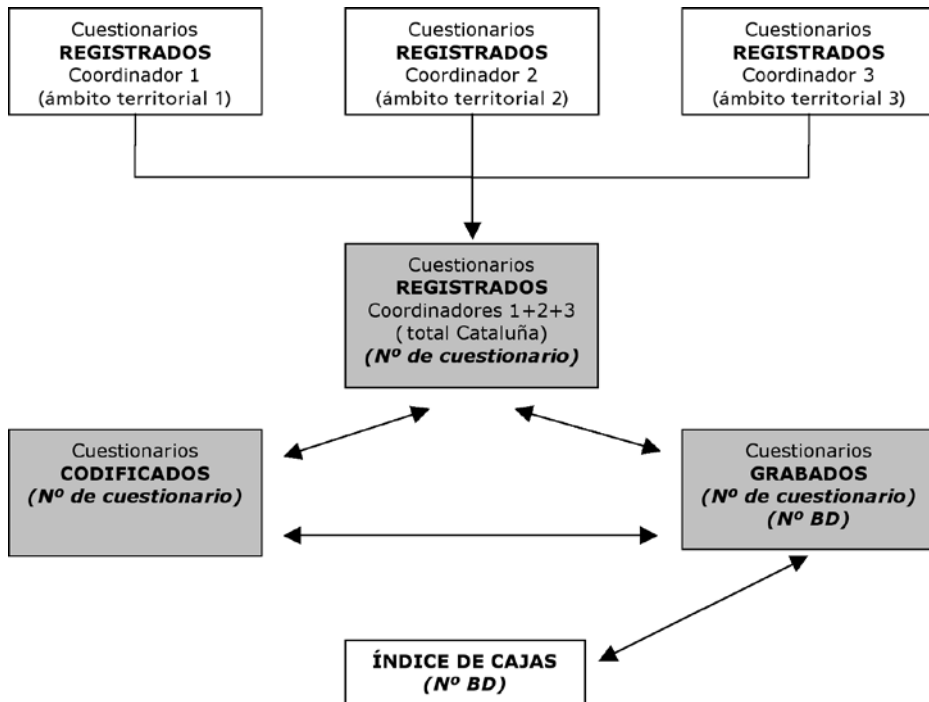
- *Entrega de muestra* (salida de cuestionarios): los coordinadores simplemente asignan fragmentos de la muestra a los encuestadores.
- *Registro de cuestionarios* (entrada de cuestionarios): a medida que los coordinadores reciben cuestionarios realizados se registran los datos reales² del cuestionario.

Para realizar el cálculo de las estadísticas de seguimiento generales, las tablas de cada coordinador se integran en una sola que aglutina toda la información. Pero aparte de la utilización de todos estos datos en la elaboración de las estadísticas de seguimiento generales, cada coordinador también dispone y hace uso de los datos concernientes al territorio que gestiona. Este hecho le permite

1 Cada microbase de datos (BD) corresponde a los cuestionarios grabados cada día por un grabador.

2 Los datos reales del cuestionario pueden diferir de los teóricos como consecuencia de errores del Padrón o de la utilización de suplentes con alguna característica diferente a las del titular.

Figura1. Estructura de la base de datos relacional utilizada en la Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006



Fuente: elaboración propia

mantener un control absoluto sobre diferentes aspectos de la producción en su territorio, tales como la cantidad de cuestionarios realizados, los cambios de cuotas que se pueden ir produciendo, el nivel de control telefónico, etc. Esta información es de gran utilidad durante el proceso de trabajo de campo, sobre todo en términos de organización y planificación estratégica.

Por su parte, la *tabla de cuestionarios codificados* se elabora a partir de los datos generados por el coordinador de codificación, y se trata de una tabla mucho más simple que la anterior, donde se almacena información sobre las siguientes variables:

- Nº de cuestionario
- Fecha de codificación (día/mes/año)
- Semana de codificación
- Nombre del codificador
- Anulado (si el cuestionario ha sido anulado en esta fase por algún motivo)

Por sí sola, esta fuente de datos permite conocer si un cuestionario determinado está codificado o no, así como el codificador y la fecha en que se codificó. Esta información, entre otras cosas, resulta imprescindible cuando por alguna razón se precisa buscar algún cuestionario. Por otro lado, también aporta información relevante para evaluar el ritmo de producción del equipo de codificación, lo cual resulta vital a la hora de planificar y de alcanzar objetivos concretos. La variable *anulada* marca si el número de cuestionario en concreto está anulado o no y, de esta forma, permite mantener el control sobre los cuestionarios que, estando ya codificados, por alguna razón no son aceptados como válidos.

La *tabla de cuestionarios grabados* tiene un formato muy parecido a la de cuestionarios codificados. La información que recoge se refiere a las siguientes variables:

- N° de cuestionario
- Fecha de grabación (día/mes/año)
- Semana de grabación
- Nombre del grabador
- N° de microbase de datos (n° BD)
- Anulado (si el cuestionario ha sido anulado en esta fase por algún motivo)

Las prestaciones que ofrece esta fuente de datos son prácticamente las mismas que en el caso anterior, pero adaptadas a la fase de grabación. Así, esta tabla aporta la información suficiente para saber si un cuestionario está o no grabado, cuándo se ha grabado y por quién, proporcionando las pistas suficientes para llevar a cabo la búsqueda de un cuestionario concreto. También permite controlar la evolución del proceso de producción e identificar los cuestionarios anulados en esta fase por diversos motivos.

Por último, la *tabla índice de cajas* desarrolla únicamente la función de facilitar la búsqueda de cuestionarios. Tras el proceso de grabación, los cuestionarios permanecen agrupados en montones —que se corresponden con las microbases creadas diariamente por cada grabador—, que quedan ubicados en cajas numeradas listas para ser depurados. En la mayoría de los casos, dichos montones oscilan entre los 20 y los 50 cuestionarios. La *tabla índice de cajas* tan sólo recoge tres variables: *n° de caja*, *n° de microbases de datos* y *cantidad de cuestionarios por microbase de datos*. Esta fuente no incluye la variable *n° de cuestionario*, ya que, de cara al proceso de depuración, son las microbases de datos las que juegan un papel más relevante en términos de identificación. Por eso en este caso los valores (ID) que se utilizan para establecer la relación con la *tabla de cuestionarios grabados* son los de la variable *n° de microbase de datos*, y no los de *n° de cuestionario* como en el resto de fuentes (véase figura 1).

2. Estadísticas de seguimiento

Como ya se ha apuntado, el seguimiento estadístico del trabajo de campo constituye una herramienta esencial en términos estratégicos y organizativos, sobre todo en un proyecto de las dimensiones de la *Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006*. En este sentido, el uso de estadísticas incide en la calidad del estudio, pero, además, el hecho de disponer de toda la información que se desprende del trabajo de campo también hace posible desarrollar diversos análisis que evalúen la calidad de los datos recogidos, aparte de otros aspectos que pueden continuar actuando de diferentes maneras en términos de calidad. Como expresa Francisco Alvira, este tipo de información y de análisis permite conocer mucho mejor los datos recogidos antes de entrar en análisis descriptivos o explicativos y ayuda a afinar más en las interpretaciones (Alvira, 2004).

Así, las estadísticas de seguimiento en este proyecto desempeñan básicamente dos funciones, ambas relacionadas con la calidad:

- Aportar en cualquier momento información sobre la situación de producción de los diferentes procesos del trabajo de campo.
- Ofrecer datos que permitan realizar un análisis de la calidad de la muestra obtenida, así como otros tipos de análisis relacionados con otros aspectos del trabajo de campo.

2.1. Control de la producción del trabajo de campo

Respecto al seguimiento que se lleva a cabo de la producción del trabajo de campo se puede diferenciar, por un lado, el control de la recogida de datos, que se realiza desde una perspectiva

territorial, y, por otro, el control del circuito de producción, realizado desde una óptica más general e integradora de los diferentes procesos que conforman el trabajo de campo.

2.1.1. Control territorial de la recogida de datos

El control territorial de la recogida de datos se realiza a través de dos grupos de tablas:

- Producción de encuestas por mes y ámbito territorial
- Producción de encuestas acumuladas por mes y ámbito territorial

Cada uno de estos grupos está formado por las mismas cuatro tablas, con la única diferencia de que, en el primer caso, los datos hacen referencia a la producción concreta de cada mes y, en el segundo, la producción se presenta acumulada. La información que recogen estas tablas es la siguiente:

- Producción teórica (por mes/acumulada) y ámbito territorial (tabla 1)
- Producción real (por mes/acumulada) y ámbito territorial (tabla 2)
- Diferencial real–teórica (por mes/acumulada) y ámbito territorial (tabla 3)
- Índices de cumplimiento de la producción (mensual/acumulada) prevista (tabla 4)

Estas informaciones, cada una de ellas recogida en una tabla diferente, se presentan con datos relativos a la totalidad del territorio objeto de estudio y desagregada para cada uno de los ámbitos territoriales que gestiona cada coordinador de campo.

El objetivo que se persigue con estas tablas es tener la posibilidad de evaluar, en cualquier momento, el volumen de producción conseguida frente a un valor de referencia teórico, para así conocer y controlar el curso de la ejecución del trabajo de campo.

El valor de referencia teórico se calcula distribuyendo el total de la muestra teórica por los meses durante los que se prevé que se desarrolle el trabajo de campo, teniendo en cuenta los días laborables de cada mes, tal como se muestra en la tabla 1 en relación a un caso ficticio. De la misma manera, se calculan también los valores teóricos de referencia acumulados.

