ESTADÍSTICA ESPAÑOLA
Vol. 47, Núm. 159, 2005, págs. 321 a 351

Construcción de una población artificial basada en el Censo Español de Población de 1991

por
JORGE SARALEGUI
MONTSERRAT HERRADOR
CARLOS PÉREZ
FLORENTINA ÁLVAREZ

Instituto Nacional de Estadística

Mª DOLORES ESTEBAN
YOLANDA MARHUENDA
DOMINGO MORALES
ÁNGEL SÁNCHEZ
LAUREANO SANTAMARÍA

Centro de Investigación Operativa
Universidad Miguel Hernández de Elche

RESUMEN

En este trabajo se describe la construcción de un fichero de datos representativo de la población española basado en el Censo de Población y Viviendas de 1991. Este fichero, se construyó bajo el marco del proyecto europeo EURAREA (Enhancing Small Area Estimation Techniques to Meet European Needs, IST-2000-26290, 2001-2003.) con la finalidad de simular de forma realista diseños muestrales de encuestas de la Estadística Pública Española (Encuesta de Población Activa y Encuesta de Presupuestos Familiares) y de evaluar distintos procedimientos de estimación en áreas pequeñas.

Palabras clave: Universo artificial, imputación, modelos lineales generalizados, censo de población, encuesta de población activa, encuesta de presupuestos familiares.

Clasificación AMS: 62E30, 62J12

1. OBJETIVOS Y APLICACIONES DE LA POBLACIÓN ARTIFICIAL

Las actividades del proyecto de investigación EURAREA (Enhancing Small Area Estimation Techniques to Meet European Needs, IST-2000-5.1.8, 2001-2003.) en su primera fase, estuvieron orientadas a la creación de una población artificial por cada uno de los países participantes, puesto que la evaluación de los resultados de la investigación descansaba principalmente en la obtención de estimaciones del sesgo y del error cuadrático medio de las estimaciones mediante simulaciones de un número elevado de reiteraciones muestrales con esquemas similares a los utilizados por las encuestas oficiales en mundo real.

Los requerimientos de la población artificial debían cumplir al menos las siguientes condiciones:

1. Responder a una estructura de microdato en la que estuvieran presentes: a) los identificadores geográficos de las unidades tipo 'área pequeña' a investigar en el proyecto; b) las variables de identificación de las unidades últimas y de las diferentes etapas de muestreo que se utilizan en las encuestas oficiales que investigan las variables objetivo, así como la información complementaria (como el estrato y las características que determinan la probabilidad de selección) para la selección de muestras simuladas; c) un conjunto mínimo de variables auxiliares para utilizar en los modelos; d) las variables objetivo.

2. Tener como cobertura todo el territorio nacional, aunque en algunas fases de la investigación se utilizase solamente un subconjunto de regiones (NUT II).

Las variables objetivo, que debían estar disponibles en el microdato de la población artificial española para el proyecto EURAREA (APES) han sido las siguientes:

- Proporción de la población económicamente activa que está desempleada según la definición de la Organización Internacional del Trabajo (OIT).
- Proporción de hogares unipersonales.
- Ingresos anuales por unidad de consumo según la escala de equivalencia utilizada en las estadísticas europeas, como por ejemplo en el Panel de Hogares de la U.E. La escala de equivalencia de los ingresos familiares tiene como finalidad ajustar dichos ingresos a los diferentes tamaños y composiciones de los hogares.

En EURAREA se utiliza la llamada escala OCDE modificada. Esta escala asigna un peso de 1.0 al primer adulto, 0.5 al segundo adulto y a las personas siguientes con edad mayor o igual a catorce años y 0.3 a los menores de catorce años que componen el hogar. El tamaño equivalente de un hogar es la suma de los pesos asignados a cada persona. Así el tamaño equivalente resulta ser

$$1 + 0.5\,(N_{\geq 14} - 1) + 0.3N_{<14}$$

donde $N_{\geq 14}$ es el número de personas con catorce o más años y $N_{<14}$ es el número de niños con menos de 14 años.

La estructura y administración de la base de datos de la población artificial, según los objetivos de la investigación, debían a su vez permitir:

– Extraer con facilidad y rapidez un número muy elevado de muestras de diseño complejo (estratificado polietápico, con probabilidades desiguales), tanto de hogares como de personas.

– Con cada muestra obtenida según un diseño predeterminado, estimar los parámetros de los modelos formulados por el equipo investigador, a partir de las variables explicativas disponibles en las unidades de la población artificial, tanto en modelos de área (en algunos casos, completadas por variables de fuentes externas a este nivel), como en modelos de individuo.

– Estimar las variables objetivo en cada simulación de muestra y modelos, utilizando un conjunto de estimadores estándar para áreas pequeñas (directos, GREG, sintéticos y combinados EBLUP), u otros modificados, desarrollados a partir de la teoría en EURAREA.

– Obtener los indicadores de calidad empíricos del sesgo y del error cuadrático medio, para su análisis y representación, utilizando como 'verdadero valor' las variables objetivo proporcionadas por la explotación exhaustiva de la población artificial en cada área pequeña objeto de estudio.

– Repetir el experimento con diferentes grados de complejidad en el diseño y con diferentes combinaciones de diseño/modelos/estimadores e indicadores de evaluación empíricos.

2. DESCRIPCIÓN DE LA POBLACIÓN ARTIFICIAL

APES es una colección de ficheros en formato texto con 40 variables y 38.872.268 registros. La longitud del registro es de 90 caracteres. La unidad del registro es la persona, residente en una vivienda familiar principal en España, en la fecha de referencia del Censo (el 1 de marzo de 1991). El hogar al que la persona pertenece también puede identificarse a través de un número de identificación común para todos los miembros del mismo.

La descripción de los ficheros APES se presenta en la Tabla A.1. del Apéndice. Las primeras 35 variables se han obtenido del Censo Español de Población y Viviendas de 1991 y se han generado 5 nuevas variables: 2 imputadas a partir de información contenida en ficheros auxiliares y 3 obtenidas mediante la transformación de las variables anteriores, no siendo estrictamente necesaria la inclusión de estas últimas. En dicha tabla las variables APES se ordenan según la primera columna y no según su posición en el fichero del Censo Español de Población y Viviendas de 1991.

Las variables imputadas son:

· *Registro en las oficinas de empleo público* (APES501), según la Encuesta de Población Activa (EPA). Esta variable, obviamente, no estaba presente en el registro original del Censo, pero es necesaria en APES como variable explicativa en los modelos para estimar con las muestras simuladas el desempleo OIT (variable objetivo, presente como variable 'real' en APES). El demandante de empleo es la persona que está registrada en la Oficina Nacional de Empleo del Ministerio de Trabajo (Instituto Nacional de Empleo, INEM) para solicitar trabajo. A las personas entrevistadas en la EPA se les pregunta si son demandantes de empleo. Esta variable se refiere a la pregunta formulada en el cuestionario de la EPA: ¿Está usted inscrito en una Oficina de Empleo Público? Si la respuesta es positiva APES501 toma el valor "1", y en caso contrario su valor es "2".

• *Ingreso neto total anual del hogar* (APES502), que se obtiene por imputación a partir de la Encuesta de Presupuestos Familiares (EPF). Esta variable, que en este caso sí es una variable objetivo de EURAREA pero que no está disponible en los censos de población españoles, se define en la EPF como el ingreso neto total debido a la renta anual monetaria del hogar en el año anterior a la entrevista. Para las aplicaciones de EURAREA, se han excluido los ingresos del capital y de la propiedad ya que estas componentes no son adecuadas para las simulaciones. También se han excluido las componentes no monetarias (como el alquiler imputado de las viviendas en propiedad, autoconsumo y autosuministro) debido a la falta de comparabilidad internacional. De esta forma, el ingreso neto total debido a la

renta anual monetaria del hogar, tal como se usa en EURAREA, se refiere a las siguientes componentes:

– Ingresos por trabajo: Sueldos y salarios, ingresos de trabajo por cuenta propia.

