TRATAMIENTOS LONGITUDINALES EN LOS CENSOS DE 2001

Izaskun Atxa, Jesús R. Marcos y Enrique Morán

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

TRATAMIENTOS LONGITUDINALES EN LOS CENSOS DE 2001

Izaskun Atxa, Jesús R. Marcos y Enrique Morán

Toledo, junio de 2004

Indice

Capítulo
1

Introducción

Para entender los tratamientos realizados sobre los Censos de 2000 (Censo de Edificios y Locales de 2000 y Censos de Población y Viviendas de 2001) por Eustat, resulta necesario subrayar que se entroncan en el marco de un proyecto más complejo de sistema de información nucleado en torno al Registro de Población puesto en marcha a partir de 1996.

En ese año Eustat, en colaboración con Ayuntamientos y Diputaciones Vascas, lleva a cabo, junto con la renovación padronal, su propio censo, identificado como 'Estadística de Población y Viviendas 1996' (EPV96), en orden a mantener series quinquenales de información censal, prácticamente disponibles para la Comunidad Autónoma de Euskadi, desde 1970. En esa operación ya se plantea la demanda histórica de los Ayuntamientos de utilizar los padrones actualizados para ejecutar la renovación. Con esa perspectiva se procedió a preimprimir los padrones de todos los Ayuntamientos y posteriormente a grabar la información actualizada.1

La experiencia de los Ayuntamientos, junto con la necesidad de definir un marco integrado de todas las estadísticas de población de Eustat y con la casi obligación de aprovechar el máximo posible de fuentes estadísticas -administrativas y propiamente estadísticas-, sirven de acicate para iniciar la construcción del citado Registro de Población (RP). Uno de los elementos de los que se parte son las propias tablas de información del censo de 1996.

Otro de los elementos que merece la pena destacar y que justificó la realización por Eustat de los Censos de Edificios y Locales de 2000 (CEL2000) –en colaboración con el INE-, y que surge de la construcción del Registro, viene a ser el llamado Subsistema de Territorio. Sin una buena base mínima cartográfica, junto con una perfecta identificación de las unidades estadísticas básicas, mal se podía pretender el mantener el núcleo del registro: las viviendas y su población.

La cartografía, la diponibilidad de directorios de edificios, viviendas y locales, así como la posibilidad de tratamientos de depuración e imputación entre las Estadística de Población y Viviendas de 1996, Censos de Edificios y Locales de 2000, otros ficheros auxiliares (estudiantes, personal docente, selectividad, autónomos, matrimonios, nacimientos, etc.) y las tablas de los Censos de 2001, fueron justificación suficiente para asumir las tareas citadas a través de un convenio firmado con el INE el 12 de mayo de 2001. A esas tareas se sumaron la colaboración en la traducción y edición de parte del material de campo, seguimiento de la recogida y escaneo de los Censos, además de completar la codificación de literales.

[1] Eustat, 1999, Principales Resultados de la Estadística de Población y Viviendas 1996. Pag. 30.

La arquitectura del Registro de Población, basada en tablas en una base de datos relacional, así como el diseño de herramientas adecuadas al citado montaje –CDR: módulo de codificación automática de literales y MIR: módulo de homogeneización, validación, de depuración e imputación- permitieron dar soporte a la nueva visión de tratamientos censales.

Capítulo
2

Arquitectura de los CPV01

El nuevo entorno planteaba, como hemos dicho, la necesidad de cambiar todas las herramientas habituales de tratamiento estadístico, a la vez que los propios formatos de archivo: pasamos de los ficheros planos a tablas de Oracle y de programas específicos en lenguaje FORTRAN o PASCAL, a aplicaciones PL-SQL y Visual Basic.

Tablas y vista: diseño

Podemos agrupar las tablas más importantes utilizadas en los procesos censales en tres tipos con relación a las funcionalidades buscadas:

CUADRO 1. Resumen de tablas empleadas en los CPV01. Eustat.

Tablas origen CPV01-INE

Cuadernos de recorrido-huecos (CPV_CRH)
Edificios (CPV_APP)
Vivienda (CPV_FSV)
Población-padrón (CPV_FPC)
Población-censo (CPV_FSC)

Tablas propias de donantes

- Tablas EPV96:

  Viviendas (EVV_EPVVIV)
  Familias (CPV_CARFAM96)
  Población (EPP_EPVPER)
  Vínculos familiares (CPV_VINPER)

- Tablas CEL2000:

  Tabla de edificios (CPV_CPVEDF) [2]

[2] La tabla CPV_CPVEDF contiene tanto la información de los Censos de Edificios y Locales de 2000, como la propia de edificio de los CPV01.

- Tablas auxiliares:

  Selectividad (CPV_SELEC)
  Alumnos de doctorado (CPP_ALUN12)
  Alumnos universitarios (CPP_ALUN3)
  Diplomados y licenciados universitarios (CPP_EGUN12)
  Personal sanitario extrahospitalario (CPV_EXTRAH)
  Personal sanitario hospitalario (CPV_HOSPIT)
  Autónomos (CPV_AUTO)
  Nacimientos (MTP_TITUBOL)
  Matrimonios (MTP_TITULAR)

- *Vistas familiares:*

  Cónyuges(V08_CPVPER_CONY)
  Padres-Hijos (V09_CPVPER_PAD)
  Madres-Hijos(V10_CPVPER_MAD)
  Hijos-Padres(V06_CPVPER_HIPAD)
  Hijos-Madres(V07_CPVPER_HIMAD)

Tablas destino CPV01-EUSTAT

Edificios (CPV_CPVEDF)
Vivienda (CPV_CPVVIV)
Población-censo (CPV_CPVPER)
Familias (AUX_VARFAM)
Territorio (CPV_TERRITORIO)

Si tenemos en cuenta las unidades de análisis, en los tres grupos de tablas encontramos tres básicas: edificios, viviendas y población. Los huecos se encuentran en las tablas INE y EUSTAT y familias, en el entorno Eustat, ya sea de donantes o de destino. Se ha descartado, por ahora, tratamientos de huecos CEL2000-CPV01, que no pasen por la caracterización del edificio donde se enclavan. En el caso de los huecos nos encontramos con una unidad de recogida, mientras que la familia resulta ser una derivada después de realizar los correspondientes tratamientos.

Completan el esquema censal la tabla de Territorio-Eustat (claves e identificaciones postales de portales y huecos) y la tabla de vínculos familiares. En esta tabla aparecen las claves personales de los individuos, las de las personas con las que tienen o han tenido un vínculo de conyugalidad (casado/pareja) o filiación, así como las fechas de establecimiento y finalización de dicho vínculo.

Para determinados tratamientos familiares, ya fueran de la propia información familiar, como de otras variables censales (el euskera como veremos más adelante), resultó y resulta adecuado generar vistas, que nos permiten poner en relación unidades de distinto nivel. Nos permiten poner en relación características de distintos miembros de la familia entre sí o tratar los individuos con características de las viviendas o viceversa. Todo ello sin necesidad de consumir recursos informáticos excesivos.

Antes de describir brevemente los procesos claves en los tratamientos censales pasaremos a resumir el diseño de las tablas.

CUADRO 2. Diseño de tabla de Eustat de información de los CPV01

Nombre	CPVPER Datos de las personas según el CPV01						
Claves	SECUENCI			**Prefijo**	CPP_		
Nº	**NOMBRE SIMBÓLICO**	**LONG**	**INTERVALO DE VALORES**	**TIPO**	**DESCRIPCIÓN**	**NOMBRE ORIGINAL**	**ORIGEN DE LA INFORMACIÓN**
					1 Claves		
1	CL_UPB	15		NUM	Clave única de habitante	IN2_CL_UPB	CPVINE
2	IDEV	8		NUM	IDEV actual asociado	IN2_IDEV1	CPVINE
3	IDEV96	8		NUM	IDEV de la vivienda de residencia a 1/5/96	EPP_IDEV96	EPVPER
4	CL_UTE	15		NUM	Clave de UTE actual	IN2_CL_UTE	CPVINE
7	SECUENCI	9		CHAR	Número secuencial de persona	FSC_SECUENCI	FSC
8	SECUEN	7		CHAR	Número secuencial de vivienda	CRH_SECUEN	CRH
					2 Datos de identificación		
15	DNIP	1	1,2,3	CHAR	Tipo de documento	FPC_TIDEN, Hom	FPC
16	NUMD	8		CHAR	Número de documento	FPC_IDEN, Nor	FPC
18	ALFA1	24		CHAR	Alfaclave 1º apellido	Gen	
19	ALFA2	24		CHAR	Alfaclave 2º apellido	Gen	
20	ALFAN	16		CHAR	Alfaclave nombre	Gen	
					3 Datos básicos		
21	F_NA	10	>1880	DATE	Fecha nacimiento	FSC_FNAC	FSC
22	I_FNA	3		CHAR	Origen de la imputación		
					4 Datos padronales		
					5 Datos familiares		
44	NFAM	2		NUM	Número de familia	FSC_NSOBRE	FSC
45	PERN	4		NUM	Número de persona	FSC_NORDF, Gen	FSC
46	REPP	1	1-7	NUM	Relación con la primera persona	FSC_PAREN, Hom	FSC
47	I_REP	3		CHAR	Origen de la imputación		
49	CONY	1	1,6	NUM	Figura el cónyuge	Gen	
50	I_CON	3		CHAR	Origen de la imputación		
51	NCON	2	1-30	NUM	Número de orden del cónyuge	Gen	
52	I_NCO	3		CHAR	Origen de la imputación		
53	CL_UPBC	15		NUM	Clave UPB del cónyuge	EPP_CL_UPBC	EPVPER
69	AMAT	4	1901-2001	NUM	Año de matrimonio o boda	EPP_AMAT	EPVPER
70	I_AMA	3		CHAR	Origen de la imputación		
					6 Datos censales		
109	LMAT	1	5-8	NUM	Lengua materna	FSC_LM, Hom	FSC
110	I_LMA	3		CHAR	Origen de la imputación		
					7 Características derivadas		
181	LNAC5P	1	1-6	CHAR	Lugar de nacimiento (proximidad)		
190	EKNG1P	1	1-7	CHAR	Nivel global de euskera		
191	EDMG	3	0-120	DEC	Edad de migración	CPP_ALLM-CPP_ANNA	
192	LPRO6P	1	1-6	CHAR	Lugar de procedencia (proximidad)		
193	LR916P	1	1-6	CHAR	Lugar de resid. a 1/3/1991 (proximidad)		
194	LSVI5P	1	1-6	CHAR	Lugar de segunda vivienda (proximidad)		

En el cuadro 2 presentamos una versión simplificada de una de las tablas principales de los censos de 2001, la tabla de personas.