Tabla 1. Producción teórica por mes y ámbito territorial (nº absolutos)

Ámbitos territoriales	Sep	Oct	Nov	Dic	Ene	Feb	Mar	Abr	May	Jun	Jul	Total
Días laborables del mes para trabajar	10	20	21	14	21	20	21	14	22	21	16	200
Ámbito X	100	200	210	140	210	200	210	140	220	210	160	2.000
Ámbito Y	100	200	210	140	210	200	210	140	220	210	160	2.000
Ámbito Z	50	100	105	70	105	100	105	70	110	105	80	1.000
Total	250	500	525	350	525	500	525	350	550	525	400	5.000

Por otro lado, una tabla similar a la 1 sirve para recoger la distribución de la producción real, tanto mes a mes como acumulada. Así, a través de una simple resta entre las casillas de ambas tablas se obtiene la diferencia entre los valores teóricos de referencia y la producción real, lo que hace posible evaluar dicho diferencial para cada mes, con datos sin acumular y acumulados.

Para el caso ficticio considerado, la tabla 2 refleja esta información sin acumular (fecha de mayo a modo de ejemplo), en la que se pueden observar diferencias importantes de producción mensual entre los diferentes territorios analizados. En este sentido, se puede destacar el caso del ámbito territorial X, que presenta, entre los meses de septiembre y diciembre, un superávit de producción respecto a los valores teóricos, mientras que a partir de enero registra, lógicamente, una disminución considerable de la producción. El territorio Y, en cambio, arrastra una producción deficitaria en todos los meses desde el inicio, salvo en el mes de abril, donde supera las previsiones teóricas, de manera que, en los tres meses que quedan de trabajo de campo (mayo, junio y julio), se deberá realizar un esfuerzo importante en dicha zona. La producción del ámbito Z también ha ido algo retrasada hasta febrero, si bien mucho menos que en el territorio Y.

Tabla 2. Diferencial real–teórico por mes y ámbito territorial (nº absolutos), 1 de mayo

Ámbitos territoriales	Sep	Oct	Nov	Dic	Ene	Feb	Mar	Abr	May	Jun	Jul	Total
Días laborables del mes para trabajar	10	20	21	14	21	20	21	14	22	21	16	200
Ámbito X	20	34	38	26	-8	-4	-8	-5	-	-	-	93
Ámbito Y	-80	-45	-48	-18	-20	-18	0	16	-	-	-	-213
Ámbito Z	-27	-25	-10	0	5	-12	32	37	-	-	-	0
Total	-87	-36	-20	8	-23	-34	24	48	-	-	-	-120

Pero para lograr una visión más amplia de la producción a nivel territorial, la lectura de estos datos sin acumular debe ser complementada necesariamente con la lectura de datos acumulados, ya que, a medida que avanza el trabajo de campo, los datos sin acumular quedan descontextualizados y se pierde la trayectoria de la producción. En cambio, los datos acumulados ofrecen una visión más amplia del recorrido del trabajo de campo y permiten conocer cuál es la situación de la producción total de cada territorio en un momento concreto, a pesar de los superávits o déficits de producción registrados mes a mes. Para ello, se calcula la diferencia entre la estimación de la producción teórica acumulada para cada mes y la producción real mensual acumulada.

Observando los datos acumulados (tabla 3) se puede apreciar, por ejemplo, que la producción real de cuestionarios en el territorio X hasta el mes de mayo supera con creces las previsiones teóricas, de manera que prácticamente no es relevante que en los últimos meses la producción haya estado algo por debajo de las previsiones (tabla 2). Por su parte, la producción del ámbito Z va según las previsiones previstas, y a 1 de mayo se ha realizado justo el número de cuestionarios de referencia teórico, mientras que los datos del territorio Y confirman el importante déficit de encuestas que sufre a 1 de mayo.

Tabla 3. Diferencial real–teórico acumulado por mes y ámbito territorial (nº absolutos), 1 de mayo

Ámbitos territoriales	Sep	Oct	Nov	Dic	Ene	Feb	Mar	Abr	May	Jun	Jul	Total
Días laborables del mes para trabajar	10	20	21	14	21	20	21	14	22	21	16	200
Ámbito X	20	54	92	118	110	106	98	93	-	-	-	-
Ámbito Y	-80	-125	-173	-191	-211	-229	-229	-213	-	-	-	-
Ámbito Z	-27	-52	-62	-62	-57	-69	-37	0	-	-	-	-
Total	-87	-123	-143	-135	-158	-192	-168	-120	-	-	-	-

La última información que se utiliza en el control territorial de la producción, tanto mes a mes como con datos acumulados, son los índices de cumplimiento de las previsiones. Este índice sintetiza la información expuesta en las tablas anteriores y permite la comparabilidad de los resultados entre los diferentes territorios. Se calcula dividiendo la producción real entre la teórica, de manera que el valor 1 representa el ajuste perfecto entre la previsión y la producción real. En cambio, los valores inferiores a 1 se interpretan como una producción deficitaria, mientras que los que superan la unidad reflejan una producción superior a las previsiones.

La tabla 4 recoge un ejemplo de este tipo de información en el que los índices de cumplimiento están calculados en base a datos acumulados. En ella se puede confirmar que, a 1 de mayo, el territorio que presenta la producción acumulada más abundante es el X, mientras que el que presenta un mayor problema de producción es el Y. En la misma fecha, el ámbito territorial Z, tal como ya apuntamos, se ajusta perfectamente a las previsiones.

Tabla 4. Índices de cumplimiento de la producción prevista acumulada, 1 de mayo

Ámbitos territoriales	Sep	Oct	Nov	Dic	Ene	Feb	Mar	Abr	May	Jun	Jul	Total
Días laborables del mes para trabajar	10	20	21	14	21	20	21	14	22	21	16	200
Ámbito X	1,20	1,18	1,18	1,18	1,13	1,10	1,08	1,07	-	-	-	-
Ámbito Y	0,20	0,58	0,66	0,71	0,75	0,78	0,82	0,85	-	-	-	-
Ámbito Z	0,46	0,65	0,76	0,81	0,87	0,87	0,94	1,00	-	-	-	-
Total	0,65	0,84	0,89	0,92	0,93	0,93	0,95	0,97	-	-	-	-

El objetivo principal de este conjunto de tablas es mitigar o regular los desequilibrios en la consecución de entrevistas a nivel territorial, incidiendo de esta manera en la calidad de los datos recogidos. En este sentido, su incidencia en la calidad tiene que ver con la importancia de mantener un cierto equilibrio en la evolución de las cuotas territoriales durante el trabajo de campo, para evitar así que los desequilibrios se mantengan también en un posible cierre precipitado de la campaña y afecten a la representatividad de la muestra.

De esta manera, a partir de la información que ofrecen estas tablas es posible detectar problemas de producción en los territorios asignados a cada coordinador, a la vez que se pueden trazar estrategias de reorganización de los recursos humanos destinados a cada territorio si se considera necesario. Por ejemplo, si tras analizar los datos se detectan problemas de producción en una zona concreta (como en el ámbito Y de nuestro caso hipotético), se pueden desplazar encuestadores (desde el territorio X, por ejemplo) u otro tipo de recursos al territorio en cuestión para reforzar la producción o tomar, según se valore la situación, otro tipo de decisiones.

La ventaja de disponer en cualquier momento de esta información actualizada, gracias a la base de datos relacional, es determinante para acelerar la respuesta a los problemas que se encuentran sobre el terreno.

Al margen de estas informaciones, los coordinadores, en la medida en que cada uno dispone de un archivo *xls* donde va almacenando los datos correspondientes a la producción efectuada en su territorio, tienen también la posibilidad de realizar observaciones personalizadas a través de la elaboración de tablas dinámicas sobre aspectos más específicos de la producción de los territorios que gestionan y que les pueden interesar en un determinado momento (como por ejemplo la producción individual de los encuestadores, los niveles de sustitución, etc.).

2.1.2. Control del circuito de producción

El otro tipo de seguimiento estadístico sobre el trabajo de campo que se realiza en la *Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006* es el control del circuito de producción. Uno de los objetivos de este seguimiento es el de disponer de información constante acerca del estado de la producción en cada uno de los procesos que conforman el trabajo de campo, con la finalidad de utilizarla también en términos organizativos. Para ello se utiliza una tabla un tanto compleja, presentada tanto en números absolutos como en porcentajes, que integra información sobre diferentes aspectos del proceso de creación de la base de datos, como la entrega de muestra a los encuestadores, los cuestionarios registrados (realizados), la revisión, el control telefónico (supervisión), la codificación y la grabación de los datos (véase tabla 5).

Aparte de controlar individualmente la producción de cada uno de los procesos que forman parte del trabajo de campo, la utilidad de esta tabla recae más bien en la valiosa información que aporta en lo relativo a la organización. Uno de los aspectos más perseguidos en el diseño de este trabajo de campo es dotar a los equipos encargados de desarrollar las tareas de los diferentes procesos de la capacidad de trabajar con cierta autonomía y de manera simultánea, pese a estar trabajando sobre el mismo material. Como es evidente, los diferentes equipos de técnicos operan sobre los mismos cuestionarios, los cuales se tienen que realizar, revisar, codificar, grabar, etc. Por lo tanto, irremediablemente, los procesos permanecen encadenados, de manera que la realización de cuestionarios permite que el equipo que revisa cuestionarios pueda revisar, la revisión de cuestionarios permite que el equipo de codificadores pueda codificar, y así sucesivamente.