– Transferencias personales.

– Transferencias sociales: Pensiones de jubilación y viudedad, otras transferencias sociales, ingresos por desempleo, prestaciones relacionadas con la familia, ingresos por invalidez o enfermedad, asistencia social.

– Otros ingresos.

3. FUENTES COMPLEMENTARIAS

Para la generación de los ficheros APES, se han seleccionado las siguientes fuentes complementarias:

- La Encuesta de Población Activa (EPA).
- La Encuesta de Presupuestos Familiares (EPF).
- El fichero GEO EURAREA.
- También se pueden considerar como fuentes complementarias a APES, para el suministro de información auxiliar agregada en modelos de área, al Registro de Demandantes del INEM y a la Renta Imponible del IRPF agregada en nivel y estructura por origen del ingreso, ambas con datos agregados a nivel de "área pequeña"–EURAREA. Ninguna de estas fuentes, obviamente, está integrada a nivel microdato en APES, sino tan solo enlazada por el código geográfico de área pequeña, y no serán tratadas aquí.

3.1 Encuesta de Población Activa (EPA)

Esta encuesta, referida al segundo trimestre de 1991, se utiliza para proporcionar valores de la variable *"Registro en las oficinas de empleo público*" (APES501). Esta variable figura de forma explícita en el cuestionario de la EPA y se recoge para todas las personas entrevistadas. La imputación de APES501 en la población artificial, se hace usando variables explicativas comunes a ambas fuentes, lo que permite que estén disponibles para cada registro de los ficheros APES. Las publicaciones metodológicas del INE contienen información detallada sobre esta encuesta.

3.2 Encuesta de Presupuestos Familiares (EPF)

Esta encuesta referida al año 1990-91 proporciona valores de la variable "*Ingreso neto total anual del hogar*" (APES502). La imputación a la población artificial se lleva a cabo a nivel de hogar. Debido a que el registro elemental en los ficheros APES (es decir, la persona) incluye, además de su propio identificador, un número de identidad común a todos los miembros del mismo hogar, se ha imputado el valor de la renta familiar a todos los miembros del hogar. En consecuencia, el valor de la renta del hogar estará disponible en todos los registros de los ficheros APES. Las publicaciones metodológicas del INE contienen información detallada sobre esta encuesta.

3.3 El fichero GEO EURAREA

Este archivo es la base para unir los códigos geográficos de los ficheros APES con los códigos geográficos estándar utilizados por el INE y, en particular, en las fuentes seleccionadas (Censo, EPA y EPF). A cierto nivel (NUT III, provincia), los códigos son exactamente los mismos en cada fuente; mientras que, a niveles inferiores, se renumeran las zonas geográficas en APES para prevenir la identificación indirecta.

Los códigos GEO estándar en España son:

- Comunidad Autónoma (NUT II). Hay 17 Comunidades Autónomas, formadas por varias provincias (algunas de ellas, sólo por una provincia).

- Provincia (NUT III). Son 52 en España, de tamaños diversos, formadas por un conjunto de municipios.

- Municipio (sin código NUT oficial). Hay 8.077 en el momento censal en España, con gran variedad de tamaños.

- Distrito. Un conjunto de zonas enumeradas en el Censo dentro de un municipio.

- Sección Censal o Unidad Primaria de Muestreo. Había 31.881 en España en el momento censal. Constituyen las áreas cerradas dentro de un municipio con un máximo de unos 1500 habitantes aproximadamente, aunque su tamaño es muy variable especialmente en zonas rurales. Tienen cierto estatus legal, como es su utilización en los procesos electorales. También es la unidad primaria de muestreo en los diseños polietápicos del INE.

Además de los códigos estándar anteriores se define, en el fichero GEO EURAREA, un código virtual para la unidad primaria de muestreo (Sección Censal) como resultado de renumerar su código original para ser incluido en APES.

Se ha creado un área pequeña, denominada "Comarca-EURAREA", que es la unidad territorial "ad hoc" de nivel NUT IV para la investigación del proyecto EURAREA. Esta división territorial consiste en un área geográfica intermedia entre los niveles de Provincia y Municipio. El INE tradicionalmente utiliza estas áreas como las áreas del control de calidad para la supervisión del trabajo de campo en los Censos, bajo la competencia de un inspector de trabajo de campo.

La formación de las Comarcas-EURAREA se apoyó fundamentalmente en los siguientes criterios:

- No más de 50 municipios.
- No más de 100.000 habitantes.
- No más de 72 unidades elementales de muestreo.

Además de estos criterios, también se tiene en cuenta algún componente estructural (p.e. social, económico).

Por otro lado, hay que destacar que actualmente, algunas de las administraciones regionales están en proceso de aprobar definitivamente las áreas subprovinciales para estadísticas oficiales del nivel NUT IV en España; las cuales, salvo en unas pocas comunidades autónomas, no estaban disponibles en el momento de creación de EURAREA. No obstante, la comarca-EURAREA responde a la misma estructura que las unidades tipo NUT IV en creación aunque no se da una coincidencia exacta y satisface plenamente los objetivos de la investigación pudiendo eventualmente incorporarse al código geográfico de EURAREA.

4. GENERACIÓN DE LAS VARIABLES ARTIFICIALES

4.1 Comentarios generales sobre la modelización de las variables imputadas

En este apartado se describen las dificultades que surgen al ajustar modelos lineales generalizados para predecir APES501 (demandante registrado en una oficina de empleo público) en el fichero de la EPA y para predecir APES502 (total neto de ingresos familiares) en el fichero de la EPF. Las monografías de McCullagh y Nelder (1998), Fahrmeir y Tutz (2001) y Dobson (1990) se han usado como referencia sobre modelos lineales generalizados. Para modelos lineales y modelos lineales mixtos, referencias de interés son Searle (1971), Searle y otros (1992) y Draper (1998).

El fichero de la EPA contiene 199.231 registros con 23 variables APES y el de la EPF contiene 21.155 registros con 21 variables APES. Algunas de las variables son discretas y en la terminología de modelos lineales se las denomina *factores*. Las variables continuas las llamamos *covariables*. Para el caso de factores con *a* niveles, se estiman a-1 parámetros (el parámetro para el último nivel es cero). Sin embargo, para las covariables sólo se estima un parámetro.

El diseño muestral de la EPA y de la EPF españolas es tal que se obtiene la independencia estadística entre Comunidades Autónomas (niveles de la variable APES102). El comportamiento de las variables APES501 y APES502 varía significativamente entre comunidades autónomas. Por lo tanto, ajustar un modelo global para todo el conjunto de datos no es una tarea fácil. No se ha encontrado un buen modelo global y las dificultades computacionales (tiempo de proceso de CPU) no son despreciables. Por tales motivos se ha optado por usar la comunidad autónoma como zona geográfica base para ajustar los modelos.

Se han tenido en cuenta varias opciones antes de elaborar la estrategia final para construir los ficheros de datos APES. Los modelos seleccionados explican alrededor del sesenta por ciento de la variabilidad de APES501 y de APES502. El tamaño de la muestra de la EPF es mucho más pequeño que el de la EPA. Por tal motivo, en el caso de comunidades autónomas uniprovinciales el modelo para APES502 no tiene una proporción adecuada entre el número de parámetros estimados y el tamaño de la muestra. Para aumentar la diferencia entre el tamaño muestral y el número de parámetros a estimar, las comunidades autónomas con una sola provincia no se tratan independientemente. La única excepción es Baleares debido a su carácter de archipiélago.