En ella se pueden apreciar cuatro grupos básicos de columnas. Por un lado tenemos el de claves, información necesaria para poder tener relacionadas las unidades de una tabla con las del resto del esquema censal. Tenemos las necesarias para relacionar las tablas derivadas de los censos de 2001 con las informaciones previas de Eustat (Cuadro 2. Apartado de tablas propias de donantes) y que son comunes al RP: Clave única de habitante –CL_UPB-, identificadores de vivienda (IDEV) y de hueco (CL_UTE). Por otro, resultó necesario crear otras claves propias para las unidades censales para poder tener relacionadas todas las unidades censales, aunque no se pudieran asociar a todas las previas de Eustat (por ser altas en el censo o por simples problemas de identificación que derivaron en una no asociación en el proceso de fusión que se resumirá luego). De este tipo se dispone del SECUEN (clave de vivienda 'censal') y del SECUENCI (clave de individuo 'censal').

Un segundo apartado lo constituyen identificadores personales (DNI y alfaclaves de nombres y apellidos –encriptación de dichos elementos-) que tienen, entre otros objetivos, el de comprobar que las asociaciones son correctas y, a veces, servir ellos mismos de claves de relación. Este fue el caso de algunas tablas que aún no disponían de las claves del RP.

El núcleo central de la tabla, obviamente, lo constituye la propia información original, encuadrada en cuatro apartados: datos básicos, padronales, familiares y censales.

El último grupo de columnas lo constituye la información derivada: presentación de la información que supone una elaboración especial, ya sea por tenerse que generar con varias columnas de las tablas (el nivel global de euskera, resumen de información de las cuatro preguntas de conocimiento o la edad de migración, diferencia entre año de llegada y año de nacimiento), o por ser agregaciones especiales (lugar de nacimiento, municipios próximos). Estas variables derivadas no sólo facilitan la futura difusión de la información, sino que sirven a todos los efectos para los propios tratamientos de depuración o imputación.

Como elemento específico de las filas, tenemos que subrayar la inclusión de marcas de imputación. Las tablas de la EPV96 ya contenían esta información, al objeto de poder diferenciar siempre la que había pasado todos los controles de validación frente a la que había sido depurada y/o imputada. Cada información sujeta a tratamiento tiene su fila correspondiente en donde se indica también a través de qué fuente ha sido imputada o modificada, en su caso.

Dada la evolución del peso de la no respuesta, parcial o total, en el conjunto de las estadísticas públicas, parece que se tiende a incluir este tipo de información en los ficheros de intercambio entre organismos estadísticos.3 A su vez Eurostat solicita para la construcción de algunos indicadores estructurales, la inclusión sólo de información no imputada.4

[3] Reglamento (CE) Nº 1983/2003 de la Comisión de 7 de noviembre de 2003 por el que se aplica el Reglamento (CE) nº 1177/2003 del Parlamento Europeo y del Consejo relativo a las estadísticas comunitarias sobre la renta y las condiciones de vida (EU-SILC) en lo referente a la lista de variables objetivo principales. Diario Oficial de la Unión Europea. L298/34-ES-17.11.2003

[4] http://europa.eu.int/comm/eurostat/newcronos/queen/display.do?screen=detail&language=en&product=YES&root=YES_copy_539019591709/strind_copy_817397594099/socohe_copy_88803726593/sc051_copy_463637803247

Procesos previos: las fusiones

Dentro del conjunto de tratamientos censales, no cabe duda que los relativos a la vinculación de unidades entre las distintas tablas y procedencias, han sido los más complejos y también más necesarios. Sin ellos los trabajos de depuración e imputación longitudinal –misma unidad con información de dos momentos distintos- o incluso transversales –unidades de distinta procedencia o de distinto tipo entre sí- no podrían llevarse a cabo.

Como dijimos en la introducción, Eustat lleva a acabo los Censos de Edificios y Locales de 2000 en colaboración con el INE, operación previa a los de Población y Viviendas, ya que suponen la identificación de las unidades espaciales –Edificios y Viviendas- en donde se va a llevar cabo la encuestación de los segundos. Sin estos censos previos no queda perfectamente garantizada la cobertura. Como decimos, en ellos se censan edificios, huecos –ya sean destinados a viviendas o a locales- y , en nuestro caso, establecimientos económicos.

Realmente Eustat en esta operación plantea una actualización del territorio contenido en su RP (que parte de la EPV96, de los padrones municipales del 96 al 99, de sus callejeros y de la planimetría incluida en su SIG) y de los establecimientos de su Directorio de Actividades Económicas –DIRAE-. Entre otras innovaciones en los Censos de 2000, Eustat diseña una recogida-actualización de información a través de un procedimiento CAPI, con gestión telemática de flujos de tablas de información. Así pues, un conjunto de agentes censales recorren horizontal –se revisan todos los portales y se recoge una información adicional- y verticalmente –se actualizan todas las identificaciones postales de los huecos y los propios huecos- de la C.A. de Euskadi, con un ordenador portátil, recibiendo por correo electrónico las unidades de revisión –cuadernos=secciones- y enviándolas por la misma vía a Eustat. A su vez se revisa y prepara la cartografía para los censos de 2001, herramienta clave una vez que han desaparecido los llamados trabajos preliminares.

El INE se plantea para el 2001 una operación similar en su formato a la de Eustat de 1996, esto es, una actualización de la información disponible con origen padronal mediante la preimpresión de esta en los cuestionarios censales. Por otro lado, al no haber realizado los Censos de Edificios y Locales en todo el territorio estatal, se diseña incluir su recogida en los propios censos de 2001, a través de utilizar los tradicionales cuadernos de recorrido también como cuestionarios censales. Esta circunstancia supone que sólo disponen información de viviendas ocupadas (no de vacías o de locales) y de edificios con viviendas ocupadas (no del resto de edificios).

Se decide, por tanto, añadir a las viviendas que el INE disponía de origen padronal (con sus identificaciones propias) el resto de huecos y portales de que disponía Eustat actualizados con los Censos de Edificios y Locales de 2000 (también con sus identificaciones postales normalizadas). Para determinar ese resto, y en orden a no trasladar a preimpresión huecos duplicados (desde Eustat sólo se podían añadir viviendas vacías, locales y edificios sin viviendas ocupadas), resulta necesario realizar un primer diseño de fusión de ficheros. En este momento se construye un sistema de fusión de ficheros padronales-INE/tablas RP-Eustat, que con ligeras modificaciones se volverá a repetir una vez finalizada la recogida, escaneo y tratamientos previos de los CPV01 por el INE y enviados los ficheros de forma definitiva el 26 de diciembre de 2002 a Eustat.

A pesar de los intentos realizados en esta primera fusión, Eustat detectó la existencia de posibles huecos duplicados, por lo que se transmitió a las dirección de campo del INE la necesidad de controlar especialmente los casos más notables (las secciones completas más afectadas).

Pasamos a continuación a resumir el proceso de fusión de personas y los resultados en los CPV01.

Para la identificación de una persona en ambas fuentes –Padrón INE/RP-EUSTAT- se han utilizado por un lado, DNI (excluyendo duplicados) nombres y apellidos normalizados (construcción de alfaclaves que obvian abreviaturas, grafías distintas, etc), fecha de nacimiento y sexo; por otro se comprueba y valora la secuencia PROV+MUN+CALL+NUM+PISO+MANO+PUERTA.

Para este último caso se evalúa el grado de aproximación geográfica al igual que se hizo en el caso de la construcción de los ficheros de preimpresión censales, mediante un contador de coincidencias. Además se crea un nuevo contador que puntúa además las coincidencias en los identificadores personales. Todas las igualdades encontradas se guardan en una tabla (CPV_FUSION), que contiene las claves de ambas fuentes y el contador citado, que se llamó FUS_PESO, que no deja de ser una valoración numérica del grado de igualdad entre las unidades de la tabla de padrón del INE y las del RP de Eustat.

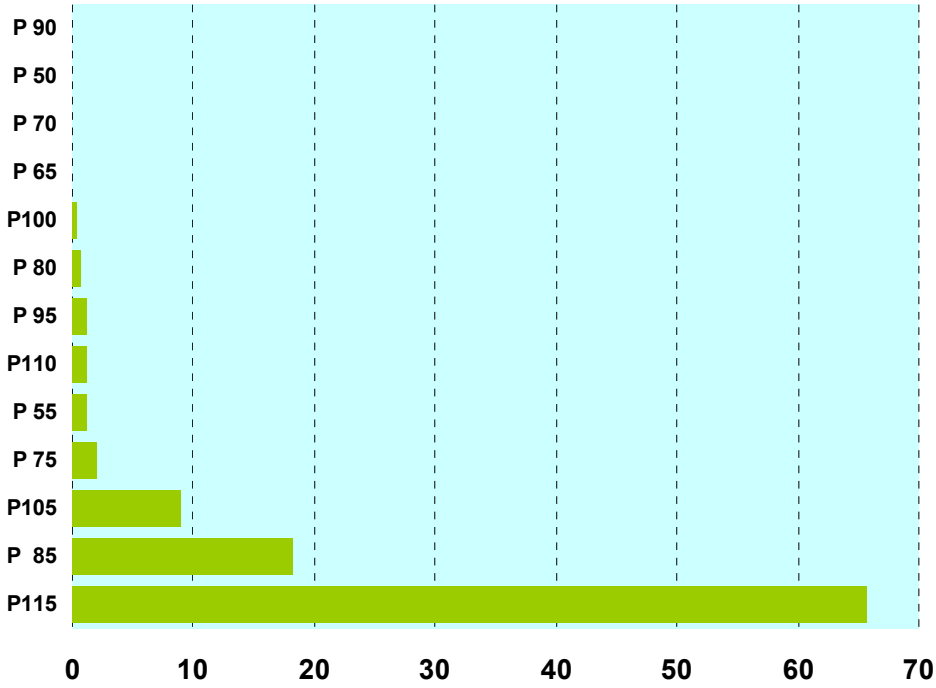
CUADRO 3 Valoración de las igualdades en el proceso de fusión de personas.