Tabla 5. Circuito de la producción (en porcentajes)*, 1 de mayo

	1	2	3	4	5 (3-4)
Ámbitos territoriales	Total cuestionarios a realizar	Entregados a los encuestadores	Realizados (registrados)	Revisados	Pendientes de revisar
Ámbito X	100,0%	81,5%	75,2%	60,5%	14,7%
Ámbito Y	100,0%	70,5%	59,8%	54,1%	5,7%
Ámbito Z	100,0%	76,5%	70,5%	60,7%	9,8%
Total	100,0%	76,1%	68,1%	58,0%	10,1%

(continúa)

	6	7 (4-6)	8	9 (6-8)
Ámbitos territoriales	Codificados	Pendientes de codificar de los revisados	Grabados	Pendientes de grabar de los codificados
Ámbito X	53,8%	6,7%	46,9%	6,9%
Ámbito Y	44,3%	9,8%	33,8%	10,5%
Ámbito Z	46,7%	14,0%	37,3%	9,4%
Total	48,6%	9,4%	39,7%	8,8%

* El cálculo de todos los porcentajes se realiza en base al total de cuestionarios a realizar en cada ámbito (columna 1)

En este sentido, cualquier incidencia o problema que limite la producción en uno de los procesos puede repercutir negativamente en el resto de fases que constituyen el trabajo de campo. De ahí la importancia de mantener, en cada uno de los procesos, márgenes de producción susceptibles de poder ser trabajados en la fase siguiente, generando así un equilibrio en el ritmo de producción entre las diferentes funciones que permita a cada equipo disponer de autonomía.

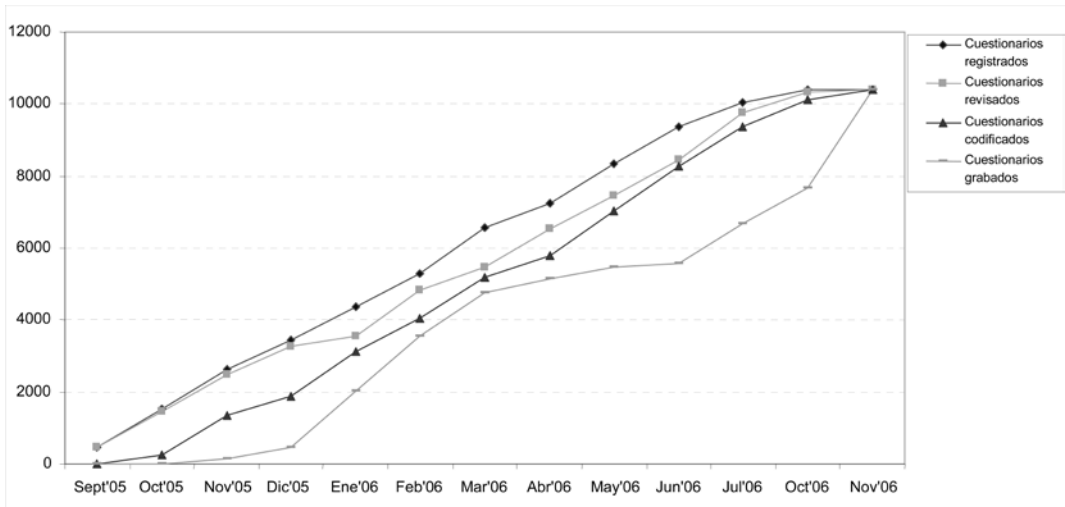
Si en la tabla 5 se atiende al total de las columnas *Pendientes de revisar de los realizados (registrados)*, *Pendientes de codificar de los revisados* y *Pendientes de grabar de los codificados*, se puede observar el porcentaje (en base al total de encuestas a realizar) correspondiente a estos márgenes de producción que permite controlar el flujo constante de cuestionarios a lo largo del circuito.

El porcentaje considerado óptimo alrededor del cual se debería intentar mantener los márgenes de producción durante el trabajo de campo de la *Encuesta* es del 10% (en base al total de encuestas a realizar). Así, si en algún proceso en concreto el porcentaje se aleja mucho de esta cifra se deben tomar las medidas necesarias en cuanto al ritmo de la producción (acelerándola o desacelerándola, dependiendo de la situación) para tratar de armonizar los márgenes en todos los procesos. En el momento del trabajo de campo hipotético al cual pertenecen los datos de la tabla 5 (1 de mayo), la situación de los márgenes de producción se encuentra bastante equilibrada, pero a lo largo del proceso no siempre es así, sobre todo durante el inicio y el cierre del trabajo de campo, cuando, por diferentes motivos, se encuentran los mayores problemas.

El gráfico 1 refleja la evolución de los márgenes de producción a lo largo de todo el trabajo de campo de la *Encuesta*. En él se perciben las considerables irregularidades que presentan dichos márgenes tanto en la fase inicial como durante los últimos meses, lo que no desmerece la utilidad de esta medida en la fase intermedia.

Estos problemas se deben sobre todo a dos cuestiones. En primer lugar, el control de los márgenes de producción requiere de un rodaje previo para comenzar a aplicarse, de ahí los desequilibrios que se observan en el comienzo del trabajo de campo. En segundo lugar, las irregularidades que se registran en los últimos meses, sobre todo en el proceso de grabación, se deben en gran parte a las constricciones de tiempo y recursos disponibles implícitas en los estudios por encuesta y en cualquier investigación en general. Como indica Cea D'Ancona, estos dos elementos,

Gráfico 1. Evolución de la producción en diferentes procesos del trabajo de campo de la ECVHP'06



Fuente: elaboración propia

tiempo y recursos disponibles, condicionan directamente, entre otras muchas cosas, determinados aspectos metodológicos. En este caso, intensificar la producción de determinados procesos en detrimento de la grabación de datos tiene mucho que ver con la falta de tiempo propia de los meses de cierre del trabajo de campo.

2.2. Control de la calidad de la muestra

El control de la muestra es una práctica ineludible en los estudios por encuesta, en la medida que implica el control de la no respuesta y constituye así el principal indicador de la calidad de los datos recogidos.

“Uno de los retos más difíciles de la investigación mediante encuesta es alcanzar la plena participación. Aunque se haya diseñado una muestra “perfecta”, que satisfaga los requisitos exigidos de tamaño y aleatoriedad en la elección de sus integrantes, si finalmente no se consigue que éstos respondan al cuestionario, todos los recursos a ella destinados se habrán malogrado. Los datos que se extraigan de la encuesta [...] no serán “representativos” de la población de interés si no se mantiene la composición de la muestra original e intervienen otros factores ajenos a la aleatoriedad. La no respuesta representa un error trascendental que deteriora la calidad de la encuesta por muestreo”

(Cea D’Ancona, 2004).

Por lo tanto, la medición y el control de la no respuesta, como fenómeno inevitable y como principal fuente de error de la técnica de encuesta, se vuelven tareas esenciales para conocer la naturaleza de los datos que finalmente se recogen.

En el trabajo de campo de la *Encuesta*, la incorporación en el sistema de seguimiento estadístico de este control y su uso a lo largo de todo el proceso de encuestación ha servido también, en cierta forma, para generar calidad en la información recogida. En este sentido, el seguimiento de la muestra se ha llevado a cabo a través de dos grupos de tablas denominados *tablas muestrales* y *tablas de calidad de la muestra*.

El primero de ellos, *tablas muestrales*, lo forman siete tablas que registran las encuestas realizadas de cada cuota de la muestra, establecidas éstas a través del conjunto de variables muestrales. Las variables muestrales son aquellas a partir de las cuales se construye la muestra y, por tanto, son las que delimitan los rasgos de la población objeto de estudio. En el caso de la *Encuesta*, la muestra se construye a partir de las siguientes variables:

- Variables territoriales: sección censal, distrito, municipio, comarca, ámbito del Plan Territorial de Cataluña y provincia³
- Sexo
- Edad
- Estrato social
- Origen geográfico

Cada una de las categorías que se derivan de estas variables constituye una cuota, que, a su vez, determina el número de individuos poseedores de un determinado atributo que deben formar parte de la muestra.

La finalidad básica del conjunto de tablas muestrales es ofrecer la referencia de producción de entrevistas en relación a cada cuota de la muestra. La posibilidad de tener acceso a este tipo de informa-

³ Las tablas tan sólo se han realizado en relación con las comarcas, los ámbitos del Plan Territorial de Cataluña y las provincias.

ción en cualquier momento, gracias a la base de datos relacional, permite también utilizarla para incidir en la estrategia de recogida de datos para favorecer la calidad de la información recogida. Esta estrategia es posible llevarla a cabo en la medida en que los coordinadores del trabajo de campo dosifican la muestra que entregan a los encuestadores para que la realicen y, por otro lado, porque los coordinadores dictan instrucciones a los encuestadores en función de las necesidades de la producción.

En este sentido, el principal objetivo durante el período de recogida de datos, en relación con las cuotas para incrementar en lo posible la calidad, es mantener un cierto equilibrio en la producción de todas ellas. El cumplimiento de este objetivo permitirá, entre otras cosas, que se controle a los colectivos más difíciles de entrevistar para que no se concentren al final del trabajo de campo, y, por tanto, evitar que un final precipitado de la recogida de datos deje una muestra excesivamente desequilibrada y mermada en términos de representatividad. Hay que decir de todas formas que lograr este objetivo resulta una tarea harto costosa y prácticamente irrealizable de manera plena, pero aún así, la existencia de una tendencia correctora en este sentido ya es suficiente para que tenga repercusiones en términos de calidad.