Los modelos finales, resumidos en las Tablas A2-A3 del Apéndice, para el proceso de imputación de las variables APES501 y APES502 en los ficheros APES, se ajustaron para las siguientes áreas regionales: (1) Andalucía + Ceuta y Melilla, (2) Aragón, (3) Asturias + Cantabria, (4) Baleares, (5) Canarias, (6) Castilla-León, (7) Castilla-La Mancha + Murcia, (8) Cataluña, (9) Valencia, (10) Extremadura, (11) Galicia, (12) Madrid, (13) Navarra + La Rioja, (14) País Vasco.

En los siguientes apartados se muestra cómo usar estos modelos para hacer la imputación de APES501 y APES502 en la provincia de La Rioja. Los trabajos de imputación se iniciaron en La Rioja que, por su tamaño, era la región más adecuada para valorar los problemas metodológicos y computacionales que posiblemente se iban a encontrar. Posteriormente estos trabajos se extendieron a las demás áreas regionales. Obviamente se podría hacer una exposición exhaustiva de los mismos pero dado que se ha aplicado la misma metodología aunque obteniendo diferentes resultados según las características propias de cada área regional, remitimos al lector a las tablas A2-A3 del Anexo mencionadas anteriormente.

4.2 Generación y diagnosis de APES501 en la provincia de La Rioja

APES501 se calcula para todos los individuos cuya edad es mayor o igual a 16 años (16) y menor que 65 años (<65). En esta sección se describe el ajuste de un modelo de regresión logística en el área geográfica (13) Navarra + La Rioja, su utilización en la imputación de APES501 en La Rioja y su validación. En el ajuste del modelo se utilizan, entonces, los registros de la EPA que verifican [APES103=26 o APES103 = 31] y [16 APES203 <65]. En cambio, en la imputación de APES501 se consideran los registros de los ficheros APES que verifican [APES103 = 26] y [16 APES203 <65].

Definimos Y=APES501. Sea Y_j=APES501(j), j=1,2,…, el valor que Y toma en el j-ésimo individuo considerado de La Rioja o Navarra. Obsérvese que Y_j=1 si el j-ésimo individuo considerado afirma (en la EPA) que está registrado en una Oficina de Empleo Público e Y_j=2, en caso contrario. Se consideran las variables aleatorias $X_j=Y_j-1$, con X_j=0 si el j-ésimo individuo considerado afirma estar registrado en una Oficina de Empleo Público y X_j=1, en caso contrario. Estas variables siguen distribuciones de Bernoulli independientes con $E[X_j]=\pi_j$, $\pi_j \in (0,1)$. Es decir, π_j es la probabilidad de que el j-ésimo individuo afirme no estar registrado en una Oficina de Empleo Público. El nexo logit es:

$$\eta_j = g(\pi_j) = \log\left(\frac{\pi_j}{1-\pi_j}\right)$$

Se obtiene el modelo de ecuación:

$$\eta_j = \beta_0 + [APES103(j)] + [APES202(j)] + [APES207(j)] + [APES208(j)] +$$

$$[APES211(j)] + [APES304(j)] + \hat{a}_{APES203}\ APES203(j), \qquad (4.1)$$

donde los parámetros se estiman mediante el software estadístico SPLUS. En la ecuación (4.1), los factores se escriben entre corchetes y las covariables no.

Hay 260.354 registros con APES103 = 26, pero sólo 169.091 registros con APES103 = 26 y 16 APES203 <65. Para todos estos registros, j=1,2,…,169.091, correspondientes a los individuos de La Rioja cuyas edades están comprendidas entre 16 y 64 años, ambos inclusive (16 APES203<65), η_j se calcula aplicando la fórmula (4.1) y los valores de la probabilidad se calculan con la fórmula

$$\pi_j = \frac{\exp\{\eta_j\}}{1+\exp\{\eta_j\}} \qquad (4.2)$$

El paso siguiente es obtener X_j, j=1,2,...,169.091, mediante el generador de números aleatorios de la distribución Bernoulli de parámetro $E[X_j]=\pi_j$. Finalmente, APES501 se calcula para cada individuo como $Y_j=1+X_j$. De esta forma, si $X_j=0$, entonces APES501(j) toma el valor "1" (registrado) y si $X_j=1$, entonces APES501(j) toma el valor "2" (no registrado).

Una vez que las variables Y=APES501 y X=Y-1 han sido imputadas, se realiza un post-análisis por estratos para comprobar si los valores obtenidos para la población son consistentes con los valores originales de la EPA. Para tal fin, se dan las siguientes definiciones:

La media y la varianza poblacional (APES) de X en el i-ésimo estrato son

$$\overline{X}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij} \qquad \text{y} \qquad S_i^2 = \frac{1}{N_i}\sum_{j=1}^{N_i}\left(X_{ij}-\overline{X}_i\right)^2$$

donde N_i (cantidad censal) es el número total de individuos del i-ésimo estrato de La Rioja con edades comprendidas entre 16 y 64 años (ambas incluidas) y X_{ij} es el valor APES de X=APES501-1 del j-ésimo individuo del i-ésimo estrato.

La media y la varianza muestral (EPA) de X en el i-ésimo estrato son

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \qquad \text{y} \qquad s_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}_i\right)^2$$

donde n_i es la cantidad muestral de personas con edades comprendidas entre 16 y 64 años (ambos incluidos) en el i-ésimo estrato de La Rioja y X_{ij} es el valor en la EPA que la variable X = APES501-1 toma en el j-ésimo individuo del i-ésimo estrato.

El intervalo de diagnosis, con factor β, para $\overline{X}_i$ es

$$I_i = \left(\hat{\mu}_i - \beta\sqrt{\frac{S_i^2}{n_i}}, \hat{\mu}_i + \beta\sqrt{\frac{S_i^2}{n_i}}\right)$$

y se espera que $\overline{X}_i \in I_i$.

En la Tabla 4.1 se presentan magnitudes de los ficheros APES y EPA para la provincia de La Rioja.

Tabla 4.1

TAMAÑO, MEDIA Y VARIANZA DE X = APES501-1 POR ESTRATOS EN LA RIOJA

Estrato	*EPA*				*APES*			
	n	*Media*	*1-Media*	*Varianza*	*N*	*Media*	*1-Media*	*Varianza*
1	872	0.916	0.084	0.077	79747	0.889	0.111	0.098
6	229	0.943	0.057	0.054	19851	0.892	0.108	0.096
7	225	0.933	0.067	0.063	19593	0.899	0.101	0.091
8	189	0.937	0.063	0.060	18163	0.900	0.100	0.090
9	314	0.914	0.086	0.079	31737	0.910	0.090	0.082

Los intervalos de diagnosis con factor β=3 y β=1.96 se presentan en la Tabla 4.2.

Tabla 4.2

INTERVALOS DE DIAGNOSIS CON FACTORES $\beta = 3$ Y $\beta = 1.96$ PARA LA RIOJA

Estrato			*Intervalo de diagnosis*			*Intervalo de diagnosis*	
	$\overline{X}_i$	$\overline{X}_i \in I_i$	*con* $\beta = 3$		$\overline{X}_i \in I_i$	*con* $\beta = 1.96$	
1	0.889	SI	0.888	0.944	NO	0.898	0.934
6	0.892	NO	0.897	0.989	NO	0.913	0.973
7	0.899	SI	0.883	0.983	NO	0.900	0.966
8	0.900	SI	0.884	0.990	NO	0.902	0.972
9	0.910	SI	0.866	0.962	SI	0.883	0.945

En la Tabla 4.3 se presentan algunos estadísticos para X=APES501-1 en los ficheros EPA y APES de la provincia de La Rioja. La variable APES2031 (nivel de edad) se obtiene por recodificación de la APES203 (edad), de forma que APES2031=1 si 16≤APES203≤24, APES2031=2 si 25≤APES203≤54, y APES2031=3 si 55≤APES203≤64.