CONCEPTO	SUBCONCEPTO	PUNTUACIÓN
PERSONAS		
DNI		30
Nombre y apellidos		
	Nombre	10
	Apellido 1	10
	Apellido 2	10
Sexo		10
Fecha de nacimiento		
	DD	5
	MM	5
	YYYY	5
RESIDENCIA		
Hasta portal		10
Hasta escalera		20
TOTAL MÁXIMO		**115**

Con las puntuaciones obtenidas se procedió a excluir las uniones no válidas, siguiendo el criterio de desechar las que para una misma clave de Eustat o del INE eran –las puntuaciones- inferiores al máximo posible de 115 y las duplicadas con valor máximo.

Después de este proceso, se procede a asignar definitivamente las claves personales de Eustat. Para un total de 2.082.587 registros procedentes de la tabla de la revisión padronal del INE de 2001, se consiguieron fusionar 2.031.020 registros, lo que supone el 97,52%.

Gráfico 1. Distribución de las puntuaciones asignadas en la la fusión INE-EUSTAT de personas CPV01 por frecuencia.Max=115.%.



Fuente: Eustat

Del total de unidades fusionadas el 93% consiguió puntuaciones de 85 y más, y hasta el 66% del máximo. Los 85 puntos indicaban que coincidían todas las claves personales comparadas pero no las correspondientes a la identificación postal.

Este resultado, próximo al obtenido en la preparación de los ficheros de preimpresión de los censos, justificaba o sustentaba suficientemente la metodología que se pretendía implementar en los tratamientos censales.

Con viviendas y edificios se procedió a realizar procesos de fusión con una metodología similar.

Resumiendo, diremos que se diseñaron procesos de carga de ficheros INE a tablas Oracle en una base de datos de Censos Eustat. Estas mantienen la misma estructura de los ficheros INE. Se ejecutan los procesos de fusión. Posteriormente se cargan las tablas Eustat de censos, tanto con la información INE como con las columnas propias Eustat (con información homogeneizada o procedente de otras fuentes).

A su vez se diseñan los procesos de vuelta a las tablas del INE, ya que los tratamientos censales se ejecutan en las tablas Eustat y luego se trasladan a las tablas INE.

Una vez definidas las relaciones y flujos entre tablas Eustat e INE se procede a realizar las tareas de codificación automática y de validación, depuración e imputación.

Herramientas de tratamiento censal: CDR y MIR

La decisión de trasladar los trabajos censales a un entorno de base de datos relacional, determinada por la puesta en marcha del RP en dicho entorno, obligó a elaborar nuevas herramientas para interactuar con dicho entorno. Los objetivos se fundamentaban en mantener todas las funcionalidades de las aplicaciones que se habían diseñado en operaciones censales anteriores, en permitir a los técnicos estadísticos trabajar con la mayor independencia posible de los técnicos informáticos y en ampliar el uso de las citadas herramientas a todos los usuarios del entorno del RP.

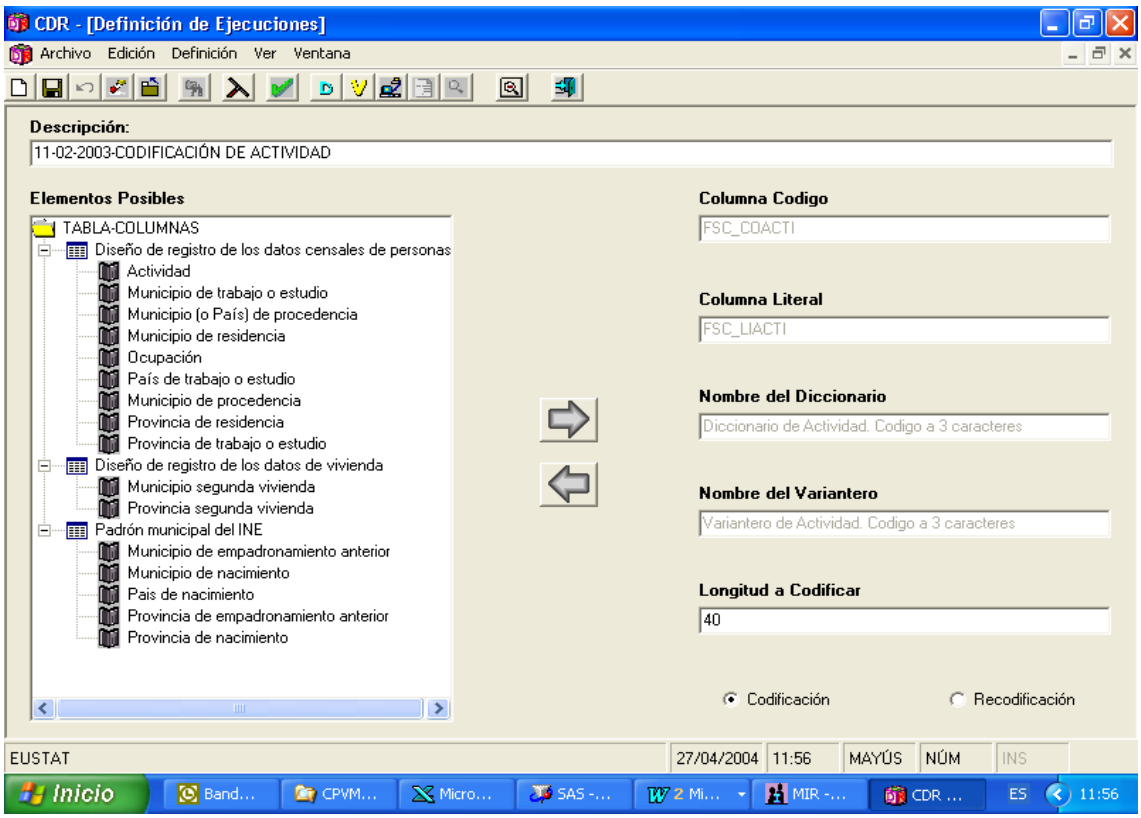
Si el primer objetivo se ha cumplido sobradamente, ya que el entorno de base de datos permite y facilita la inclusión de funcionalidades nuevas, del segundo no puede decirse lo mismo sin matizaciones. Aunque las herramientas nuevas permiten una independencia en el diseño y ejecución de tratamientos estadísticos, su mantenimiento y gestión, así como las propias incógnitas que plantea toda innovación, han impedido la autonomía que se deseaba. Con respecto al tercer objetivo, aunque se han utilizado en alguna operación menor, aún no se han generalizado.

Dos han sido las herramientas diseñadas para los tratamientos censales y que resumimos brevemente: el módulo de codificación automática (CDR-codificación del registro) y el de homogeneización, validación, depuración e imputación (MIR-módulo de imputación del registro).

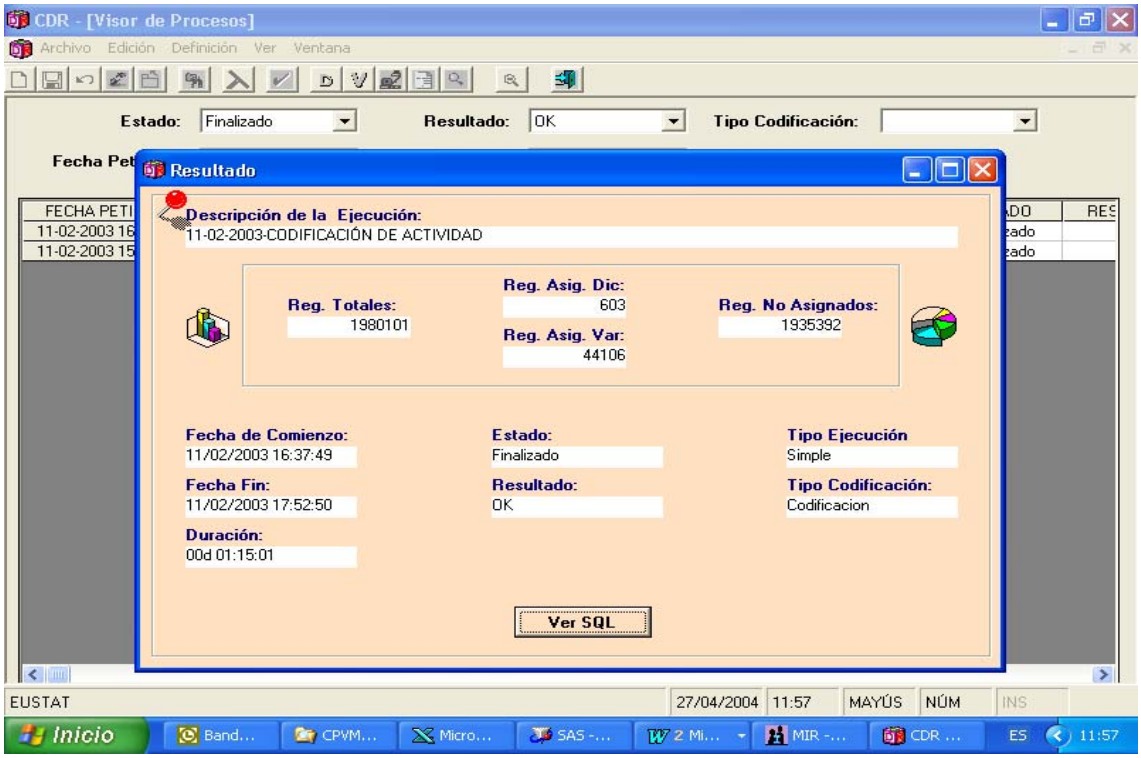
El módulo CDR diseñado en PL-SQL y Visual BASIC (hay ya una versión en PuntoNET), incluye aplicaciones para codificación automática de literales simples o condicionados (provincia-municipio, etc.) por el método de igualdad, de codificación manual asistida para literales no incluidos en los varianteros previos, así como las propias de gestión.

No obstante hay que decir que en estos censos la codificación de literales se realizó en su mayor parte en las propias oficinas del INE, quedando sólo un resto para las CCAA que lo tenían previsto en su convenio. Recordaremos que Eustat envió en su día sus propios varianteros (derivados de los procesos censales de la EPV96) al INE. En el caso de la C.A. de Euskadi, se han codificado 61.566 literales de actividad (un 7,9% del total), 67.260 de profesión (8,7%) y 59.600 correspondientes a campos geográficos.

Pantalla 1. CDR-Selección de Ejecuciones.



Pantalla 2. CDR-Resultado de la Ejecución.