La tabla 6 es un ejemplo de este tipo de tablas, donde se presenta la producción de cuestionarios del caso hipotético considerado en relación con la variable edad de los individuos encuestados. En ella se puede observar un cierto desequilibrio en la producción de entrevistas a individuos pertenecientes a los grupos de edad mayor y menor de la muestra, y, por tanto, se debería considerar la necesidad de priorizar la realización de entrevistas a los individuos pertenecientes a estos grupos.

Tabla 6. Control cuotas variable edad, 1 de mayo

Grupo de edad	Cuestionarios a realizar	Cuestionarios realizados	% realizados
20-29 años	966	520	53,8
30-39 años	983	711	72,3
40-49 años	864	619	71,6
50-59 años	813	633	77,9
60-69 años	719	570	79,3
70-79 años	655	352	53,7
Total	5.000	3.405	68,1

El otro grupo de tablas relacionadas con el control de la muestra lo constituyen el conjunto de *tablas de calidad muestral*, las cuales recogen, básicamente, los datos que hacen referencia a las suplencias de la muestra y los motivos que las provocan.

El diseño de la muestra de la *Encuesta* establece una serie de titulares, todos ellos individuos de unas determinadas características, a los cuales se les debe pasar el cuestionario. Pero además, la muestra cuenta con un grupo de suplentes para cada titular, normalmente siete, de las mismas características que éste, y que, lógicamente, tan sólo son entrevistados en aquellos casos en que no se puede acceder al titular por el motivo que sea. En este sentido, teniendo en cuenta las dimensiones y la complejidad del trabajo de campo de este proyecto, y haciendo de nuevo referencia a la anterior cita de Cea d'Ancona, resulta inevitable que se produzcan suplencias en la muestra titular y, por lo tanto, que el detrimento de aleatoriedad perjudique la calidad de la muestra. Precisamente por esa razón, es imprescindible controlar la cantidad y los motivos de suplencia por los cuales se producen estas sustituciones, para así conocer el grado de calidad de los datos, teniendo en cuenta el sesgo que se produce con respecto a la muestra original y las razones que han originado las suplencias.

Para ello se elabora una tabla de distribución de la muestra realizada en función de la persona a la que se ha realizado el cuestionario (titular o suplente) y una tabla con los motivos de suplencia. La

primera informa de la proporción de titulares y suplentes, mientras que la segunda refleja la distribución de los contactos que se han producido sobre la muestra por parte de los encuestadores atendiendo a su resultado. Es decir, cuando un encuestador se dirige a un individuo de la muestra con la intención de entrevistarle se considera un contacto, y éste puede acabar con una encuesta realizada o con un contacto fallido. En el segundo caso, se produciría una suplencia y se registraría el motivo que la ha provocado.

Así, cuanto más bajo sea el porcentaje de encuestas realizadas al titular, mayor será la proporción de contactos fallidos, lo que implicará un mayor sesgo de la muestra original, ya que, aunque se trate de individuos de características similares a los titulares, disminuye la aleatoriedad con la que los individuos entrevistados son seleccionados.

La tabla 7 es un ejemplo de distribución entre titulares y suplentes que recoge los datos de la muestra final realizada en el caso hipotético analizado, donde se observa que gran parte de la muestra, más del 80%, se ha realizado a titulares, suplentes 1 o suplentes 2. Es más, el 45,8% de los entrevistados son titulares.

Tabla 7. Distribución entre titulares y suplentes de la muestra final

Titular/Suplente	Frecuencia	Porcentaje	Titular/Suplente	Frecuencia	Porcentaje
Titular	2.290	45,8%	Suplente 7	45	0,9%
Suplente 1	1.170	23,4%	Suplente 8	20	0,4%
Suplente 2	665	13,3%	Suplente 9	10	0,2%
Suplente 3	425	8,5%	Suplente 10	5	0,1%
Suplente 4	185	3,7%	Suplente 11	5	0,1%
Suplente 5	110	2,2%	Total	5.000	100,0%
Suplente 6	70	1,4%			

El motivo de suplencia explica la razón por la que no se realiza la entrevista al titular o a los sucesivos suplentes, y aporta información sobre los motivos del sesgo que se está produciendo en la muestra titular. Por ello, interesa desagregar las diferentes causas que provocan que un contacto no pueda acabar en una entrevista realizada y, por lo tanto, en un cuestionario listo para ser explotado.

La tabla 8 recoge este tipo de información, resumiendo la distribución de contactos en números absolutos y en porcentajes, además de la distribución entre los diversos motivos de suplencia de los contactos fallidos.

En la *Encuesta*, esta tabla también se elabora atendiendo a la desagregación territorial habitual, para mantener el control de este sesgo también a nivel territorial.

Otro de los aspectos importantes que se pueden analizar a través de los motivos de suplencia es el nivel de aceptación que tiene la encuesta entre la población contactada, ya que se cuantifica el número de suplencias provocadas por la negación a participar en la encuesta. En el caso concreto de la *Encuesta de condiciones de vida y hábitos de la población de Catalunya, 2006*, la *negativa* aparece como el tercer motivo de suplencia más importante, por debajo del *cambio de domicilio e ilocalizable*, de lo que se deduce que dicho trabajo de campo ha tenido una óptima aceptación por parte de la población.

Este análisis de los motivos que provocan las suplencias es la aportación más clara en términos de calidad que se desprende del uso de este tipo de tablas. Lo lógico sería detectar los problemas que han originado la no repuesta y que sean susceptibles de ser solucionados en futuros proyectos. Se podría decir quizá que es la más peculiar de las aportaciones a la calidad que se ha presentado en

Tabla 8. Motivos de suplencia

Motivo de suplencia	Contactos en la muestra (nº absolutos)	Contactos en la muestra (%)	Contactos fallidos (%)
Encuesta realizada	5.000	44,8%	-
Negativa	1.840	16,5%	29,9%
Cambio de domicilio	2.175	19,5%	35,3%
llocalizable	1.316	11,8%	21,4%
Error censal (no se encuentra la dirección, el edificio ya no existe...)	234	2,1%	3,8%
Enfermedad	178	1,6%	2,9%
Defunción	145	1,3%	2,4%
Es su segunda residencia	78	0,7%	1,3%
No procede (hotel, residencia, colectividad)	33	0,3%	0,5%
Mismo domicilio que otro encuestado	11	0,1%	0,2%
Otros motivos	145	1,3%	2,4%
Total	11.155	100,0%	100,0%

este artículo; ya que, en general, el efecto que produce tiene lugar *a posteriori* de la campaña a la cual pertenecen los datos. Ahora bien, dependiendo del tipo de motivo de suplencia de que se trate y de la capacidad de respuesta de la que se disponga, se podría plantear también una intervención en el mismo trabajo de campo.

Por otro lado, dadas las variables disponibles, sería posible también realizar análisis más sofisticados de la no respuesta, por ejemplo cruzando determinados perfiles sociales de los motivos de suplencia, para conocer así si hay algún motivo de suplencia asociado a un colectivo social determinado. Uno de los casos más claros para ilustrar esta posibilidad en la edición de 2006 de la *Encuesta* es el del colectivo inmigrante extracomunitario, que precisamente ha presentado uno de los índices de no respuesta más elevados. El análisis de los motivos de suplencia en relación con este colectivo desvela que registra una negativa bajísima a colaborar con la encuesta en comparación con la población comunitaria (nacidos en la UE15, incluida la población española) y su no respuesta viene explicada principalmente por el elevado porcentaje de cambio de domicilio que presenta. Este resultado permite determinar que, en la medida en que se controle el factor del domicilio en individuos que pertenezcan a este colectivo, se podrá incrementar notablemente su participación y, por tanto, disminuir su no respuesta.

Esta función no se puede incluir dentro del seguimiento estadístico, pero sí se puede considerar como una aportación más a la calidad generada a partir de los datos que se recogen durante el trabajo de campo, gestionados por la base de datos relacional, y desde los que se realizan las estadísticas de seguimiento.

3. Conclusiones

Recapitulando, se puede decir que la gestión y el uso intensivo de la información que emana de un trabajo de campo, ya sea en forma de seguimiento estadístico o como análisis posterior a la recogida

da de datos, incrementan las posibilidades de mejorar la calidad de los resultados de la técnica de encuesta. En este sentido, es importante resaltar también la agilidad y eficiencia que aporta la base de datos relacional como instrumento de gestión de esta información.

A lo largo de este artículo se han expuesto las prácticas más relevantes destinadas a incrementar la calidad de los datos. Éstas se podrían resumir en:

- a) El control de la producción a nivel territorial a través del seguimiento estadístico favorece la consecución de una producción territorialmente equilibrada. De esta manera se reduce el riesgo de perder representatividad en los datos en casos extremos como, por ejemplo, el cierre precipitado del trabajo de campo.
- b) En trabajos de campo presenciales que abordan un territorio de estudio amplio, los cuales suelen ser llevados a cabo por un equipo técnico complejo, el seguimiento estadístico aporta información muy valiosa sobre la evolución de los diferentes procesos que forman parte de la campaña. Esto permite mantener un control exhaustivo y constante de la producción de los diferentes procesos, ofreciendo la posibilidad de optimizar la organización interna del equipo y dar respuesta inmediata a los problemas que puedan surgir en la construcción de la base de datos.
- c) El control de cuotas, considerado también como parte de las estadísticas de seguimiento, permite introducir tendencias correctoras que disminuyan los sesgos que se originan en el proceso de consecución de la muestra. El objetivo consistiría en intentar mantener un cierto equilibrio en las cuotas muestrales durante todo el proceso de recogida de datos, de manera que se reduzca el riesgo de pérdida de representatividad de los resultados.
- d) Ante los casos inevitables de no respuesta, el control de cuotas se erige como un elemento clave para conocer la naturaleza y el nivel de calidad de los datos obtenidos finalmente. En este sentido, con los datos recogidos durante todo el proceso de campo es posible realizar un informe de campo minucioso que analice el potencial y las limitaciones de la información recogida. Un informe de este tipo facilitaría a cualquier investigador, durante la fase de análisis, la información suficiente para conocer en profundidad los datos, aspecto que se considera imprescindible para una correcta interpretación de los resultados.
- e) Un análisis posterior más sofisticado de los datos de campo, por ejemplo de los motivos de supencias, puede dar pistas sobre cómo solucionar problemas relacionados con la no respuesta en futuros proyectos.