Tabla 4.3

ESTADÍSTICOS DESCRIPTIVOS DE X=APES501-1 EN LOS FICHEROS EPA Y APES DE LA RIOJA

		EPA			*APES*		
APES104 Estrato	*APES 2031 Nivel de edad*	*Media*	*Desviación estándar*	*n*	*Media*	*Desviación estándar*	*N*
1	1	0.84946	0.3586	186	0.8164	0.3872	17136
	2	0.92922	0.2567	551	0.9008	0.2990	50205
	3	0.95556	0.2068	135	0.9433	0.2312	12406
	Total	0.91628	0.2771	872	0.8893	0.3138	79747
6	1	0.88889	0.3187	36	0.8173	0.3865	4291
	2	0.94839	0.2220	155	0.9046	0.2938	12015
	3	0.97368	0.1622	38	0.9402	0.2372	3545
	Total	0.94323	0.2319	229	0.8921	0.3103	19851
7	1	0.91111	0.2878	45	0.8058	0.3957	4124
	2	0.93443	0.2486	122	0.9138	0.2807	11875
	3	0.94828	0.2234	58	0.9558	0.2057	3594
	Total	0.93333	0.2500	225	0.8987	0.3017	19593
8	1	0.94595	0.2292	37	0.8055	0.3959	3881
	2	0.91346	0.2825	104	0.9132	0.2815	10408
	3	0.97917	0.1443	48	0.9572	0.2025	3874
	Total	0.93651	0.2445	189	0.8996	0.3006	18163
9	1	0.75862	0.4317	58	0.8083	0.3936	6016
	2	0.94512	0.2284	164	0.9205	0.2706	17677
	3	0.95652	0.2050	92	0.9616	0.1922	8044
	Total	0.91401	0.2808	314	0.9096	0.2867	31737
Total	1	0.85635	0.3512	362	0.8127	0.3902	35448
	2	0.93339	0.2495	1096	0.9074	0.2898	102180
	3	0.95957	0.1972	371	0.9508	0.2164	31463
	Total	0.92346	0.2659	1829	0.8956	0.3057	169091

4.3 Generación y diagnosis de la variable APES502 en la provincia de La Rioja

En esta sección se describe el ajuste de un modelo de regresión en el área geográfica (13) Navarra + La Rioja, su utilización en la imputación de APES502 (Total neto de ingresos anuales del hogar) en La Rioja y su validación. Esta variable se obtiene de la EPF y contiene información sobre la unidad familiar. Se estudió también la posibilidad de aplicar modelos de regresión múltiple asumiendo la normalidad de la variable objetivo. Sin embargo, los residuos estandarizados de tales modelos poseían una distribución con una clara asimetría a la derecha. Para evitar este problema se decidió modelar el logaritmo de APES502. Para el resto de las áreas geográficas se utilizó el mismo modelo, aunque seleccionando en cada caso el conjunto de variables explicativas más apropiado.

El modelo log-normal que se presenta en este apartado, relaciona APES502 con las características de la Persona de Referencia. La Persona de Referencia se determina mediante la condición APES206=1. En el ajuste del modelo se utilizan los registros de la EPF que verifican [APES103=26 o APES103 = 31] y [APES206=1]. En cambio, en la imputación de APES502 se consideran los registros del fichero APES que verifican APES103 = 26 y APES206=1.

Una aplicación estricta del modelo lineal ajustado permite imputar sólo el valor de APES502 en los casos en que APES206=1. Sin embargo, para completar la variable APES502 en los ficheros APES, a todos los miembros de un hogar se imputa el mismo valor que el de su Persona de Referencia. Esto significa que todos los registros que en los ficheros APES tienen el mismo valor en las variables APES103, APES104, APES106, APES401 y APES402, también tendrán el mismo valor de APES502. En los ficheros APES el hogar está definido por el siguiente conjunto de variables: APES103 (provincia), APES104 (estrato), APES106 (unidad principal de muestreo dentro del estrato), APES401 (número de domicilio de dentro de la unidad principal de muestreo) y APES402 (número de hogar dentro de la vivienda).

Definimos Y=APES502, por lo que Y_j=APES502(j), para j=1,2,…, denota los valores que toma Y en el j-ésimo individuo de La Rioja o Navarra. Se asume que X_j =log(Y_j) son variables aleatorias independientes con $E[X_j]=\mu_j$ y $V[X_j]=\sigma^2/w_i$; es decir, los w_i son los pesos muestrales de la EPF que dentro de cada estrato son constantes. La varianza σ^2 se toma del error cuadrático medio de la Tabla ANOVA asociada al modelo ajustado. La media μ_j se calcula aplicando la ecuación del modelo

$$\mu_j = \hat{a}_0 + [APES403(j)] + [APES405(j)] + [APES410(j)]$$

$$+ [APES303*APES306(j)] + \beta_{APES409}\ APES409(j), \qquad (4.3)$$

donde los parámetros se han estimado usando el software estadístico SPSS. En la ecuación (4.3), los factores se escriben entre corchetes y las covariables no.

Para todos los registros, j=1,2,...,85.086, correspondientes a los individuos con APES103=26 y APES206=1, μ_j se calcula aplicando la fórmula (4.3). Así, tenemos todos los elementos que necesitamos para generar el valor de la distribución teórica de $X_j=\log(Y_j)$; es decir, para simular de una distribución normal con media $E[X_j]=\mu_j$ y varianza $V[X_j]=\sigma^2/w_i$.

Si se dispusiera de una estructura geográfica más detallada en los datos de la EPF española (por ejemplo, se incluyeran las comarcas como variables GEO), sería más apropiado un modelo multinivel. Así, el modelo lineal binivel con efectos aleatorios en las comarcas es apropiado para la utilización de estimadores EBLUP (véase Rao(2003)) en las simulaciones que se realicen con el fichero APES. Sin embargo, los citados datos geográficos no están disponibles y el modelo log-lineal (4.3) se considera aceptablemente bueno para la imputación de la variable APES502. Dado que los ficheros APES se construyen con el objetivo de realizar simulaciones para verificar el comportamiento de estimadores de áreas pequeñas, es conveniente introducir un poco de variabilidad entre comarcas. Sin embargo, en el modelo para log(APES502), se tiene la misma ecuación de regresión para cada comarca.

Por ello, en lugar de contar sólo con la variación debida a los individuos, se puede introducir alguna variación entre comarcas simulando condiciones aleatorias a ese nivel de agregación. Esto se consigue descomponiendo la estimación de la varianza σ^2 en dos partes, de modo que (1-p)% de la varianza se asigne al nivel individual y p% al nivel comarcal. Los efectos aleatorios de la comarca se generan con distribución normal de media 0 y varianza $p\sigma^2$, es decir, $U_d \sim N(0,p\sigma^2)$. Para cada individuo se genera una perturbación normal $V_{dj} \sim N(0 ,(1-p)\sigma^2)$. Finalmente, se aplica la fórmula

$$X_{dj} = \log Y_{dj} = \mu_{dj} + (w_{dj})^{1/2}\, U_d + (w_{dj})^{1/2}\, V_{dj}\ .$$

Como resultado de este proceso, el logaritmo de los ingresos tiene correlación $\rho=p$ dentro de la comarca. Los valores de APES502 se imputan mediante la fórmula $Y_{dj} = \exp(\ X_{dj}\ )$.

Después de haber imputado APES502 a la persona de referencia, el mismo valor se asigna a todas aquellas personas que pertenecen a su hogar. Para la selec-

ción del *p%* de variabilidad asignable al nivel comarcal se efectuaron las siguientes consideraciones:

1. El valor de *p* debe ser inferior a 0.1 (10%) con objeto de no desviarse demasiado del modelo (4.3) ajustado en la muestra EPF.