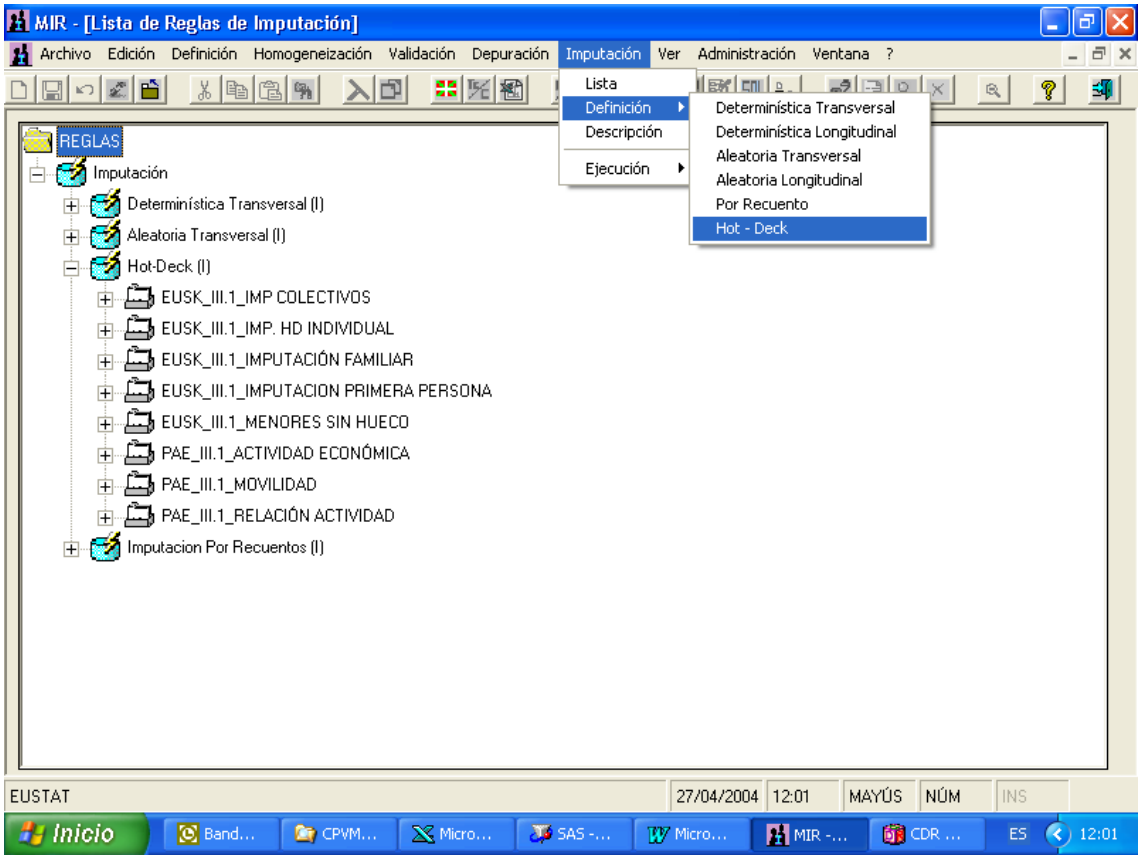


Eustat

El módulo MIR programado en los lenguajes utilizados en CDR y también diseñado para trabajar con tablas de Oracle, incluye las funcionalidades básicas para poder realizar todos los tratamientos censales clásicos. Estos se agrupan en submódulos: de homogeneización, validación, depuración e imputación. Dentro de este último se incluyen procedimientos determinísticos, aleatorios y de HOT-DECK. Además se han previsto submódulos de generación de tablas de recuento –recuento de distintas variables por unidades básicas como vivienda, familia, edificio, etc., a fin de poderse utilizar como variables de imputación-, generación de vistas, de generación de columnas, de números aleatorios y algunas otras de menor entidad.

Como dijimos, la gestión y mantenimiento de dichas herramientas, así como de la propia base de datos y sus componentes, en ambos casos situaciones novedosas para Eustat, han supuesto la mayor fuente de dificultades, no así la 'amigabilidad' de las propias herramientas y las funcionalidades previstas.

Pantalla 3. MIR-Selección del Módulo objeto de trabajo.



Capítulo
3

Calidad de la información recogida en los CPV01

Desde un entorno de base de datos relacional se puede dividir la calidad de la información censal en tres apartados: calidad de contenido, calidad de estructura y cobertura.

La primera se configura por el nivel de falta de respuesta y de errores de contenido, la segunda por el grado de consistencia entre los distintos ficheros censales (portales/edificios-huecos-personas), por la idoneidad de los identificadores, por la coherencia de los propios procesos de construcción de los ficheros, cuando, como en este caso, las fuentes son diversas y por la correcta definición de campos y contenidos.

Con respecto a la cobertura, aunque se han realizado comparaciones con los resultados en cuanto a edificios, huecos (viviendas y locales) obtenidos en el CEL2000, aún no se tienen resultados definitivos. A ese efecto, además de medir los propios errores de contenido, Eustat realizó, como ya viene siendo tradicional desde 1986, una Encuesta de Validación de los Censos. Una vez que se cierren los tratamientos censales se procederá a explotar dicha encuesta.

Asimismo, de forma experimental, se está procediendo a validar la información censal disponible con los ficheros de la PRA (Encuesta de Población con Relación a la Actividad), en orden a valorar la posibilidad de utilizar en el futuro exclusivamente dicha encuesta.

Calidad de estructura

A parte de diferencias de formatos o cambios en relación a los ficheros de preimpresión y diseños previos, las principales carencias estructurales fueron: las derivadas de incoherencias entre ficheros/tablas: viviendas sin asociar a huecos (12.326), personas sin hueco (20.782). Se detectan personas en viviendas de baja o viviendas ocupadas en huecos vacíos o cuyo uso no es vivienda. Aparecen 3.244 edificios con huecos sin APP y 936 edificios con APP pero sin huecos. A estas carencias hay que sumar las habituales en todo fichero de población: personas duplicadas, ficticias (identificadores no válidos), menores (5.027) sin vivienda o sin familia, etc.

Estas carencias se van resolviendo teniendo en cuenta que no se podían variar la población base de individuos ni sus características padronales (se mantienen, por ejemplo, los duplicados encontrados).

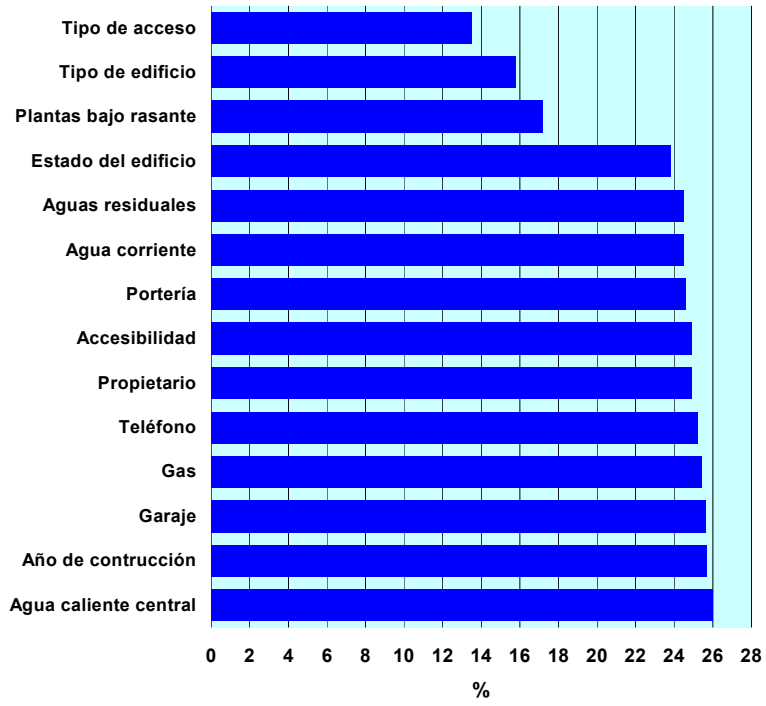
Calidad de contenido: la falta de respuesta.

Como hemos dicho, la consistencia de la información resulta preciso ser medida por los tradicionales métodos de validación a través de encuestación repetida. Quedan, por otro lado, las carencias debidas a la falta de respuesta (parcial o total) y las asociadas a las incoherencias o valores no válidos entre los datos de una misma unidad censal (o entre unidades censales). De estas últimas se dará una visión general en el caso que posteriormente se expondrá.

Hay varias vías de analizar la falta de respuesta. Comenzaremos por la temática y seguiremos por la espacial.

Dado que aún no se ha definido el censo de edificios derivado de los CPV01, no daremos datos definitivos de falta de respuesta, aunque sí unos provisionales. La falta de respuesta en las características de los edificios (recordemos que debía ser recogida por los agentes censales en los Cuadernos de Recorrido), se aproximaría al 25% en la C.A. de Euskadi.

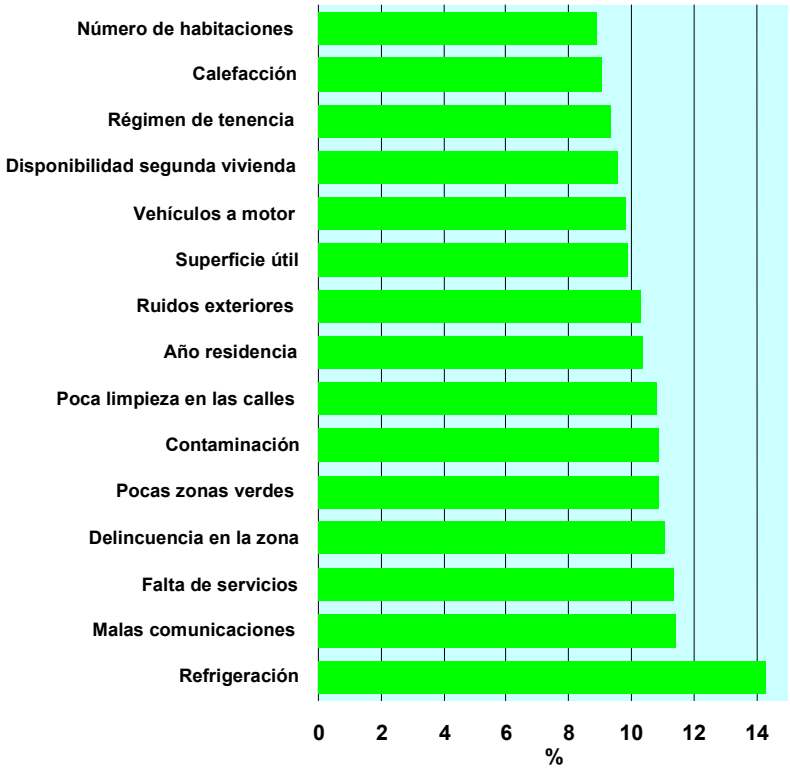
Gráfico 2. Caracteríscas de edificio según la falta de respuesta.C.A. de Euskadi. CPV01.%.