Aparte de las prácticas expuestas en este artículo, las posibilidades que ofrece el uso de las estadísticas de seguimiento pueden ir más allá. Todo depende de la ambición con que se plantee su uso y, sobre todo, de los recursos que se dispongan para llevarlos a cabo.

Bibliografía

ALVIRA, F. «La encuesta: una perspectiva general metodológica». *Cuadernos Metodológicos* [Madrid: Centro de Investigaciones Sociológicas], (2004).

CEA D'ANCONA, M. A. *Métodos de encuesta. Teoría y práctica, errores y mejora*. Madrid: Ed. Síntesis, 2004.

CEA D'ANCONA, M. A. «La senda tortuosa de la "calidad" de la encuesta». *Revista Española de Investigaciones Sociológicas*, n.º 111 (2005), p. 75-104.

CODD, E. F. *The Relational Model for Database*. Oxford: Addison-Wesley Publishing Company, 1990.

05 CREACIÓN Y DEPURACIÓN DE LA BASE DE DATOS

José Luis Flores (Instituto de Estudios Regionales y Metropolitanos de Barcelona)

Introducción

1. Fases en la creación de la base de datos provisional

- 1.1. Codificación de los cuestionarios
- 1.2. Grabación de los datos
- 1.3. Base de datos provisional

2. Depuración de la base de datos

- 2.1. Primera depuración
- 2.2. Control continuo de la base de datos

3. Base de datos definitiva

Introducción

"Por muy buenas que sean las técnicas de análisis utilizadas, si los datos, fuente de información, son incorrectos, los resultados serán incorrectos"

GIGO: Garbage In - Garbage Out

Con dicho acrónimo (GIGO, que en castellano se podría resumir: *Si se guarda basura, sale basura*), algunos programadores de datos intentan concienciar a cualquier investigador en ciencias sociales que, si se introduce información incorrecta en un programa, el resultado que se obtiene será también erróneo. El ordenador no podrá mejorar la "esencia" de los datos o de la información con que se trabaja; a lo sumo, podrá mantenerla. No son pocos los casos donde se recurre, por ejemplo, a datos de muy mala calidad y se pretende que la computadora, después de procesar dichos datos, brinde información de altísima calidad.

En este sentido, y para garantizar la fiabilidad de la información obtenida a través del trabajo de campo, es necesario llevar a término un riguroso y constante trabajo de control de las diferentes etapas del proceso de recogida y de tratamiento de los datos. Y es que no podemos olvidar que los datos obtenidos son la fuente de información de todos los análisis que se realicen posteriormente, y que, por ello, el control de calidad es un elemento omnipresente en todo el proceso de creación de la base de datos de la *Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006*.

1. Fases en la creación de la base de datos provisional

Una vez los cuestionarios son revisados por los jefes de equipo —a través de un repaso visual pormenorizado de todos ellos que permite controlar la calidad de la información recogida por los encuestadores, y que en algunas ocasiones comporta un control telefónico—, y registrados en las estadísticas y tablas de seguimiento —lo que permite controlar la buena marcha y la calidad del trabajo de campo, para que no se produzcan desviaciones importantes de la muestra inicial—, continúan su circuito, pasando a la fase de codificación.

1.1. Codificación de los cuestionarios

En esta fase del proceso de realización de la *Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006* consiste en codificar los cuestionarios, uno por uno, para facilitar posteriormente la grabación de la información recogida. Para llevar a cabo dicha fase existe un equipo formado por ocho codificadores, todos ellos con un amplio conocimiento del cuestionario, y un jefe de equipo que se dedica a supervisar todo el proceso, así como a resolver las dudas que puedan ir apareciendo.

El proceso de codificación no se realiza de una manera automática, sino que los codificadores revisan el cuestionario a medida que lo van codificando, comprobando la coherencia de la información recogida y reforzando así la revisión sistemática realizada por los jefes de equipo de los diferentes encuestadores. De esta manera se ralentiza algo el proceso, si bien dicho tiempo queda compensado por la mayor calidad de los datos obtenidos.

Cada cuestionario es codificado por una sola persona, con la única excepción de las preguntas 35 (referente a la ocupación concreta que realiza o realizaba el entrevistado y otros miembros del hogar) y 38 (que hace referencia a la actividad de la empresa o institución donde trabaja o trabajaba el entrevistado). Estas dos preguntas de todos los cuestionarios son codificadas, ambas con tres dígitos, por una misma persona, que utiliza la Clasificación Catalana de Ocupaciones 1994 y la Clasificación Nacional de Actividades Económicas 1993, consultando, cuando es necesario, la página Web del INE (Clasificaciones. Ayuda a la codificación).

Cada codificador apunta diariamente los números de los cuestionarios codificados en una hoja personal de seguimiento que recoge semanalmente el jefe de codificadores. Los cuestionarios totalmente codificados, incluidas las preguntas 35 y 38, se van marcando (con diferentes tramas para cada semana) en una hoja de control general, lo que permite saber en todo momento si un cuestionario está o no codificado y la semana en la que se codificó. Una vez controlados, los cuestionarios pasarán a grabación, pero, antes de seguir con el circuito, detengámonos un momento en los criterios utilizados en esta fase.

En la mayoría de los casos, para llevar a término su trabajo, los codificadores transfieren directamente el código de la respuesta marcada por el encuestador en el cuestionario, como sería el caso de las preguntas 4, 24, 94 o 168. Ahora bien, en el caso de posibilidades de respuesta no consideradas en el cuestionario —como ocurre en las preguntas 20, 84, 138 o 152, entre otras—, las personas que codifican consultan el *Manual de codificación*, con sus anexos correspondientes, donde encontrarán, para cada una de las variables, un abanico de respuestas posibles, cada una con su código correspondiente, más amplio que el del cuestionario.

En el caso de las escasas preguntas totalmente abiertas del cuestionario —como la 155 o la 161—, las posibilidades de respuesta que se han incluido en el *Manual de codificación* lo han sido en función de los resultados obtenidos en anteriores ediciones del presente estudio o en otras encuestas realizadas por el IERMB, así como a partir de la consulta a diferentes investigadores expertos en diversas materias. De todas formas, en alguna pregunta se han ido añadiendo posibilidades de

respuesta a medida que ha ido transcurriendo el trabajo de campo, para lo cual se han tenido que abrir nuevos códigos en el *Manual de codificación*, para incorporar, por ejemplo, nuevas tendencias o aspectos no tenidos en cuenta a la hora de confeccionar dicho *Manual*.

En esta edición de la *Encuesta*, se han dejado por primera vez en blanco las posiciones de las preguntas que no se deben realizar, es decir, los “no procede” (hasta esta edición se codificaban como 0). Esto ha aligerado el proceso, puesto que muchas preguntas no se han de efectuar, como por ejemplo la batería de preguntas a las personas ocupadas, desempleadas o inactivas, en función de la situación laboral actual del entrevistado.

Figura 1. Diferentes tipos de preguntas a la hora de codificar

Pregunta sin posibilidades de respuesta añadidas en el manual de codificación

Pregunta del cuestionario	Posibles respuestas en el cuestionario	Manual de codificación
24. ¿Cuál es su nivel de conocimiento del catalán?	1. Lo habla y lo escribe 2. Lo habla, pero no lo escribe 3. Lo entiende, pero no lo habla 4. No lo entiende	1. Lo habla y lo escribe 2. Lo habla, pero no lo escribe 3. Lo entiende, pero no lo habla 4. No lo entiende 9. NS/NC

Pregunta con posibilidades de respuesta añadidas en el manual de codificación

Pregunta del cuestionario	Posibles respuestas en el cuestionario	Manual de codificación
84. ¿Cuándo cree que podrá volver/comenzar a trabajar?	1. Cuando acabe los estudios 2. Cuando mis hijos crezcan 3. Cuando encuentre un trabajo interesante 4. Cuando tenga necesidad económica 5. Otra posibilidad: _____ 6. No creo que pueda trabajar	1. Cuando acabe los estudios 2. Cuando mis hijos crezcan 3. Cuando encuentre un trabajo interesante 4. Cuando tenga necesidad económica 5. Otra posibilidad: _____ 6. No creo que pueda trabajar 7. Cuando deje de cuidar a los nietos 8. Cuando tenga los permisos en regla 9. NS/NC


Pregunta totalmente abierta sin posibilidades de respuesta en el cuestionario

Pregunta del cuestionario	Posibles respuestas en el cuestionario	Manual de codificación
155. Cuando está en casa, ¿a qué dedica el tiempo libre? (5 respuestas como máximo)	1. _____ 2. _____ 3. _____ 4. _____ 5. _____	00. No tiene tiempo libre 01. Leer 02. Ver la televisión, el vídeo o DVD 03. Escuchar música 04. Escuchar la radio etc..... 26. Estar con los nietos 27. Hacer pasatiempos 99. NS/NC

Fuente: elaboración propia

La primera parte del cuestionario (de la portada a la pregunta 40) se codifica en una plantilla donde se van apuntando los números correspondientes (véase figura 2). Esta plantilla está realizada para simplificar y aligerar el proceso y para facilitar el momento de la grabación, puesto que hasta la pregunta 40 sería difícil codificar en el propio cuestionario debido a la estructura del mismo. A partir de dicha pregunta, y hasta el final, la codificación se realiza en el propio cuestionario, en los espacios reservados para ello que quedan a la derecha de las diferentes preguntas (véase figura 3).