2. En los ficheros APES (y no en los EPF) de las zonas geográficas (1)-(14) se ajustaron modelos ANOVA unifactoriales con efectos fijos en las comarcas-EURAREA para evaluar la variabilidad entre ellas. El análisis comparativo de las tablas ANOVA permitió asignar los valores $0<p<0.1$ a las distintas zonas. En concreto, para La Rioja se seleccionó el valor 0.096.

Finalmente se realiza un post-análisis para comprobar si los valores de X=log(APES502) obtenidos para la población son consistentes con los valores originales de la EPF. En la Tabla 4.4 se presentan magnitudes de los ficheros APES y EPF para la provincia de La Rioja. Los intervalos de diagnosis con factor β=3 y β=1.96 se presentan en la Tabla 4.5.

Tabla 4.4

TAMAÑO, MEDIA Y VARIANZA DE X EN EL ESTRATO i DE LA RIOJA

	EPF			*APES*		
Estrato	*n*	*Media*	*Varianza*	*N*	*Media*	*Varianza*
1	178	14.416	0.568	38990	14.347	0.290
6	46	14.245	0.257	9775	14.259	0.286
7	43	14.239	0.277	9660	14.222	0.296
8	40	14.352	0.443	9397	14.142	0.283
9	50	14.329	0.423	17264	14.076	0.302

Tabla 4.5

INTERVALOS DE DIAGNOSIS CON FACTOR β=1.96 Y β=3 PARA LA RIOJA

Estrato			*Intervalo de diagnosis*			*Intervalo de diagnosis*	
	$\overline{\chi}_i$	$\overline{\chi}_i \in I_i$	*con* $\beta = 3$		$\overline{\chi}_i \in I_i$	*con* $\beta = 1.96$	
1	14.347	SI	14.246	14.585	SI	14.305	14.527
6	14.259	SI	14.021	14.469	SI	14.098	14.391
7	14.222	SI	13.998	14.480	SI	14.082	14.396
8	14.142	SI	14.036	14.668	NO	14.146	14.558
9	14.076	SI	14.053	14.605	NO	14.149	14.509

En la Tabla 4.6 se presentan algunos estadísticos de X=log(APES502) en los ficheros EPF y APES. La variable APES4051 (nivel de ocupación de la familia), se obtiene por recodificación de APES405 (número de personas ocupadas) de la siguiente forma APES4051=0 si APES405=0, APES4051=1 si APES405=1, APES4051=2 si APES405=2, y APES4051=3 si APES405≥3.

Tabla 4.6

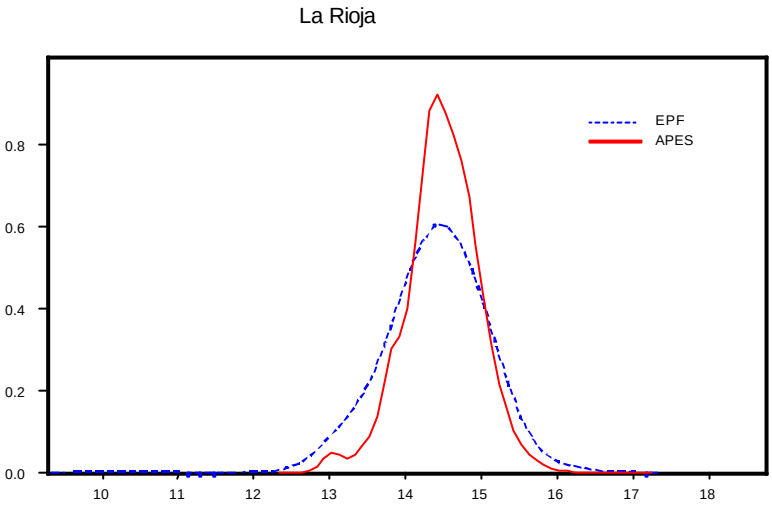
ESTADÍSTICOS DESCRIPTIVOS DE X=LOG (APES502) EN LOS FICHEROS EPF Y APES DE LA RIOJA

		EPF			*APES*		
APES104 Estrato	*APES4051 Nivel de ocupación*	*Media*	*Desviación estándar*	*n*	*Media*	*Desviación estándar*	*N*
1	0	13.764	0.555	39	13.701	0.407	10269
	1	14.383	0.759	82	14.382	0.287	16518
	2	14.855	0.380	39	14.787	0.254	10087
	3	15.026	0.588	18	15.098	0.280	2116
	Total	14.416	0.754	178	14.347	0.539	38990
6	0	13.912	0.563	13	13.631	0.375	2809
	1	14.287	0.356	18	14.334	0.275	4064
	2	14.479	0.472	13	14.704	0.254	2364
	3	14.503	0.793	2	15.009	0.266	538
	Total	14.245	0.507	46	14.259	0.535	9775
7	0	13.682	0.354	14	13.597	0.391	2740
	1	14.399	0.334	21	14.300	0.287	4447
	2	14.725	0.210	5	14.712	0.261	1916
	3	14.909	0.316	3	14.997	0.294	557
	Total	14.239	0.526	43	14.222	0.544	9660
8	0	13.689	0.327	13	13.549	0.355	2995
	1	14.212	0.272	11	14.240	0.251	3888
	2	14.998	0.638	7	14.636	0.228	1957
	3	14.978	0.230	9	14.918	0.225	557
	Total	14.352	0.666	40	14.142	0.532	9397
9	0	13.734	0.467	15	13.522	0.353	6225
	1	14.374	0.471	22	14.199	0.264	6869
	2	14.936	0.552	10	14.615	0.214	2930
	3	14.949	0.215	3	14.907	0.256	1240
	Total	14.329	0.651	50	14.076	0.549	17264
Total	0	13.757	0.486	94	13.619	0.390	25038
	1	14.361	0.610	154	14.316	0.287	35786
	2	14.805	0.460	74	14.728	0.256	19254
	3	14.967	0.477	35	15.010	0.281	5008
	Total	14.353	0.678	357	14.245	0.551	85086

En la Figura 4.1 se presentan los estimadores kernel con núcleo normal de la función de densidad de la variable aleatoria *log(APES502)* de los ficheros APES y EPF.

Figura 4.1

ESTIMADORES KERNEL PARA LOG (APES502) EN APES Y EPF DE LA RIOJA



4.4. Consideraciones de carácter computacional

En este apartado se describen algunos aspectos computacionales referentes a la asignación de la variable APES502 para los registros correspondientes al mismo hogar. Según se ha comentado previamente, cuando se imputa el valor de APES502 a la persona de referencia se puede obtener de esta persona el valor de las variables APES103, APES104, APES106, APES401 y APES402 y buscar todas aquellas personas que tengan los mismos valores en dichas variables. Esto significa que cada vez que se imputa una persona de referencia se tiene que recorrer todo el fichero APES de la zona geográfica considerada para buscar aquellas personas con los mismos valores en esas variables. Desde un punto de vista computacional el coste de este procedimiento es n^2, donde n es el número de registros del fichero. Se puede mejorar el algoritmo parando la búsqueda cuando se haya encontrado a todos los miembros del hogar, lo que implica grabar también en

los ficheros APES la variable APES409 (tamaño del hogar) para cada persona de referencia. Sin embargo, el coste asintótico continúa siendo n^2, ya que esta mejora divide el coste de ejecución por una constante positiva y, por ello, no modifica el coste asintótico. Aplicando esta mejora al fichero APES restringido a la provincia de La Rioja se estimó que el tiempo de ejecución sería de un mes en un PC con procesador Pentium III a 500 Mhz.

El algoritmo anterior puede ser mucho más rápido si el fichero se ordena previamente. En este caso el coste asintótico de la asignación de la variable APES502 es n. No obstante, el coste asintótico de ordenar el fichero es n^2 sin ninguna constante que disminuya el tiempo de cálculo. En el caso anterior, el coste disminuye por el hecho de no recorrer el fichero entero, ya que cuando se encuentran todos los miembros del hogar se detiene la búsqueda. El coste también disminuye porque el proceso se realiza solamente para las personas de referencia (40% de la población). Para realizar la ordenación del fichero APES de la provincia de La Rioja (n=263.354), el coste estimado es prohibitivo en un PC con las características de hardware antes mencionadas.