Fuente: Eustat, CPV01

La falta de respuesta en las características de viviendas se sitúa cerca del 10% en casi todos los casos, para un total de 759.830. Frente a las preguntas de carácter objetivo (nº de habitaciones, calefacción, superficie útil, etc.), las de mayor contenido opinático (malas comunicaciones, falta de servicios, etc.) superan ligeramente la media de la falta de respuesta.

Gráfico 3. Caracteríscas de vivienda ocupada según la falta de respuesta.C.A. de Euskadi. CPV01.%.



Fuente: Eustat, CPV01

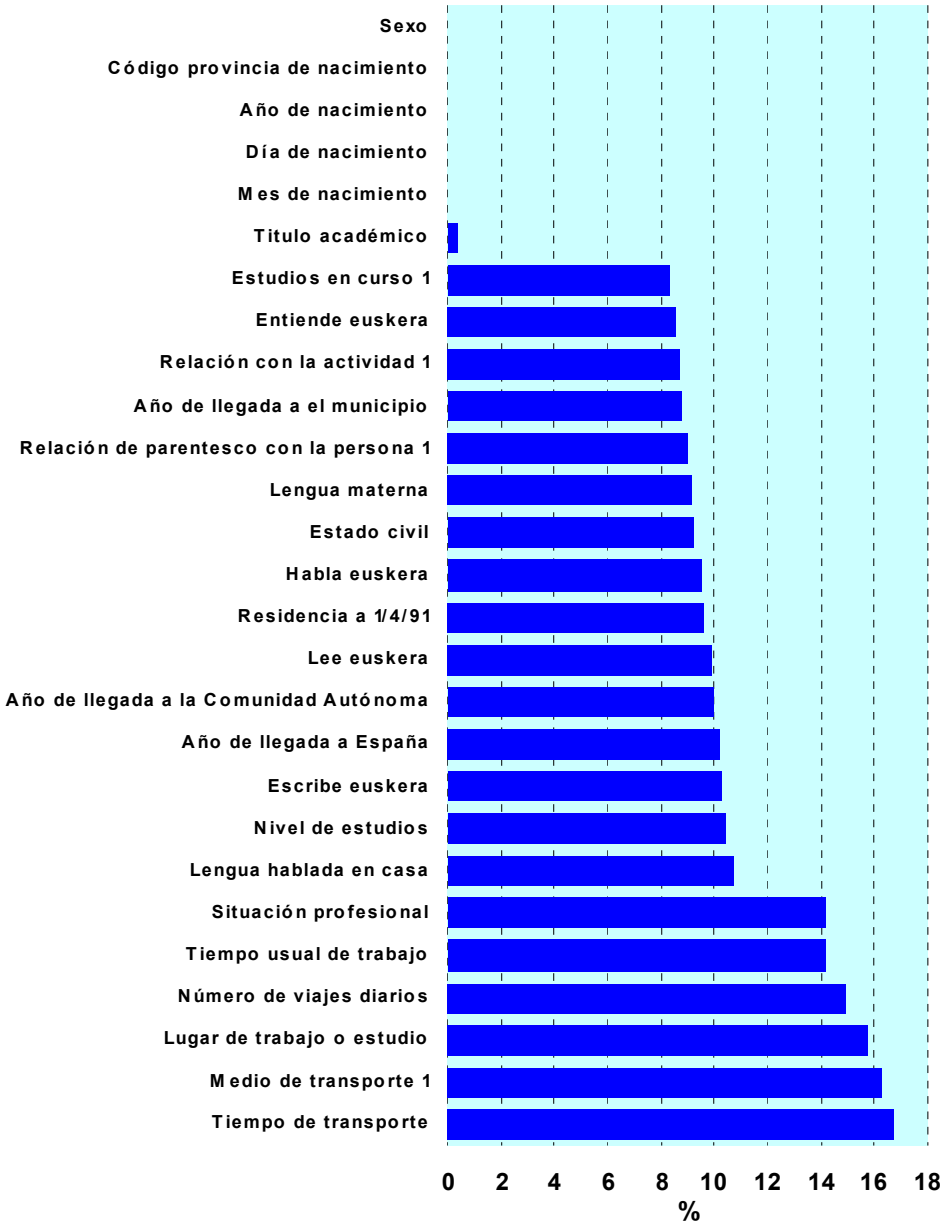
Comparemos con algunos datos de los Censos de 1991.

El número de habitaciones en 1991 faltó en el 1,3% de las viviendas ocupadas y en el 40,1% de las vacías. Ponderando el peso de ambos grupos daría una falta de respuesta del 8,1%. En este caso las diferencias entre censos no son apreciables.

La superficie útil en 1991 tuvo una falta de respuesta del 10,7% (en viviendas ocupadas un 4,8% y en vacías un 38,6%), cifra sustancialmente inferior a la de 2001, que ofrece un resultado cercano al 10% para la vivienda ocupada.

Como se puede apreciar, existen diferencias importantes de falta de respuesta entre un censo y otro.

Gráfico 4. Caracteríscas de personas según la falta de respuesta.C.A. de Euskadi. CPV01.%.



Fuente: Eustat, CPV01

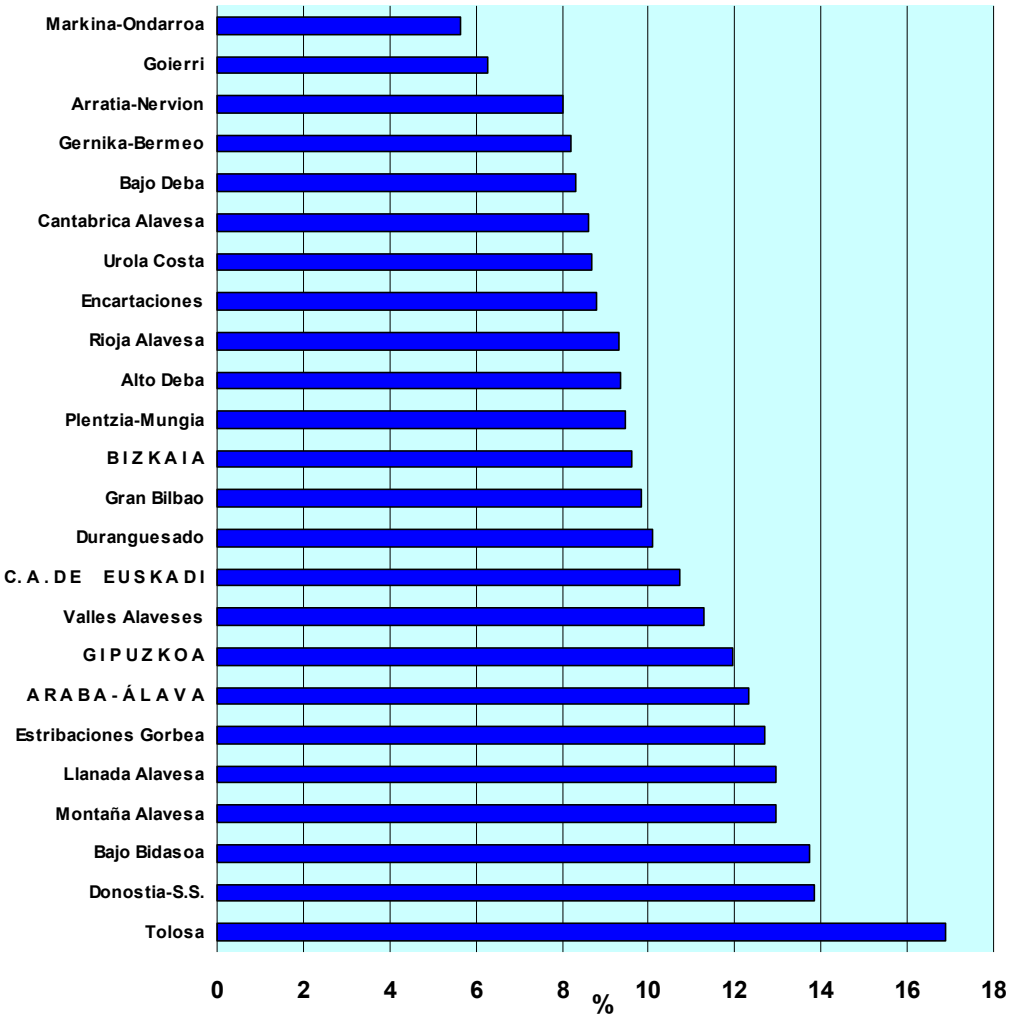
En la cobertura de la información de personas se pueden diferenciar tres grandes bloques: información procedente de los cuestionarios de revisión padronal (y de los propios padrones municipales), información del cuestionario censal familiar y del cuestionario censal individual.

En cuanto a las variables padronales hay que decir que han llegado sin falta de respuesta en la mayor parte de los casos. Sólo se apreciaron pequeñas faltas de respuesta e incoherencias entre día y mes de nacimiento y una pésima calidad de la variable titulación (aparecen 10 veces más analfabetos de los esperados, etc.).

Las variables derivadas de los cuestionarios familiares se aproximan al 10% de falta de respuesta. Desde el 8,6% en la pregunta sobre nivel de conocimiento del euskera (entiende) al 10,7% de la lengua hablada en casa.

Estas últimas variables en 1991 se situaron en 1,85% la primera y en 2,68% la segunda. Este fuerte aumento de la falta de respuesta se podría deber sobre todo, al gran número de individuos no censados mantenidos o añadidos a los ficheros censales desde los padrones, aunque no hay que descartar el aumento de las negativas.

Gráfico 5. No respuesta a lengua hablada en casa por provincia y Comarca.CPV01.%.



Fuente: Eustat, CPV01

Las falta de respuesta más significativa se encontraría en las variables recogidas en el cuestionario individual. Hay que precisar que se incluye en la falta de respuesta la indefinición de los propios subgrupos a los que afectaría la carencia de información. Aún así se pueden apreciar entre 4 y 7 puntos más de falta de respuesta.