Figura 2. Plantilla de codificación



**Enquesta sobre condicions de vida i hàbits
de la població de Catalunya 2006**

Codificador: _____
 Data: _____
 Control: _____

IDENTIFICACIÓ										ESTRUCTURA FAMILIAR																																							
Num. qüest.					Territori					Producció					Nom					Parentiu																													
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Any naixement										Sexe																																							
51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
Estat Civil										Convivència					Fills					Any naixement fills																													
91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140
Convivència fills										Desig					PROCEDENCIA GEOGRAFICA																																		
Provincia naixement										LLENGUA					NIVELL D'ESTUDIS																																		
Municipi										Arribada a Catalunya/al municipi					Origen progenitors					Llengua pròpia					Us llengües					Nivell d'estudis																			
141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190
Títol universitat										Estudis registrats en curs					Tipus de centre					Localització centre																													
191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240
SITUACIÓ LABORAL										Origen ingressos					Ocupació					Situació feina					Responsabilitat					Activitat empr.					Tipus														
Situació laboral										Categoría professional					Situació feina					Responsabilitat					Activitat empr.					Tipus																			
241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290
Categoría professional										Situació feina					Responsabilitat					Activitat empr.					Tipus																								
291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340

Figura 3. Fragmento de cuestionario con espacios para codificación

Per a totes les persones entrevistades		
122. En els darrers cinc anys, han hagut de reduir despeses a la llar?		
1. Sí		(631)
6. No	⇒ passi a la pregunta 124	
Només per als qui han hagut de reduir despeses a la llar en els darrers cinc anys		
123. Quin tipus de despesa han reduït principalment?		
01. Alimentació		(632,633)
02. Begudes i tabac		
03. Vestit i calçat		
04. Mobles i objectes per a la llar		
05. Béns i serveis mèdics o farmacèutics		
06. Oci i lleure		
07. Una altra possibilitat. <i>Especifiqui-la:</i> _____		
I. Habitatge		
Seguidament li faré una sèrie de preguntes relatives al seu habitatge i al seu barri		
124. Quin és el règim de tinença de l'habitatge en viu habitualment?		
01. En propietat, per herència o per donació familiar	⇒ passi a la pregunta 126	(634,635)
02. En propietat, totalment pagada	⇒ passi a la pregunta 126	
03. En propietat, amb pagaments ajornats (hipoteca o similar)		
04. Lloguer amb contracte indefinit		
05. Lloguer a terminis		
06. Cedida per familiars o amics	⇒ passi a la pregunta 126	
07. Cedida per una empresa en què treballa algun membre de la família	⇒ passi a la pregunta 126	
08. Subarrendament (relogats)		
09. En precari (sense contracte, ocupació, etc.)		
10. Una altra possibilitat. <i>Especifiqui-la:</i> _____		
125. Quina quantitat van gastar a la seva llar en el darrer MES pel lloguer o la hipoteca d'aquest habitatge?		
TARGETA 9		
01. Fins a 20 euros		(636,637)
02. De 21 a 50 euros		
03. De 51 a 100 euros		
04. De 101 a 200 euros		
05. De 201 a 300 euros		
06. De 301 a 600 euros		
07. De 601 a 900 euros		
08. De 901 a 1.200 euros		
09. De 1.201 a 1.500 euros		
10. De 1.501 a 1.800 euros		
11. Més de 1.800 euros		
126. En total, quants metres quadrats té aquest habitatge?		
01. Fins a 50 m ²		(638,639)
02. De 51 a 60 m ²		
03. De 61 a 70 m ²		
04. De 71 a 80 m ²		
05. De 81 a 90 m ²		
06. De 91 a 100 m ²		
07. De 101 a 110 m ²		
08. De 111 a 120 m ²		
09. De 121 a 130 m ²		
10. De 131 a 150 m ²		
11. De 151 a 200 m ²		
12. De 201 a 300 m ²		
13. Més de 300 m ²		

la estructura modelo de la base de datos, que contiene información esencial sobre las variables del cuestionario: las columnas donde se hallan en el fichero, las etiquetas que nombran a las variables y a sus atributos (categorías de respuesta o valores numéricos), el formato de cada variable, la especificación de la no respuesta (no sabe/no contesta) y la de los valores perdidos (véase figura 5). La aplicación de esta estructura modelo a la base grabada en formato ASCII permite obtener un único fichero en formato de tabla con los valores de cada variable (véase figura 6).

La presente fase del trabajo se realiza a través de un sistema de doble grabación, que consiste en introducir los datos de todos los cuestionarios dos veces, para garantizar que no se produzcan errores de tecleo por parte del equipo de grabación.

Figura 6. Base de datos en formato tabla

	s1	s2	v001	v002a	v003a	v004a	v005a	v006	v007	v008	v009	v010	v011a	v011b	v011c	v011d
1	5	8019	3	3	1961	6	1	1		1985	0	2	1986	1989		
2	5	8019	1	1	1922	6	3	6				5	9999	9999	9999	9999
3	5	8019	2	1	1975	1	5	1		2004	0	0				
4	5	8019	2	4	1971	1	1	6	6			0				
5	5	8019	4	2	1965	6	2	1		1993	1994	2	1998	2000		
6	5	8019	2	1	1937	1	2	1		1959	1959	2	1959	1964		
7	5	8019	3	2	1948	1	2	1		9999	9999	4	9999	1977	1979	1980
8	5	8019	4	3	1965	6	4	1		1991		2	1994	1994		
9	5	8019	2	3	1981	6	1	1		2004		0				

1.3. Base de datos provisional

Un aspecto crucial en la realización de una base de datos es la comprobación de que las etiquetas identificadoras de las variables y de las respuestas sean correctas, tanto con relación a posibles faltas de ortografía como, principalmente, a que transmitan la información que realmente han de transmitir, lo que facilitará la interpretación correcta de los datos, sin posibilidad de errores. Así, por ejemplo, la etiqueta de una variable ha de incluir la selección de la muestra a la que afecta dicha pregunta, el período de tiempo al que hace referencia, la medida utilizada a la hora de recoger la información, ... Veamos, a modo de ejemplo, las etiquetas que deberían acompañar a las preguntas 43 y 68 del cuestionario.

La variable 43 pregunta, a las personas ocupadas que llevan menos de cinco años en la empresa actual, si han estado paradas en los últimos cinco años, así como, en caso afirmativo, las veces que lo han estado. Por su parte, la variable 68 pregunta, a las personas paradas que han trabajado anteriormente, por las veces que han estado en dicha situación de paro durante los últimos cinco años, incluyendo la actual. Así, el siguiente etiquetaje sería totalmente incorrecto, ya que las dos variables aparecerían igual:

Pregunta 43. Veces en paro
 Pregunta 68. Veces en paro

Las siguientes etiquetas continuarían siendo incorrectas, ya que la pregunta no se realiza a todos los ocupados o parados y, además, se refiere a un período de tiempo concreto:

Pregunta 43. Veces en paro (ocupados)
 Pregunta 68. Veces en paro (parados)

Si se añade el período de tiempo, las etiquetas van mejorando, si bien todavía no reflejan correctamente lo que deberían:

Pregunta 43. Veces en paro durante últimos cinco años (ocupados)

Pregunta 68. Veces en paro durante últimos cinco años (parados)

Finalmente, si acotamos a qué ocupados y parados se realizan las preguntas, obtendremos las siguientes etiquetas, que, si bien pueden pecar de largas, nos informan correctamente de cuáles son los datos que nos muestran dichas variables:

Pregunta 43. Veces en paro durante últimos cinco años (ocupados con menos de 5 años en su empresa actual)

Pregunta 68. Veces en paro durante últimos cinco años (parados que han trabajado anteriormente)

Además, siempre se puede abreviar el etiquetaje, si bien se ha de ir siempre con cuidado para no cambiar el significado.

Pregunta 43. Veces en paro últimos 5 años (ocupados < 5 años en empresa actual)

Pregunta 68. Veces en paro últimos 5 años (parados que han trabajado antes)

Otras veces no es tan complicado, ya que simplemente se trata de añadir una/s palabra/s o alguna explicación entre paréntesis. Éste es el caso de la variable 54, que pregunta sobre el tiempo que se invierte habitualmente en el desplazamiento al trabajo (en minutos), contabilizando sólo el viaje de ida. Así, la siguiente etiqueta podría llevar a error, ya que alguna persona podría interpretar que se trata del total de tiempo invertido en todos los viajes laborales que se realizan al día, tanto de ida como de vuelta:

Pregunta 54. Tiempo en desplazamiento al trabajo

Situación que mejora claramente de la siguiente forma:

Pregunta 54. Tiempo invertido habitualmente (en minutos) en el viaje de la vivienda al trabajo

Otras veces son las etiquetas de las respuestas las que pueden llevar a error, en muchos casos por querer simplificarlas tanto que al final no se interpreta correctamente lo que significan en realidad.