Una solución alternativa es construir una base de datos con los ficheros APES y crear un fichero índice con las variables que definen el hogar (APES103, APES104, APES106, APES401 y APES402). Como en el fichero índice cada registro debe ser único, incluimos APES206 (relación con la persona de referencia) para tener la persona de referencia en primer lugar y detrás todos los miembros del hogar. En el caso de una familia con más de un hijo, el valor de APES206 para los hijos es igual a 3, por lo que continúa repitiéndose. Para asegurarnos que el registro es único incluimos la variable APES201 (número de orden de la persona en el hogar).

De esta forma, en la base de datos se define una tabla con todas las variables de los ficheros APES y, como índice de esta tabla, las variables que se han enumerado en el párrafo anterior. Para pasar los ficheros de texto a bases de datos, se ha desarrollado un programa en lenguaje C++ que lee los ficheros APES y para cada línea del fichero guarda el correspondiente registro en la base de datos. Cada vez que se introduce un nuevo registro se actualiza automáticamente el fichero de índices, por lo que podemos decir, que la base de datos ordena cada nuevo registro introducido en relación con los que ya han sido introducidos. Este procedimiento se ha ejecutado con el fichero APES de La Rioja en un PC con las características mencionadas anteriormente.

Una vez ordenado un fichero APES dado, cada vez que encontramos una persona de referencia calculamos e imputamos el valor de la variable APES502. El mismo valor también se asigna a todos los miembros del hogar (que están exactamente después de la persona de referencia al estar ahora el fichero APES ordena-

do). Este proceso se llevó a cabo con el fichero APES de La Rioja con un PC con las especificaciones hardware descritas anteriormente y tardó apenas tres minutos.

5. CONCLUSIONES

5.1. Sobre los ficheros de datos APES

La población artificial española, generada por el Instituto Nacional de Estadística (INE) y la Universidad Miguel Hernández (UMH) para el proyecto EURAREA, es un conjunto de 14 ficheros de datos, correspondiente a cada una de las 14 regiones enunciadas en la Sección 4.1. El fichero único APES debería ser construido uniendo este conjunto de ficheros. Si se hiciera así tendríamos 38.872.268 registros, 40 variables por registro y un tamaño total aproximado de 3.5 Gigabytes (GB), y se presentan limitaciones con respecto al tamaño máximo de 2 GB de los ficheros en el sistema operativo MS Windows.

Por ello se ha optado por la división según regiones. Los ficheros APES toman 35 variables del fichero del censo español de 1991. Las variables APES501 y APES502 están imputadas utilizando la Encuesta de Población Activa (EPA) referida al segundo trimestre de 1991 y la Encuesta de Presupuestos Familiares 1990-91 (EPF), respectivamente. Las tres últimas variables se obtienen mediante transformación de algunas de las variables anteriores.

Para el proceso de imputación se ha usado el "método de regresión", es decir, se ha ajustado un modelo de regresión para calcular los datos. Este método une las variables de respuesta a las covariables del conjunto de datos de la población. Se han ajustado 14 modelos de regresión logística con los datos de la EPA y 14 modelos de regresión lognormal con los datos EPF.

En relación con el proceso de imputación, el indicador del área pequeña de los ficheros APES relativo a la comarca (equivalente al nivel NUT IV) no aparece en los ficheros de encuesta. Por este motivo no se usaron modelos multinivel, sino modelos individuales de efectos fijos que no permiten introducir variación entre comarcas. Para resolver este problema se ha desarrollado y aplicado un método para introducir alguna variación entre comarcas en la variable log(APES502). Alrededor del 10% de la varianza que el modelo ajustado asigna al nivel individual fue transferido al nivel comarcal. Por esta razón log(APES502) tiene una varianza individual menor en los ficheros APES que en el fichero EPF.

5.2. Sobre el post-análisis

Para realizar un post-análisis con la finalidad de validar las variables imputadas (APES501 y APES502) en los 14 ficheros de APES se han usado tres métodos: (1) Intervalos de Diagnosis, (2) Comparación de datos (APES contra EPA o EPF), (3) Comparación de los estimadores kernel no paramétricos de las funciones de densidad de log(APES502) en APES y EPF.

Aunque APES501 y APES502 no están idénticamente distribuidas a lo largo de los registros, se han utilizado las fórmulas estándar de los intervalos de confianza para calcular los Intervalos de Diagnosis. Teniendo en cuenta que la población artificial solamente es una herramienta para hacer simulaciones, la principal finalidad del estudio de validación es detectar errores y grandes discrepancias entre los datos de las encuestas y los imputados. Como resultado de este estudio se han ajustado nuevos modelos de regresión para APES501 en Valencia y Asturias–Cantabria. También se han ajustado nuevos modelos para APES502 en Canarias y Madrid.

Los Intervalos de Diagnosis y la comparación de datos muestran que la variable imputada APES501 es bastante similar a la variable APES501 de la EPA

Con referencia a la imputación de log(APES502), el principal problema consiste en la reducción de la varianza con respecto a la existente en el fichero de la EPF. Esto es debido a la creación de una variabilidad artificial a nivel de comarca. Por esta razón es necesario ser cautelosos en la interpretación de los resultados de futuras simulaciones, ya que el comportamiento de los estimadores de la media de APES502 en áreas pequeñas será previsiblemente mejor en el "mundo APES" que en el mundo real. No obstante, este hecho no es problemático si estamos principalmente interesados en la comparación de diferentes estimadores.

Otra razón para obtener una mejor diagnosis de APES501 que de APES502 es que los tamaños muestrales en la EPA son mucho más grandes que en la EPF, lo cual permite, con el mismo grado de robustez, ajustar modelos con un número mayor de variables auxiliares.

5.3. Aplicaciones del fichero APES

La generación del fichero APES se realizó durante el primer año (2001) del proyecto EURAREA. En los años posteriores (2002-2003) APES se utilizó como población artificial para la ejecución de experimentos de simulación en los que se evaluaron distintos procedimientos de estimación en áreas pequeñas. En los citados experimentos se consideraron:

1. Los tres parámetros objetivo descritos en la Sección 1: proporción de la población económicamente activa que está desempleada, proporción de hogares unipersonales e ingresos por unidad de consumo.

2. Diseños muestrales complejos; en especial, los diseños de la EPA y de la EPF.

3. Estimadores de medias y proporciones directos, asistidos por modelos y basados en modelos.

4. Sesgo y error cuadrático medio como medidas de eficiencia en las reiteraciones muestrales.

La construcción de la población artificial APES ha permitido realizar simulaciones altamente realistas en el proyecto EURAREA. Los resultados del proyecto y sus conclusiones se presentarán en una conferencia final de proyecto.

REFERENCIAS

DOBSON, A.J. (1990). «An Introduction to Generalized Linear Models». Chapman and Hall.

DRAPER, N.R. (1998). «Applied regression analysis», John Wiley & Sons.

FAHRMEIR, L. y TUTZ, G. (2001). «Multivariate Statistical Modelling Based on generalized Linear Models». Springer-Verlag.

MCCULLAGH P. y NELDER J.A. (1998). «Generalizad Linear Models». Chapman and Hall.

RAO, J.N.K. (2003). «Small Area Estimation». John Wiley.

SEARLE, S.R. (1971). «Linear Models». John Wiley.

SEARLE, S.R., CASELLA, G. y MCCULLOGH, C.E. (1992). «Variance Components». John Wiley.

«Encuesta de Población Activa. Informe Técnico. Publicaciones del Instituto Nacional de Estadística.