Para el análisis espacial de la falta de respuesta tomamos una variable, la lengua hablada en casa, que se sitúa en la media de la falta de respuesta de la variables incluidas en el cuestionario familiar. Cabe esperar que se dé una fuerte asociación, en cuanto a falta de respuesta, en otras variables del mismo tipo.

La distinta cobertura por provincias, que se puede apreciar en el Gráfico 4 sobre resultados de campo (92,6% en Gipuzkoa, 93% en Araba-Álava y 95,3% en Bizkaia), vuelve a apreciarse en la falta de respuesta. Por encima del 10,7% del conjunto de la C.A. de Euskadi se sitúan Gipuzkoa con un 12% y Araba-Álava con un 12,3%; Bizkaia se queda en un 9,6%.

Por comarcas, la de Tolosa presenta los peores resultados con una de cada seis personas sin información –16,9%-, seguida con la de Donostia-San Sebastián con un 13,8% y Bajo Biodasoa con un 13,7%. Entre un 13 y un 11,3% están cuatro comarcas alavesas: la Montaña Alavesa, la Llanada, Valles Alaveses y Estribaciones del Gorbea.

Markina-Ondarroa y el Goierri son las que mayor respuesta tuvieron, con un 5,7 y un 6,3% respectivamente.

Si descendemos a los municipios tenemos que algo más que una cuarta parte de los municipios superan la media de falta de respuesta para la variable utilizada, 65 en total.

De ellos hay algunos afectados sensiblemente de tamaño mediano: en Urnieta para más de un tercio de la población no hay datos -para el 35,1%-, en Tolosa y Pasaia para uno de cada cinco –21,9% y 20,4% repectivamente-. Las tres capitales también superan la media de falta de respuesta, destacando Donostia-San Sebastián –13,7%- y Vitoria-Gasteiz –13,1%-.

Dada la estabilidad de algunas variables censales (la propia lengua hablada en casa resulta ser una de ellas) y la necesidad de mantener series censales coherentes -como mínimo a nivel municipal y para las variables sociodemográficas básicas-, la ejecución de tratamientos de imputación longitudinales –copia de información no imputada de 1996- se convierte en un método central de tratamiento.

Asimismo, la disponibilidad de las marcas de imputación y el origen de las fuentes utilizadas, permitirá a los analistas dilucidar con mayor precisión las líneas de evolución intercensal, sobre todo en los casos de una falta alta de respuesta.

Cuadro 4. Municipios con falta de respuesta por encima de la media en lengua hablada en casa. C.A. de Euskadi. CPV01.%

		POBLACIÓN	NR%
1	URNIETA	5518	35,1
2	GAZTELU	152	29,6
3	BERROBI	566	27,4
4	ZIZURKIL	2820	25,4
5	VILLABONA	5672	22,2
6	TOLOSA	17642	21,9
7	RIBERA ALTA	522	21,8
8	PASAIA	15962	20,4
9	SAMANIEGO	308	19,5
10	ZUIA	1906	18,1
11	LAUKIZ	995	17,8
12	LOIU	2199	17,2
13	ORENDAIN	143	16,8
14	IRUÑA DE OCA	1953	16,7
15	HONDARRIBIA	15044	16,7
16	BALIARRAIN	97	16,5
17	CAMPEZO	1071	16,2
18	OROZKO	2116	16,2
19	ALEGRIA-DULANTZI	1533	15,9
20	RIBERA BAJA	698	15,9
21	DURANGO	25003	15,8
22	HERNANI	18287	15,6
23	ANDOAIN	13814	15,0
24	MOREDA DE ALAVA	261	14,6
25	IRURA	910	14,5
26	PLENTZIA	3643	14,5
27	LEGUTIANO	1359	14,0
28	ARRAIA-MAEZTU	717	13,8
29	OIARTZUN	9179	13,8
30	ZARAUTZ	21078	13,8
31	DONOSTIA-SAN SEBASTIAN	178377	13,7
32	IGORRE	3857	13,7
33	SONDIKA	3978	13,5
34	ALONSOTEGUI	2662	13,3
35	OYON-OION	2464	13,2
36	ELGOIBAR	10440	13,1
37	BERANGO	5311	13,1
38	VITORIA-GASTEIZ	216852	13,1
39	IRUN	56601	12,9
40	VALDEGOVIA	952	12,9
41	LEZAMA	2113	12,9
42	VALLE DE ARANA	334	12,9
43	BARRIKA	1230	12,8
44	ARRASATE-MONDRAGON	23118	12,5
45	USURBIL	5257	12,4
46	AIA	1610	12,1
47	MUNGIA	13807	11,8
48	ABANTO	9036	11,8
49	IBARRA	4208	11,7
50	PEÑACERRADA-URIZAR	240	11,7
51	LASARTE-ORIA	17195	11,6
52	BILBAO	349972	11,5
53	GETXO	82285	11,5
54	ZALDUONDO	139	11,5
55	AMOREBIETA-ECHANO	16182	11,5
56	LEZO	5834	11,5
57	SOPUERTA	2245	11,2
58	NAVARIDAS	223	11,2
59	ZESTOA	3100	11,2
60	ZALDIBAR	2877	11,2
61	LAGRAN	197	11,2
62	ORIO	4421	11,1
63	GUERNIKA-LUMO	15264	11,0
64	GETARIA	2406	11,0
65	ARMIÑON	166	10,8

Fuente: Eustat, CPV01

Capítulo
4

Tratamientos longitudinales y transversales: el caso del Euskera

A modo de ejemplo y en orden a valorar los resultados de los tratamientos longitudinales ejecutados en los CPV01 por Eustat, presentaremos los procedimientos y las fases en que fueron aplicados a las preguntas relativas al euskera, incluidas en los cuestionarios familiares de los CPV01. Se incluyen el resto de tratamientos en el orden que fueron aplicados.

Aunque el conjunto de tratamientos se agrupa en cinco capítulos, cabe subrayar el que se desarrollan dos tipos de metodologías: tratamientos individuales -se depura, copia y/o imputa en base a información individual, agregada o no) y tratamientos familiares – se depura, copia y/o imputa en base a informaciones familiares-. Las variables del euskera, algunas como su propia definición indica 'lengua materna', contienen fuertes asociaciones en el entorno familiar, y por supuesto es variable con el conocimiento y el uso. Lo mismo sucede con las viviendas con relación al edificio y en otros casos.

Estas realidades obligan a tenerlas en cuenta a la hora de los propios tratamientos censales, máxime cuando se vienen realizando estudios de transmisión lingüística, que suponen una red de coherencias, por una parte de las relaciones familiares y por otra de la información del euskera según esas mismas relaciones.

Hay que decir que después de los procesos de fusión y centrándose en los individuos comunes, tienen las mismas claves el 93,6% de la población censada en 2001. Este será el conjunto general de donantes, del que se seleccionan los que son aptos para donar (en su caso valores no imputados en 1996 o por características personales que no afectan a la evolución del fenómeno a estudiar.

Después de unos breves comentarios relativos a las validaciones genéricas y específicas se presentan los grupos de tratamientos y sus fases.

Las variables referentes al conocimiento del Euskera (EKEN- Entiende, EKHA- Habla, EKES- Escribe, EKLE- Lee) así como las referentes a su uso (LHAB- Lengua Hablada en casa) o la primera lengua hasta los tres años (LMAT- Lengua Materna) son sometidas a un procedimiento inicial de validación genérica, así como de diferentes validaciones específicas, que nos permitirán enfrentar todo registro al conjunto de criterios establecidos.

Asimismo, este procedimiento nos revela el conjunto de registros con falta de respuesta en uno o varios campos.

Cuadro 5. Distribución de falta de respuesta en variables relativas al euskera por Territorio-C.A. de Euskadi. CPV01.%

	Entiende	Habla	Lee	Escribe	Leng. materna	Leng. hablada
ÁLAVA	9,5	11,0	11,2	11,6	10,0	12,3
GIPUZKOA	10,5	10,9	11,3	11,6	11,2	12,0
BIZKAIA	7,2	8,4	8,7	9,1	7,7	9,6
TOTAL	**8,6**	**9,6**	**9,9**	**10,3**	**9,2**	**10,7**

Fuente: Eustat, CPV01

Cuadro 6. Distribución de falta de respuesta en variables relativas al Euskera por Territorio-C.A. de Euskadi. CPV01.

	Entiende	Habla	Lee	Escribe	L. materna	L. hablada	POBLACIÓN
ÁLAVA	27.272	31.559	32.183	33.329	28.652	35.243	286.387
GIPUZKOA	70.455	73.594	75.954	78.396	75.382	80.506	673.563
BIZKAIA	80.721	93.967	97.439	101.787	86.936	107.974	1.122.637
TOTAL	**178.448**	**199.120**	**205.576**	**213.512**	**190.970**	**223.723**	**2.082.587**

Fuente: Eustat, CPV01

Tratamientos individuales

Procedimientos iniciales

En aquellos registros que carecen de información en una o más variables, pero no en todas, se aplicarán imputaciones determinísticas o, en su caso, aleatorias en base a la múltiple casuística de la información facilitada por el individuo, teniendo por objeto este procedimiento eliminar las inconsistencias existentes manteniendo al máximo la información original.

Los criterios establecidos siguen un orden jerárquico, estableciéndose en primer lugar, como variable básica definitoria del resto, la variable EKEN (entendimiento de la lengua); ésta determina en gran medida el grado de habla, lectura y escritura. En definitiva, se aceptan como más veraces los niveles de conocimiento más bajos declarados.

En estos tratamientos iniciales de depuración-imputación se han visto alterados 108.637 registros, de los cuales 5.975 corresponden a EKEN, 79.397 a EKHA, 93.627 a EKLE y 103.738 a EKES, afectando esta cifra a un 5,2 % sobre el total de la población. Del mismo modo, se han visto modificados 52.827 casos referentes a la Lengua Hablada.

Completando este proceso, se dispone de tablas auxiliares como fuente externa de información; éstas son:

- Datos de selectividad. (Año de referencia 1994-2002)
- Diplomados y licenciados universitarios. (Año de referencia 1991-02)

- Personal docente no universitario. (Año de referencia 1998-2002)

Estas tablas auxiliares se han utilizado para realizar imputaciones determinísticas exclusivamente en las variables referentes al conocimiento.

La utilización de fuentes externas de información afectó a 4.281 registros.

Al finalizar esta fase, los registros en todas las variables referentes a conocimiento dispondrían de información o carecerían de ella completamente.