A medida que se van grabando los cuestionarios se va realizando una lectura de los resultados parciales obtenidos, que, a su vez, pueden desglosarse territorialmente para Barcelona y para cada uno de los ámbitos territoriales que después se analizarán, así como por encuestador/a. De esta manera se puede controlar, por ejemplo, si existen variables en las cuales se va acumulando un porcentaje de NS/NC superior al X%, o si se dan diferencias importantes en relación con este tipo de respuesta entre territorios o entre encuestadores, lo cual nos permite ver si alguno de ellos tiene unos porcentajes de NS/NC muy altos o por encima de la media. Ello permitirá que los jefes de equipo de los diferentes encuestadores puedan comunicarles la existencia de dicho problema si lo consideran necesario.

2. Depuración de la base de datos

Una vez llegado a este punto, es necesario llevar a cabo un control exhaustivo del trabajo realizado hasta entonces. Se trata de depurar la base de datos obtenida, para lo cual se procede a controlar los códigos fuera de rango (códigos incorrectos o inexistentes en las posibilidades de respuesta de cada una de las preguntas del cuestionario), así como toda una serie de relaciones entre las respuestas de diferentes variables, lo cual se realiza a partir del denominado *Programa de inconsis-*

tencias. Se trata, pues, de un paso más en el proceso que intenta detectar y solucionar el máximo de errores posibles para minimizar su impacto, y así obtener una buena base de datos que permita realizar análisis y estudios con total fiabilidad.

2.1. Primera depuración

Este primer control se realiza cada vez que se graban alrededor de 500 nuevos cuestionarios (que conforman varias microbases de datos), de tal manera que se lleva a cabo paralelamente al trabajo de los encuestadores, lo cual permitirá que, una vez finalizado el trabajo de campo, se puedan tener los datos en un menor espacio de tiempo. Además, esta metodología permite mejorar el *Programa de inconsistencias* a medida que se va poniendo en práctica durante todo el tiempo que dura el proceso. Al final del trabajo de campo se pasará nuevamente el programa al conjunto de las 10.398 encuestas realizadas.

En concreto, y como ya se ha mencionado, se trata de controlar toda una serie de códigos, de relaciones y correspondencias entre variables que, de entrada, se deberían cumplir. Para ello, el equipo de investigadores de la *Encuesta* realizó, a partir de la estructura del cuestionario y de la experiencia de ediciones anteriores, un *Manual de relaciones entre variables*, con toda una serie de relaciones entre variables/posiciones del cuestionario que serán posteriormente traducidas a lenguaje informático por el equipo de estadística del IERMB.

El mencionado *Manual* ocupa casi 20 hojas, si bien a continuación se muestran algunos fragmentos¹ que afectan a diferentes partes del cuestionario y que pueden ayudarnos a entender mejor el proceso/circuito seguido en esta fase del trabajo.

Fragmentos del *Manual de relaciones entre variables*:

A.-2 "Pregunta 3": Año de nacimiento

- a) El número de respuestas de la v003 (de la p50 a la p86) ha de coincidir con la v001 (p28), controlar que no haya algún "blanco/missing/no procede" mal situado
- b) si p30=1 -> p50=p54
- c) normalmente, la diferencia de edad entre CF (cabeza de familia) y su pareja no puede ser mayor de 20 años
- d) p50 siempre <=1990 (tomando como año de realización del trabajo de campo el 2006).
- e) normalmente p54 <=1986

D.-1 "Pregunta 26": Nivel de estudios finalizados

- a) si p239=14, 15, 16 -> se ha de contestar p247
si p247 está contestada -> p239=14, 15, 16
- b) si p110=1 -> p241 se ha de contestar
si p241 está contestada -> p110=1
- c) p245 se ha de contestar si el CF tiene pareja (par_cf=1)
- d) si p30=1 -> p239=p243 y p241=p245
- e) si p30=2,3 -> p239=p245 y p241=p243
- h) si p237 o p238=1 -> p239≠1
si p239=1 -> p237 y p238≠1

¹ La letra inicial corresponde al apartado del cuestionario al que hace referencia la relación, después aparece el número de la/s principal/es pregunta/s involucrada/s, así como el tema a que hace referencia. Los diferentes apartados —a), b), c), d),...— son los que indican las relaciones concretas existentes entre las variables (v) o posiciones (p) del cuestionario involucradas, que ayudarán posteriormente a construir las inconsistencias.

E1.-1 "Preguntas 41 a 63": Personas ocupadas

- a) si $p338 \neq 1 \rightarrow$ no se ha de contestar el bloque E1
si $p338 = 1 \rightarrow$ se ha de contestar el bloque E1
si se contesta el bloque E1 $\rightarrow p338 = 1$
- b) si $p410 \geq 60 \rightarrow$ no se han de contestar p413 y p415
si se contestan p413 y p415 $\rightarrow p410 < 60$
- c) si $p410 < 60 \rightarrow$ se han de contestar p413, p415; $p415 \neq 0$
si no se contestan p413 y p415 $\rightarrow p410 \geq 60$
- d) siempre $p418/p417 < 15$ (no se puede trabajar una media de 15 o más horas al día)
- e) si $p420 = 2 \rightarrow$ normalmente $p418 < 30$
si $p420 = 1 \rightarrow$ normalmente $p418 \geq 30$
- f) si $p420 = 2$ se ha de contestar p421
si $p420 = 1$ no se ha de contestar p421
si p421 está contestada $\rightarrow p420 = 2$
- g) si $p380 = 1$ se ha de contestar p424
si $p380 \neq 1$ no se ha de contestar p424
si p424 se contesta $\rightarrow p380 = 1$
- h) si $p427 \neq 1$ se han de contestar p432, p435, p436
si $p427 = 1$ no se han de contestar p432, p435, p436
- i) si $p443 = 1$ se ha de contestar p444
si $p443 \neq 1$ no se ha de contestar p444

J.-1 "Preguntas 136 a 142": Llegada a la vivienda, motivo del cambio y previsiones de cambio

- a) si $p676 \neq 0 \rightarrow p676 \geq p50$
- b) si $p676 = 0 \leftrightarrow$ no se han de contestar p680, p685, p687
- c) si $p676 = 0 \rightarrow p213 = 0$
- d) siempre $p676 \geq p213$
- e) si $p676 \neq p213 \rightarrow p680 = 2,4$
- f) si $p689 = 1 \rightarrow$ se han de contestar p690 y 692
si p690 y 692 se contestan $\rightarrow p689 = 1$

A partir de dicho *Manual*, los estadísticos proceden a traducir las diferentes relaciones a lenguaje informático y las convierten en una sintaxis inteligible para el programa SPSS (lo que denominamos *Programa de inconsistencias*), y que en muchos casos será justamente la inversa. Es decir, si la relación E1.-1f) del *Manual de relaciones entre variables* indica que si la posición 420 es igual a 2 se ha de contestar la posición 421, a la hora de informatizar dicha relación para el *Programa* se tendrá que reconvertir para que el ordenador extraiga los cuestionarios donde no se cumple dicha condición, es decir, los cuestionarios donde $p420 = 2$ y $p421 = \text{missing}$. A continuación se muestra, junto con algunos ejemplos sobre el control de los fuera de rango, cómo se reconvierten algunas otras de las relaciones del *Manual* para poder ser leídas electrónicamente.

Ejemplos de sintaxis para controlar los fuera de rango de diferentes preguntas del cuestionario:

Pregunta 2 (v002)

```
do repeat v002= v002a v002b v002c v002d v002e v002f v002g v002h v002i v002j/  
control= r30 r32 r34 r36 r38 r40 r42 r44 r46 r48.  
if not(range(v002a,1,23,99,99)) control=1.  
end repeat.
```

Pregunta 3 (v003)

```
if not(range(v003a,1900,1990,9999,9999)) r50=1.  
if not(range(v003b,1900,1986,9999,9999)) r54=1.
```



```
do repeat v003= v003c v003d v003e v003f v003g v003h v003i v003j/
control= r58 r62 r66 r70 r74 r78 r82 r86.
if not(range(v003,1900,2006,9999,9999)) control=1.
end repeat.
```

Pregunta 4 (v004)

```
do repeat v004= v004a v004b v004c v004d v004e v004f v004g v004h v004i v004j/
control= r90 r91 r92 r93 r94 r95 r96 r97 r98 r99.
if not(any(v004,1,6,9)) control=1.
end repeat.
```

Ejemplos de sintaxis para controlar las incoherencias del *Manual de relaciones*:

```
A.-2:b
do if (v002a=1 ).
if v002a≠v002b r002=concat('b',r002).
if v003a≠v003b r003=concat('b',r003).
if v004a≠v004b r004=concat('b',r004).
if v005a≠v005b r005=concat('b',r005).
end if.
```

```
A.-2:d
if v003a>1990 r003=concat('d',r003).
```

```
D.-1:a
if (any(v026a,14,15,16) and (sysmis(v027) or not(range(v027,1,28,100,143,999,999)))) or
(not(any(v026a,14,15,16)) and not(sysmis(v027))) r026=concat('a',r026).
```

```
D.-1:d
if (v002a=1 and (v026a≠v026c or v026b≠v026d)) r026=concat('d',r026).
```

```
D.-1:e
if (any(v002a,2,3) and (v026a≠v026d or v026b≠v026c)) r026=concat('e',r026).
```

```
E1.-1:a
if v033a≠1 and (
not(sysmis(V041)) or not(sysmis(V042)) or not(sysmis(V043)) or not(sysmis(V044)) or
not(sysmis(V045)) or not(sysmis(V046)) or not(sysmis(V047)) or not(sysmis(V048)) or
not(sysmis(V049)) or not(sysmis(V050)) or not(sysmis(V051)) or not(sysmis(V052)) or
not(sysmis(V053)) or not(sysmis(V054)) or not(sysmis(V055)) or not(sysmis(V056)) or
not(sysmis(V057)) or not(sysmis(V058)) or not(sysmis(V059)) or not(sysmis(V060)) or
not(sysmis(V061)) or not(sysmis(V062)) or not(sysmis(V063)) or not(sysmis(V064))
) r041=concat('a',r041).
```

```
if v033a=1 and (
(sysmis(V041)) or (sysmis(V042)) or (sysmis(V045)) or (sysmis(V046)) or
(sysmis(V047)) or (sysmis(V049)) or (sysmis(V050)) or (sysmis(V052)) or
(sysmis(V053)) or (sysmis(V057)) or (sysmis(V058)) or (sysmis(V059)) or
(sysmis(V060)) or (sysmis(V061)) or (sysmis(V063))
) r041=concat('a',r041).
```

```
E1.-1:g
if (v036a≠1 and not(sysmis(v051))) or (v036a=1 and (sysmis(v051))) r041=concat('g',r041).
```

J.-1:a

if (v136≠0000 and not(range(v136,v003a,2006))) r0136=concat('a',r0136).