«Encuesta de Presupuestos Familiares *1990-91»*. Metodología. Publicaciones del Instituto Nacional de Estadística.

«Renta, pobreza y exclusión social». Publicación de estadística social de EUROSTAT – 2001.

«Censo de Población 1991. Metodología». Publicaciones del Instituto Nacional de Estadística.

APÉNDICE

Tabla A.1

POBLACIÓN ARTIFICIAL EURAREA, ESPAÑA. APES (Continúa)

Variable	*Posición*	*Nombre y descripción*	*Valores*	*Disponible*
Características geográficas				
APES101		*País*	Es	EPA, EPF
APES102	1-2	*Comunidad Autónoma*		EPA, EPF
		Andalucía	01	
		Aragón	02	
		Asturias	03	
		Baleares	04	
		Canarias	05	
		Cantabria	06	
		Castilla León	07	
		Castilla La Mancha	08	
		Cataluña	09	
		Valencia	10	
		Extremadura	11	
		Galicia	12	
		Madrid	13	
		Murcia	14	
		Navarra	15	
		País Vasco	16	
		La Rioja	17	
		Ceuta Y Melilla	18	
APES103	3-4	*Provincia*		EPA, EPF
		Álava	01	
		Albacete	02	
		Alicante	03	
		Almería	04	
		Ávila	05	
		Badajoz	06	
		Baleares	07	
		Barcelona	08	

Tabla A.1

POBLACIÓN ARTIFICIAL EURAREA, ESPAÑA. APES (Continuación)

Burgos	09
Cáceres	10
Cádiz	11
Castellón	12
Ciudad Real	13
Córdoba	14
Coruña, La	15
Cuenca	16
Gerona	17
Granada	18
Guadalajara	19
Guipúzcoa	20
Huelva	21
Huesca	22
Jaén	23
León	24
Lérida	25
La Rioja	26
Lugo	27
Madrid	28
Málaga	29
Murcia	30
Navarra	31
Orense	32
Asturias (Oviedo)	33
Palencia	34
Palmas Las	35
Pontevedra	36
Salamanca	37
Santa Cruz de Tenerife	38
Cantabria (Santander)	39
Segovia	40
Sevilla	41
Soria	42
Tarragona	43
Teruel	44
Toledo	45
Valencia	46
Valladolid	47
Vizcaya	48

Tabla A.1

POBLACIÓN ARTIFICIAL EURAREA, ESPAÑA. APES (Continuación)

		Zaragoza	50	
		Ceuta	51	
		Melilla	52	
APES104	5	*Estrato*		EPA, EPF
		capital de provincia	1	
		municipios auto representados que son importantes en relación con la capital	2	
		municipios auto representados que son importantes en relación con la capital o que tienen más de 100.000 habitantes	3	
		municipios de 50.000 a 99.999 hab.	4	
		municipios de 20.000 a 49.999 hab.	5	
		municipios de 10.000 a 19.999 hab.	6	
		municipios de 5.000 a 9.999 hab.	7	
		municipios de 2.000 a 4.999 hab.	8	
		municipios de menos de 2000 hab.	9	
APES105		*Area Pequeña*		GEO
	6-7	Comarca	01-99	
	8-9	Zona	01-99	
APES106	10-13	*Unidad elemental de muestreo (UEM)*	0001-9999	GEOS
Características personales				
APES201	22-25	*Nº de orden de la persona en el hogar*		
		persona uno	01	
		...	...	
		persona treinta y cinco	35	
APES202	26	*Sexo*		EPA
		Varón	1	
		Mujer	6	
APES203	27-29	*Edad*		EPA
		0 años	000	
		1 año	001	
		...	...	
		120 años	120	
APES206	30-31	*Relación con la persona de referencia*		EPA
		Persona de referencia	01	
		Cónyuge o pareja	02	
		Hijo o hija	03	
		Yerno, nuera	04	
		Nieto/a	05	

Tabla A.1

POBLACIÓN ARTIFICIAL EURAREA, ESPAÑA. APES (Continuación)

		Padre o madre	06	
		Otro parentesco	07	
		Servicio doméstico	08	
		No emparentado	09	
APES207	32	*Estudios de más alto nivel completados*		EPA
		Menor de 10 años	B	
		Analfabetos	1	
		Sin estudios	2	
		Primaria	3	
		EGB o bachiller elemental	4	
		Formación profesional primera etapa	5	
		Formación profesional segunda etapa	6	
		BUP, bachiller y COU	7	
		Diplomado y equivalentes	8	
		Licenciado y equivalentes	9	
APES208	33	*Relación con la actividad*		EPA
		Ocupado	1	
		Parado	2	
		Inactivo	3	
		Servicio militar o equivalente	4	
		Menor de 16 años	5	
APES210	34	*Situación profesional*		EPA
		Profesional o empresario sin asalariados	1	
		Empleador	2	
		Asalariado	3	
		Ayuda familiar	4	
		No aplicable	5	
APES211	35-36	*Condición socioeconómica*		EPA
		Empresarios agrarios con asalariados	01	
		Empresarios agrarios sin asalariados	02	
		Miembros de cooperativas agrarias	03	
		Directores y jefes de explotaciones agrarias	04	
		Resto de trabajadores agrarios	05	
		Profesionales, técnicos y asimilados que ejercen su actividad por cuenta propia, con o sin asalariados	06	
		Empresarios no agrarios con asalariados	07	
		Empresarios no agrarios sin asalariados	08	
		Miembros de cooperativas no agrarias	09	

Tabla A.1

POBLACIÓN ARTIFICIAL EURAREA, ESPAÑA. APES (Continuación)

		Directores y gerentes de establecimientos no agrarios, altos funcionarios de la Admón. Pública, CC.AA. y corporaciones locales	10	
		Profesionales, técnicos y asimilados que ejercen su actividad por cuenta ajena	11	
		Profesionales en ocupaciones exclusivas de la administración pública	12	
		Resto del personal administrativo y comercial	13	
		Resto del personal de los servicios	14	
		Contramaestres y capataces de establecimientos no agrarios	15	
		Operarios cualificados y especializados de establecimientos no agrarios	16	
		Operarios sin especialización de establecimientos no agrarios	17	
		Profesionales de las fuerzas armadas	18	
		No clasificables por condición socioeconómica	19	
Características de la persona de referencia				
APES301	37	*Sexo*	1,6	EPF,EPA
APES302	38-40	*Edad*	001-120	EPF,EPA
APES303	41	*Estudios de más alto nivel completados*	B,1-9	EPF,EPA
APES304	42	*Relación con la actividad*	1-5	EPF,EPA
APES306	43	*Situación profesional*	1-5	EPF,EPA
APES307	44-45	*Condición socioeconómica*	01-19	EPF,EPA
Características del hogar				
APES401	14-19	*Nº de la vivienda en la sección*	000010 – 159999	-
APES402	21	*Nº del hogar dentro de la vivienda*		-
		primer hogar	1	
		...	...	
		octavo hogar	8	
		hogar transitorio	0	
APES403	46-47	*Tipo de hogar*		EPA,EPF
		Unipersonal, varón de 15 a 64 años	01	
		Unipersonal, mujer de 15 a 64 años	02	
		Unipersonal, varón mayor de 64 años	03	
		Unipersonal, mujer mayor de 64 años	04	
		Dos adultos de 15 a 64 años	05	
		Dos adultos, al menos uno mayor de 64	06	

Tabla A.1

POBLACIÓN ARTIFICIAL EURAREA, ESPAÑA. APES (Continuación)