Por otro lado, se procede a la generación de la variable de conocimiento global (EKNG1P) en función de las cuatro variables, anteriormente mencionadas, de conocimiento. Siendo utilizada esta nueva variable para los tratamientos posteriores.

Fase 2. Copia de valores de la EPV96

Tras un exhaustivo análisis y contraste evolutivo de la información existente en CPV01 y EPV96, se realizó un proceso de copia, para aquellos individuos mayores de 20 años a fecha 31-12-2000, de los valores existentes en EPV96 a CPV01 siempre que se cumpliesen dos criterios:

- Los registros EPV96 deberían ser originales y válidos, es decir, sin tratamiento de imputación.

- Los registros CPV01 deberían carecer de información en todos los campos.

En los Gráficos 6 y 7 presentamos las distribuciones de la falta de respuesta en la variable nivel global de euskera según la respuesta sin imputar a la EPV96 y la distribución en 2001 de esa misma variable en el Gráfico 8. Se puede apreciar el claro sesgo de la no respuesta de los erdaldunes (no entienden el euskera). Si repitiésemos los procesos clásicos de imputación, que tienden a mantener las distribuciones obtenidas con la falta de respuesta, estaríamos sesgando la información, hinchado artificialmente el censo de euskaldunes.

Esta copia determinística de información de 1996 afectó a 92.875 registros.

Asimismo, se procedió a la copia de 5.265 registros en lo relativo a conocimiento para aquellos individuos que, teniendo respondidas la Lengua Materna y la Lengua Hablada en CPV01, y siendo éstas coincidentes a sus correspondientes en EPV96, cumpliesen los requerimientos anteriormente mencionados.

Gráfico 6.- Distribución de no respuesta en conocimiento de euskera en 2001 según respueta en 1996 sin imputar. Edad a 31/12/2000. C.A. de Euskadi. EPV96-CPV01.%.

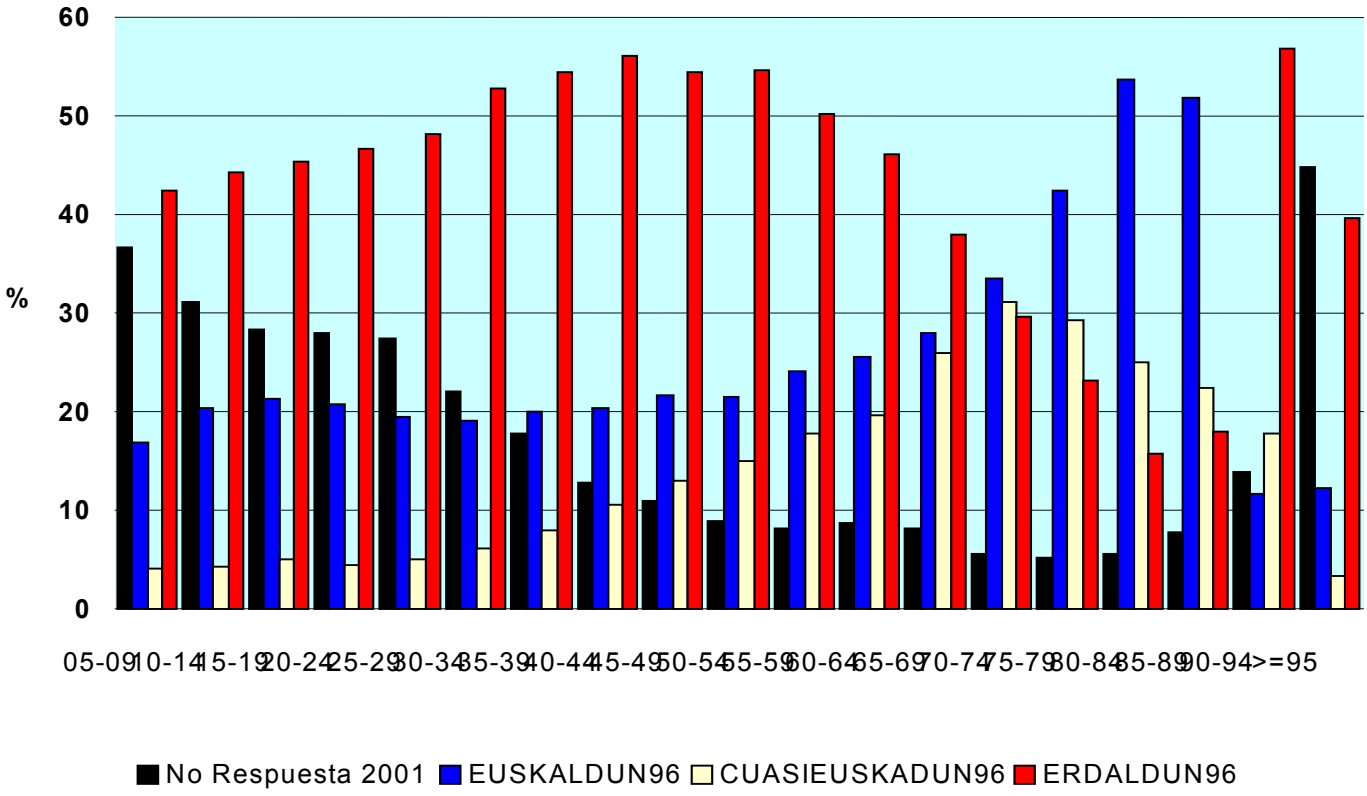
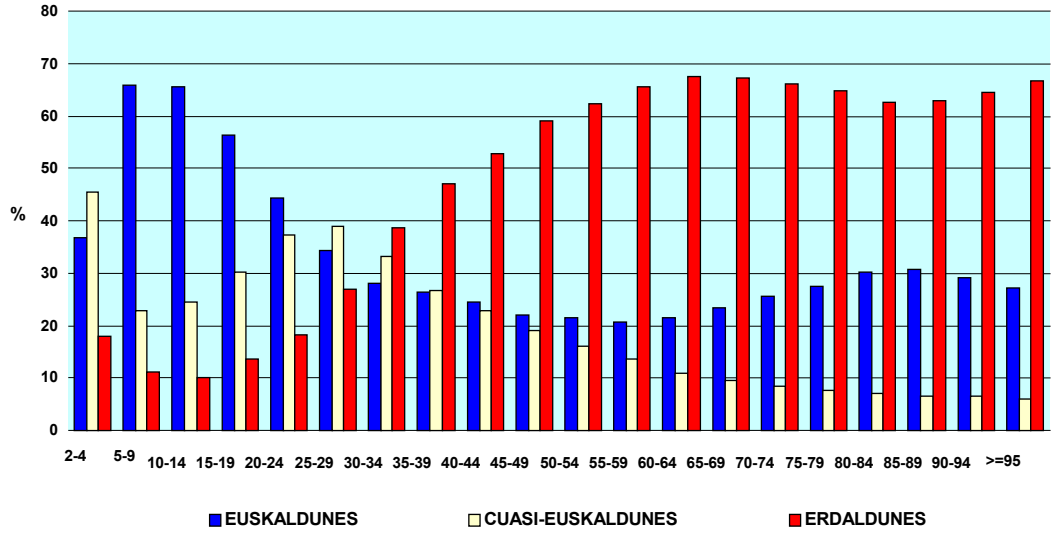


Gráfico 7. Población de 2 y más años por edad y el nivel global de euskera. 2001.%



Fuente: Eustat, CPV01

Fase 3. Interrelación conocimiento-uso

En esta tercera fase se finaliza el tratamiento individual, procediendo con aquellos registros que no contienen información en la variable conocimiento (EKNG1P) y sí en las variables del uso de la lengua o viceversa.

Para ello, se realizó una imputación HOT-DECK atendiendo a variables geográficas (Provincia y Municipio de Residencia) y de edad; mediante la cual se completó la información en 3.182 registros.

Asimismo, para aquellos registros que disponían de información en todas las variables excepto en Lengua Hablada, se realizó por el mismo método HOT-DECK la imputación, basada en Nacionalidad, Año de Nacimiento, Provincia y Municipio de Residencia, de 2.927registros. La Lengua Materna, por su lado, recibió similar tratamiento para un total de 2.126 registros.

Al finalizar el tratamiento individual, obtenemos un saldo final de 68.838 registros carentes de información en todas las variables relativas al conocimiento del Euskera.

Tratamientos familiares

Fase 1. Utilización de vistas familiares.

Las vistas se generan para poder ver dos o más veces la información de una misma tabla en una única sentencia, permitiéndonos ello, mediante la utilización de claves familiares, el recrear relaciones familiares básicas.

Para el caso que nos ocupa, se generan cinco vistas interrelacionadas:

- Padres - Hijo/as
- Madres - Hijo/as
- Cónyuges
- Hijo/as - Padres
- Hijo/as - Madres

En estas Vistas, aparte de las variables objeto de estudio y de las claves familiares necesarias para establecer las relaciones (UPB- Clave única de Habitante, UPBP- Clave UPB del Padre, UPBM- Clave UPB de la Madre, UPBC- Clave UPB del Cónyuge), se recogen variables de identificación de la vivienda y del individuo.

Se procede, con la información contenida en ellas, a la validación de las mismas, siendo utilizadas posteriormente para efectuar imputaciones atendiendo a criterios de coherencia familiar.

La secuencia utilizada en el proceso de imputación se inició con la vista procedente de los Cónyuges, a continuación se utilizó la generada con las Madres y seguidamente la obtenida por las claves de Padre, finalizando el proceso descrito la vista compuesta por aquellos registros que conforman la condición de Hijo/as.

En esta fase se procede a la imputación de 19.157 registros en todas las variables relativas al Euskera, asimismo, se implementa la información en 29.257 registros relativos a Lengua Materna y en 27.546 referentes a Lengua Hablada.

Fase 2. Tablas de recuento

Mediante la clave de vivienda y número de familia se generan tres tablas de recuento, una para cada una de variables (EKNG1P, LHAB, LMAT).

En estas tablas disponemos, además de otra información, de la moda de las variables anteriormente mencionadas por unidad familiar.

En aquellos miembros del grupo familiar que carecen de información se procede a la imputación de ésta por la moda familiar.

Este proceso afecta a 18.121 registros.

Una vez finalizada esta fase los registros carentes de información familiar ascienden a 31.560. De los cuales 11.045 corresponden a individuos residentes en colectivos y 20.515 a individuos que conforman 11.783 grupos familiares. De lo cual, se deduce que el mayor número de no respuesta corresponde a grupos familiares unipersonales.