J.-1:e

if v136≠v018 and not(any(v137,2,4)) r0136=concat('e',r0136).

A partir del mencionado *Programa de inconsistencias*, el ordenador detectará, por apartados del cuestionario, aquellas encuestas donde se dan posibles incoherencias en las relaciones establecidas entre las diferentes variables.

Una vez pasado el programa sobre la base de datos provisional, se obtienen unas listas con las diferentes incidencias detectadas, los números de identificación de los cuestionarios afectados y las posiciones que ocupan en la base de datos los posibles códigos erróneos. Para poder identificar, cuando salgan los resultados, los problemas e inconsistencias detectados con el programa, se procede a dar las órdenes precisas y necesarias al ordenador. La manera de llevarlo a cabo ha sido la siguiente:

- **Fuera de rango:** se identificarán a partir de una "x" junto a la variable o posición del cuestionario donde se ha localizado el posible error.
- **Relaciones incoherentes:** se controlarán a partir de la inconsistencia observada (identificada por el apartado y la relación concreta a la que hace referencia), que aparecerá junto al número del cuestionario y las posibles posiciones/variables afectadas en la incoherencia.

Finalmente, a partir de todo este proceso, se obtiene lo siguiente:

Incidencias del apartado A del cuestionario (puede haber diversos subapartados: A.-1, A.-2,...)

Incidencia a): no se cumple la condición lógica a

Incidencia b): no se cumple la condición lógica b

Incidencia c): no se cumple la condición lógica c

x: fuera de rango

Nº cuestionario

Inconsistencia detectada*

Conjunto de posiciones
posiblemente afectadas en la
incoherencia

(*La letra de después del apartado indica la incidencia concreta que afecta al cuestionario concreto. Una x nos indica las posiciones que están fuera de rango)

Veamos un ejemplo real:

D.-1 "Pregunta 26": Nivel de estudios finalizados

a) si p239=14, 15, 16 -> se ha de contestar p247

si p247 está contestada -> p239=14, 15, 16

b) si p110=1 -> p241 se ha de contestar

si p241 está contestada -> p110=1

c) p245 se ha de contestar si el CF tiene pareja (par_cf=1)

d) si p30=1 -> p239=p243 y p241=p245

e) si p30=2,3 -> p239=p245 y p241=p243

h) si p237 o p238=1 -> p239≠1

si p239=1 -> p237 y p238≠1

Salida programa ordenador:

Cuest	Incidencia	p30	p110	p237	p238	p239	p241	p243	p245	p247
1358 D.-1:ax	1	2x	1	.	15	.	15	.	.	.
4454 D.-1:d	1	1	2	.	6	8	5	8	.	.
5349 D.-1:x	4	6	7x	.	14	.	5	4	24	.
7835 D.-1:b	2	1	3	.	3	.	3	.	.	.

A partir de las citadas listas, se revisa incoherencia por incoherencia para ver si es o no necesario un cambio de la respuesta grabada, lo cual conlleva a consultar en muchos casos el cuestionario afectado, donde se comprueban los valores originales. En algunas ocasiones se puede tratar de un error a la hora de grabar las diferentes posiciones del cuestionario; en otras, el problema viene del proceso de codificación, en el que se traspasó mal algún código (como, por ejemplo, las ya comentadas respuestas fuera de rango) o no se detectó alguna incoherencia en la relación entre diferentes variables (como pasa en algunos filtros del cuestionario o en las inconsistencias comentadas anteriormente).

Cuando es necesario realizar cambios, éstos se realizan manualmente en las listas, sobre el papel, para posteriormente pasarlos a una hoja EXCEL que permitirá realizar los cambios en la base de datos (véase figura 7).

Figura 7. Depuración. Hoja de incidencias (cambios)

(Grupo n: cuestionarios de la microbase de datos n)					
	Cuestionario	Posición	Variable	Dígito inicial (sin corregir)	Dígito corregido
1	1358	110	v006	2	6
2	1358	247	V027	.	125
3	4454	243	v026c	5	6
4	5349	237	V024	7	1
5	7835	241	v026b	.	2

Así, la rectificación no se realiza directamente sobre el archivo de datos, sino que la corrección de los errores se registra en un archivo, donde queda recogida la siguiente información: el número del cuestionario, la posición/variable que se ha de cambiar, el valor observado en la base y el valor final/corregido. Finalmente, una macro de SPSS se encarga de leer el archivo de errores y hacer las modificaciones en la base de datos.

De todas formas, y debido a la exhaustividad con que se ha realizado el *Manual de relaciones*, hay casos en que, una vez revisada una posible incoherencia concreta de un cuestionario, se observa que no lo es, de manera que no se debe cambiar nada. Esto es debido a que, si bien la mayoría de las inconsistencias del manual definen relaciones imposibles de darse (así, por ejemplo, y tal como hemos visto, el entrevistado no puede tener menos de 16 años, es decir, no puede haber nacido después de 1990), en otros casos se trata de controlar situaciones o relaciones poco comunes aunque no imposibles, lo que podríamos denominar inconsistencias preventivas (éste sería el caso de aquellas relaciones/inconsistencias entre variables que en el *Manual* incorporan la palabra "normalmente", como aquella que indica que no es común que el cabeza del hogar tenga menos de 20 años, aunque puede ser posible en algunos casos, o la que indica que tampoco es

común que la diferencia de edad entre él/ella y su pareja supere los 20 años). En dichos casos, si finalmente se observa que no hay ninguna incoherencia (por ejemplo, que el cabeza de familia tiene realmente 19 años), es conveniente llevar un control que nos permita recordar que, en realidad, en el cuestionario X no se da dicha inconsistencia, y que, por lo tanto, cuando se vuelva a pasar el programa de inconsistencias al final del proceso no se deberá tener en cuenta. Para ello se realiza una tabla donde se apunta el grupo (1,.....,n) de 500 cuestionarios al que pertenece la microbase de datos donde está grabado el cuestionario afectado, el número concreto de dicho cuestionario y la incoherencia (que no lo es en su caso). (Véase figura 8).

Figura 8. Depuración. Hoja de incidencias (cambios)

Grupo de 500 cuestionarios *	Número cuestionario	Código de la inconsistencia
Grupo 1	1301	A.-2:c
Grupo 1	2617	A.-3:c
Grupo 1	2783	A.-7:f
Grupo 1	2848	G.-1:d
Grupo 1	4893	J.-1:x
Grupo 1	7219	L-1:a

* Grupo de cuestionarios al que pertenece la microbase de datos donde están grabados los cuestionarios afectados.

2.2. Control continuo de la base de datos

Finalizadas las correcciones pertinentes de la base de datos, se vuelve a pasar el *Programa de inconsistencias* al grupo de 500 encuestas para comprobar si los cambios se han realizado correctamente. Lógicamente, dicha comprobación se realiza sin tener en cuenta las inconsistencias del listado mencionado al final del apartado anterior, donde, como vimos, se controlan las incoherencias que en realidad no lo son. El objetivo final se conseguirá cuando al pasar el programa de inconsistencias no aparezca ninguna, lo que demostrará que la depuración se ha realizado correctamente.

Una vez se dispone de la base de datos más o menos definitiva de las 500 encuestas, se trata de ver los resultados parciales de este grupo de cuestionarios, así como los del conjunto de encuestas acumuladas. Dicha lectura de frecuencias, donde se hace especial hincapié en los filtros entre preguntas, junto con algún cruce simple entre variables, permite asegurarnos de la fiabilidad de la base de datos obtenida a través de todo el proceso.

3. Base de datos definitiva

Una vez dada por buena la base de datos obtenida se realiza la tabulación. Paralelamente se le pasa la base de datos al equipo que ha diseñado la muestra para que realice las tareas de validación (para determinar el nivel de discrepancia en la generalización de los resultados a la población de la que se extrae la muestra), ponderación *a posteriori* (de acuerdo con el diseño muestral previsto), elevación (ponderación de las frecuencias muestrales para expresarlas en términos de magnitudes poblacionales) y para que obtenga los errores muestrales finales (según la muestra realmente obtenida).

Una vez realizadas estas operaciones, se vuelve a efectuar una lectura de frecuencias de la base de datos, con objeto de verificar una vez más los datos y poder dar por finalizado el proceso. Se obtiene así la base de datos definitiva y se puede dar por cerrado el proceso de trabajo de campo.