		Un adulto varón, con uno o más niños de menos de 15 años	07	
		Un adulto mujer, con uno o más niños de menos de 15 años	08	
		Dos adultos con un niño<15	09	
		Dos adultos con dos niños<15	10	
		Dos adultos con tres niños <15	11	
		Dos adultos con cuatro o más niños <15	12	
		Tres o más adultos con uno o más niños <15	13	
		Tres o más adultos sin menores de 15	14	
		Todos los miembros son menores de 15	15	
APES404	48-49	*Número de activos en el hogar*	00-35	EPA,EPF
APES405	50-51	*Número de personas ocupadas*	00-35	EPA,EPF
APES406	52-53	*Número de personas paradas*	00-35	EPA,EPF
APES407	54-55	*Número de personas menores de 16 años*	00-35	EPA,EPF
APES408	56-57	*Número de personas mayores de 64 años*	00-35	EPF
APES409	58-59	*Tamaño del hogar*	00-35	EPA,EPF
APES410	60	*Calefacción*		EPF
		Calefacción colectiva	1	
		Calefacción individual	2	
		Sólo algún aparato fijo o móvil para calentar	3	
		No tiene calefacción	4	
		Transeúntes	B	
APES411	61	*Refrigeración de la vivienda*		EPF
		Si	1	
		No	2	
		Transeúntes	B	
APES412	62-65	*Superficie útil de la vivienda(m^2)*	0000-9999	EPF
APES413	66	*Régimen de tenencia de la vivienda*		EPF
		Propiedad por compra totalmente pagada	1	
		Propiedad por compra con pagos pendientes	2	
		Propiedad por herencia	3	
		Facilitada gratuita o semigratuitamente	4	
		En alquiler	5	
		Otras formas	6	
		hostal, posada	B	
APES414	67-70	*Año de construcción*	0000-9999	EPF

Tabla A.1

POBLACIÓN ARTIFICIAL EURAREA, ESPAÑA. APES (Conclusión)

APES415	71	*Tipo de residencia*		EPF
		Residencia principal	1	
		Residencia estable	6	
		Residencia provisional	7	
APES416	20	*Número de habitantes*		
		Un habitante	1	
		….	…	
		Ocho habitantes	8	
		Transeúntes	0	
Variables imputadas				
APES501	72	Registrado en una oficina de empleo público	1,2	EPA
		Registrado	1	
		No registrado	2	
APES502	73-80	*Total de ingresos del hogar*	00000000-99999999	EPF
Variables calculadas				
APES503	81	*Desempleo ILO*		
		Si	1	
		No	0	
APES504	82	*Hogar Unipersonal*		
		Si	1	
		No	0	
APES505	83-87	*Nº total de miembros familiares normalizado*	00-35	

Tabla A.2

FACTORES Y COVARIABLES, DEL FICHERO EPA DE 1991, QUE APARECEN EN LOS MODELOS LOGIT DE APES501

			Factores														*Covariables*			
Comunidades Autónomas	**n**	**(%)**	103	104	202	206	207	208	210	211	301	303	304	306	307	403	203	405	409	**pará-metros**
Andalucía, Ceuta y Melilla	23.016	49.88	X	X	X		X	X	X	X		X	X	X	X		X		X	88
Aragón	5.508	57.20	X		X		X	X	X	X		X	X				X		X	50
Asturias y Cantabria	6.735	63.23	X	X	X	X	X	X		X					X	X	X		X	76
Baleares	2.466	53.35		X			X	X	X	X		X				X	X		X	57
Canarias	5.912	50.26	X		X		X	X	X				X				X			21
Castilla-León	12.595	61.99	X		X	X	X	X	X	X							X			52
Castilla-La Mancha y Murcia	12.063	58.40	X	X	X	X	X	X	X				X				X			40
Cataluña	13.306	72.16	X	X	X	X	X	X		X			X				X			53
Valencia	10.783	56.57	X		X		X	X	X	X	X		X	X			X		X	47
Extremadura	4.994	46.96	X	X	X	X	X	X	X	X							X	X		52
Galicia	8.591	60.10	X	X	X		X	X	X	X	X						X			47
Madrid	6.284	77.17				X	X	X								X	X			31
Navarra y La Rioja	4.722	53.28	X		X		X	X		X			X				X			35
País Vasco	7.349	60.43	X		X	X	X	X	X	X	X						X			47

Factores		*Covariales*	
APES 103	Provincia	APES 203	Edad
APES 104	Estrato	APES405	Número de personas ocupadas
APES 202	Sexo	APES 409	Tamaño del Hogar
APES 206	Relación con la persona de referencia		
APES 207	Estudios de más alto nivel completados		
APES 208	Relación con la actividad		
APES 210	Situación profesional		
APES 211	Condición socioeconómica		
APES 301	Sexo de la Persona de Referencia (PR)		
APES 303	Estudios de más alto nivel completados por PR		
APES 304	Relación con la actividad de PR		
APES 306	Situación profesional de PR		
APES 307	Condición socioeconómica de PR		
APES 403	Tipo de Hogar		

Tabla A.3

FACTORES Y COVARIABLES, DEL FICHERO EPA DE 1991, QUE APARECEN EN LOS MODELOS LOGNORMALES DE APES503

			Factores															*Covariables*			
Comunidades Autónomas	**n**	**R^2**	103	104	301	303	304	306	403	404	405	406	407	408	410	411	413	302	409	412	**parámetros**
Andalucía, Ceuta y Melilla	3.895	0.580	X	X		X	X	X	X		X	X	X	X	X	X	X		X	X	79
Aragón	1.105	0.702	X	X		X		X	X	X	X	X		X	X	X	X			X	55
Asturias y Cantabria	805	0.584		X	X	X	X	X	X		X	X				X		X	X	X	48
Baleares	429	0.641			X	X			X		X	X			X			X			34
Canarias	771	0.575	X		X	X			X	X	X					X		X		X	60
Castilla-León	3.157	0.625	X	X		X	X	X	X		X				X				X	X	51
Castilla-La Mancha y Murcia	2.220	0.608	X	X		X		X	X		X				X				X		46
Cataluña	1.642	0.645		X		X	X	X	X		X				X	X	X		X	X	52
Valencia	1.706	0.589	X			X	X	X	X	X			X	X		X		X	X		50
Extremadura	829	0.510		X		X		X	X		X						X	X	X		42
Galicia	829	0.550	X	X		X		X	X		X				X				X	X	44
Madrid	762	0.605		X		X	X	X	X		X								X	X	38
Navarra y La Rioja	724	0.612				X		X	X		X				X				X		54
País Vasco	1.694	0.594		X	X	X	X		X		X				X					X	41

Factores		*Covariables*	
APES 103	Provincia	APES 302	Edad de PR
APES 104	Estrato	APES 409	Tamaño del Hogar
APES 301	Sexo de la Persona de Referencia (PR)	APES 412	Superficie útil de la vivienda (m^2)
APES 303	Estudios de más alto nivel completados por PR		
APES 304	Relación con la actividad de PR		
APES 306	Situación profesional de PR		
APES 403	Tipo de Hogar		
APES 405	Número de personas ocupadas		
APES 406	Número de personas paradas		
APES 407	Número de personas menores de 16 años		
APES 408	Número de personas mayores de 64 años		
APES 410	Calefacción		
APES 411	Aire acondicionado		
APES 413	Régimen de tenencias de la vivienda		

CONSTRUCTION OF AN ARTIFICIAL POPULATION BASED ON THE SPANISH POPULATION CENSUS OF 1991

ABSTRACT

In this work we describe the construction of a data file representing the Spanish population and based on the Spanish population census of 1991. The mentioned file, called APES, was built in the framework of the European project EURAREA (Enhancing Small Area Estimation Techniques to Meet European Needs, IST-2000-5.1.8, 2001-2003) in order to simulate in a realistic way sampling designs of Spanish official Statistics surveys (like Labour Force or Family Budget surveys) and to evaluate small area estimation procedures.

Key words: Artificial universe, imputation, generalised linear models, popularion census, Labour Force Survey, Family Budget Survey.

AMS Classification: 62E30, 62J12.