Fase 3. Imputación a la persona de referencia

Mediante procedimiento HOT-DECK, se procede a la imputación de la persona principal de los 11.783 grupos familiares sin información. En esta fase se utilizan como variables básicas de imputación las relativas al lugar de nacimiento y la edad.

Tras completar la información de estos 11.783 registros, se repite el proceso de creación de tablas de recuento para la imputación por la moda de los 8.732 registros restantes.

Fase 4. Imputación de individuos en colectivos

Se utiliza como en la fase anterior el módulo HOT-DECK y se procede a la imputación de todos los individuos carentes de información residentes en colectivos atendiendo básicamente a variables referentes al lugar de nacimiento y edad.

Completando con ello los 11.045 registros.

De esta manera finaliza el proceso de imputación, quedando todos los registros con información en las variables referentes al Euskera.

Imputación por una variable sintética: el nivel global de euskera

Como ya indicamos anteriormente, trabajar con una variable de Competencia Lingüística global (EKNG1P) nos obliga a realizar la imputación de las variables objeto de estudio inicial, es decir, las variables EKEN, EKHA, EKLE y EKES.

Para ello se procedió mediante el módulo HOT-DECK, utilizando como registros donantes las variables EKNG1P, Lengua Hablada y Lengua Materna, condicionando todo ello a la edad de los individuos.

Corrección de errores de imputación

En los procesos de imputación, descritos en las fases precedentes, se produjeron, de forma inevitable, pequeñas inconsistencias; éstas afectaron, en el 98% de los casos, a los menores de seis años. Para subsanar esta deficiencia, se sometió toda la información a los controles utilizados en la primera fase.

En este proceso se modificó la información en 3.317 registros de la variable EKHA, 4.801 afectaron a EKES, 4.773 en lo referente a EKLE, completándose la fase con la modificación de 47 casos en lo que respecta a Lengua Hablada.

Asimismo, en este apartado de correcciones, se procede a la modificación de EKLE y EKES en 890 de los 13.383 registros que figuran como analfabetos en el Nivel de Instrucción.

Tratamientos de incoherencia entre conocimiento y uso

Fase 1. Vistas familiares

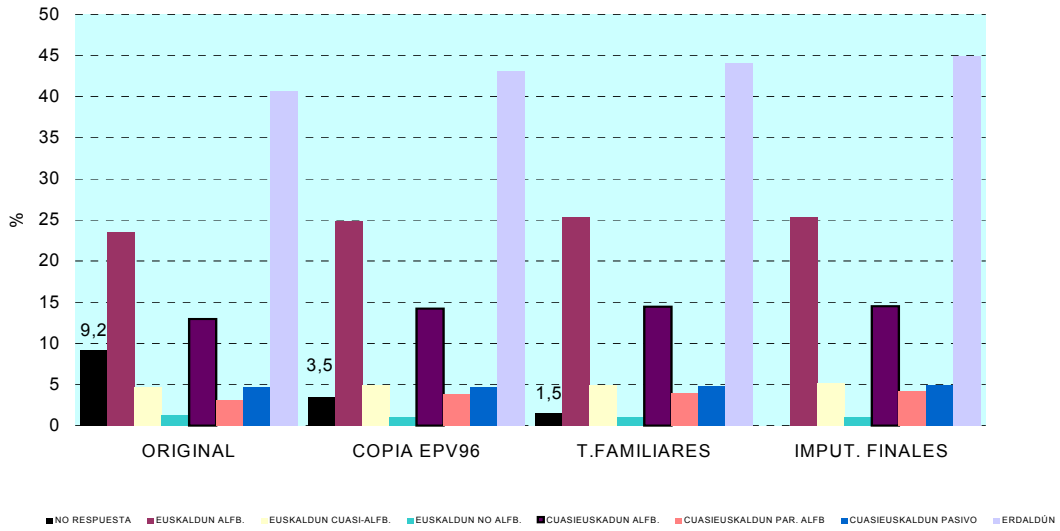
Utilizando las vistas familiares anteriormente mencionadas, se realizan las validaciones pertinentes, analizando la coherencia interna de la información en base a criterios familiares, dando lugar a una serie de depuraciones e imputaciones en la información entre las que cabe destacar la modificación de la variable Lengua Materna Euskera o Euskera y Castellano a Lengua Materna Castellano en aquellos registros en los cuales el Padre y la Madre (es necesaria la presencia de ambos) son Erdaldunes o Cuasi-Euskaldunes pasivos. Esta regla supuso alterar la información en 15.328 registros.

Fase 2. Tablas de recuento

Han sido utilizadas para analizar la coherencia de la lengua hablada entre los diferentes individuos de la familia.

Si existe más de una persona en la familia y una única persona declara hablar sólo en Euskera o sólo en Castellano, se corrige a uno de los miembros que afirma hablar en Euskera a Lengua hablada Euskera y Castellano, en este proceso se han visto afectados 13.483 individuos.

Gráfico 8. Peso de los tratamientos de imputación en el nivel global de euskera.
C.A. de Euskadi.CPV01.%.



Capítulo

5

Conclusiones

El esfuerzo de integrar una operación estadística compleja, como son unos censos de población y vivienda, dentro de un entorno configurado por un Registro de Población con carácter estadístico, con vocación de continuidad en el tiempo y de incrementar sus fuentes estadísticas, se ha justificado suficientemente por dos razones fundamentales.

Por un lado va a permitir mantener para los analistas las series censales quinquenales de Eustat en unas condiciones muy similares a las anteriores, al poder imputar en mejores condiciones –fundamentalmente con la copia de información de la Estadística de Población y Viviendas de 1996- un censo especialmente afectado por la falta de respuesta en población, acentuada si tenemos en cuenta el territorio. Esta demanda, ya no sólo de los efectivos que da la 'foto censal', sino de la evolución de la información en el tiempo se va a ver enriquecida por la posibilidad de realizar análisis longitudinales, que permitirán ver –por agregados de personas tomadas de una a una- con mayor precisión los cambios en las variables medidas y no solo como saldo de movimientos.

Teníamos 2.098.055 personas en 1996, 2.082.587 en 2001 y un subconjunto de 1.948.473 personas comunes a ambos censos. La evolución de un fenómeno en cinco años dependerá de los cambios –o de la estabilidad- de los elementos comunes, más los que aporten las altas posteriores al primer período y menos las de los individuos que desaparecen entre censos.

Los cambios metodológicos y técnicos necesarios para poder implementar estos nuevos sistemas de información, a su vez, nos plantean problemas nuevos. Sin ir mas lejos nos añade una nueva fuente de incoherencias –los cambios no permitidos, imposibles o improbables en un período de tiempo-.

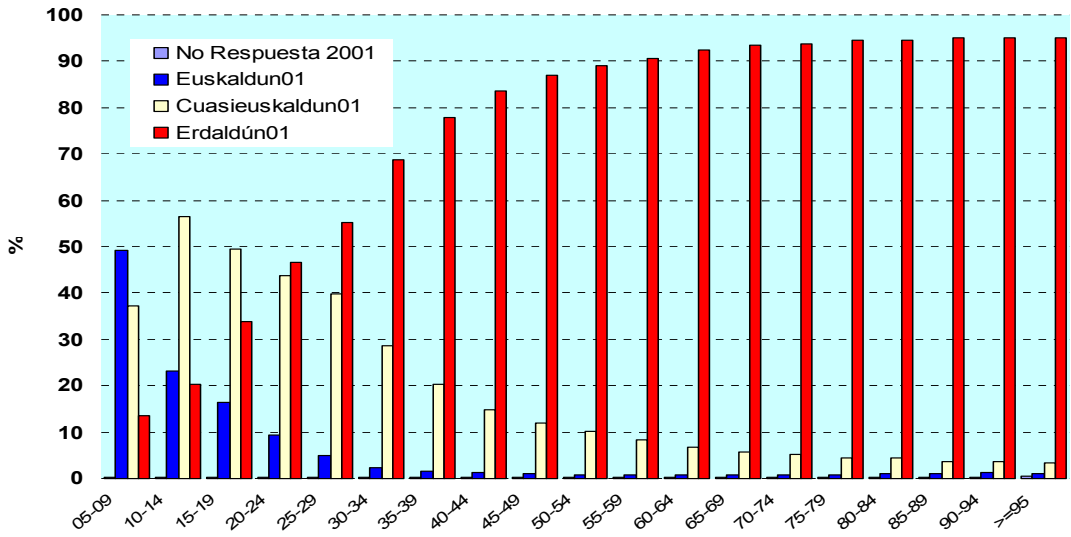
En el Gráfico 9 podemos apreciar dos circunstancias; por un lado la movilidad del euskera –el aprendizaje fundamentalmente en la escuela- en las primeras edades. Además se aprecia una pérdida de conocimiento en las edades poseducativas. Esto significa que esos grupos de edad no son susceptibles de ser utilizados como donantes. En su caso sí que nos serviría la distribución de la evolución para imputar.

Por otro la gran estabilidad de los grupos de mayores edades, salvo uno pequeños cambios de los llamados ilógicos o improbables: la euskaldunización a partir de los 50 años. Un buen análisis longitudinal nos llevaría a imputar con información que ha superado todos los controles en 2001 la información que no lo hizo en 1996, por otro, a pesar de pasar en ambos casos todos los controles, resulta necesario corregir los casos manifiestamente incoherentes, que aparecen nuevos.

Cabe subrayar la aparición de los llamados errores de estructura, derivados de las tareas de integración y relación, que en función del tipo de información que tratemos pueden ser determinantes. Una mala interrelación entre familias y viviendas puede producir una pésima cifra de viviendas vacías o una mala asignación de personas a viviendas, formas familiares completamente ficticias. Estas constataciones nos ponen en la pista de los principales problemas que siempre han subyacido en la estadística poblacional, pero que se obviaban, dadas las dificultades en la búsqueda de soluciones.

Por último, y más importante, la ejecución de los censos en un entorno relacional y con unas tablas de información histórica y auxiliares, ofrecen esperanzas fundadas para poder materializar censos sin necesitar recogidas generales de información, muy costosas y, como se comprobó en estos últimos, muy difíciles de llevar a cabo con los grados de calidad de épocas pasadas.

Gráfico 9 Erdaldunes en 1996 por nivel de euskera en 2001. Edad a 31/12/2000.



Fuente: Eustat, EPV96-CPV2001