Proyecto EURAREA
(Enhancing Small Area Estimation Techniques to meet European Needs)

Diciembre, 2005

1 Introducción

El proyecto EURAREA (Enhancing Small Area Estimation Techniques to meet European needs) ha sido realizado dentro del marco del 5º Programa de I+D desarrollado por la Unión Europea en los años 2000 - 2004, y en él ha participado España junto con otros 6 países europeos (Reino Unido, Italia, Suecia, Noruega, Finlandia y Polonia).

El objetivo de este proyecto es evaluar la eficiencia de los estimadores estándar para áreas pequeñas (sintéticos, GREGs y combinados). Los estudios realizados hasta ahora estuvieron basados en diseños muestrales con probabilidades iguales de selección. Por ello ha sido necesario llevar a cabo un estudio de la teoría existente así como desarrollar nuevas teorías que faciliten la obtención del estimador y de su error cuadrático medio cuando se utilizan otros esquemas muestrales más similares a los aplicados en las encuestas oficiales en el mundo real. Finalmente toda la teoría desarrollada ha sido implementada en una aplicación informática programada en SAS cuyo manejo ha sido ampliamente documentado para que cualquier usuario pueda aplicarlo a sus propios datos.

El proyecto concentra la investigación principalmente en cuatro temas:

1) El uso de información auxiliar procedente del pasado.
2) El uso de información auxiliar procedente de otras zonas geográficas.
3) La adaptación de los estimadores estándar a diseños muestrales complejos, es decir, con uso de probabilidades desiguales y, en particular, con selección de conglomerados.
4) La obtención de estimaciones para clasificaciones cruzadas.

Desde comienzos del 2001 el Instituto Nacional de Estadística (INE) ha participado en colaboración con la Universidad Miguel Hernández de Elche (UMH) en el desarrollo del tema 3. Por lo tanto, este documento se centra en la descripción de los trabajos realizados bajo esta perspectiva y, más concretamente, bajo el objetivo de estudiar el impacto de los pesos muestrales en los estimadores de áreas pequeñas.

Este documento está estructurado en 13 secciones con sus correspondientes subsecciones en algunos casos. En la sección 2 se describe el proceso seguido para la creación de una población artificial sobre la que llevar a cabo diferentes experimentos de simulación que nos permitan evaluar los estimadores en las pequeñas áreas. En la sección 3 se dan los conceptos aplicados de área pequeña. En la sección 4 se ilustran las fuentes complementarias a la población artificial creada que proporcionan información auxiliar agregada a nivel de área. En las secciones siguientes 5, 6 y 7 se definen los parámetros poblacionales objeto de estudio, los diseños muestrales aplicados para estimarlos y los estimadores estándar utilizados. En la sección 8 se proporcionan las medidas de evaluación calculadas en los experimentos de simulación realizados para valorar los estimadores. En la sección 9 se presenta la nueva teoría desarrollada para el cálculo de los estimadores EBLUP (Empirical Best Linear Unbiased Predictor), asistidos por modelos lineales mixtos con un factor de área aleatorio, cuando los pará-

metros del modelo son estimados haciendo uso de los pesos individuales y de los pesos del área. Toda esta teoría ha sido posteriormente implementada en SAS/IML tal y como se describe en la siguiente sección, la 10.

En el año 2002 comenzaron los experimentos de simulación para probar los estimadores estándar y más adelante, en el año 2003, continuaron con la aplicación a diseños más complejos y también a nuevos estimadores basados en la nueva teoría. Entre los experimentos desarrollados merece la pena destacar los siguientes:

- Cálculo de los estimadores estándar, aplicando los diseños tipo EPA y ECV descritos en la sección 6, para la estimación de tres parámetros poblacionales.
- Cálculo de los estimadores estándar y de los nuevos estimadores para la estimación del paro OIT aplicando un diseño modificado del tipo EPA, que proporciona una muestra no autoponderada a nivel de estrato, lo que resulta beneficioso para el estudio del impacto del uso de los pesos muestrales.
- Cálculo de los estimadores estándar para la estimación de la renta aplicando el diseño tipo ECV y una colección de covariables, diferente a la usada con los estimadores estándar, que incluye a la renta imponible del IRPF, lo que es muy útil para analizar el efecto de esta covariable en las estimaciones obtenidas.
- Cálculo de los estimadores estándar para la estimación de la renta aplicando el diseño tipo ECV pero con incorporación de un mecanismo de falta de respuesta correlado con la renta del hogar, de esta manera pudimos estudiar el impacto del uso de pesos informativos.

Todos los experimentos y resultados obtenidos, tanto en el año 2002 como en el año 2003, son descritos en las secciones 11 y 12 respectivamente. Finalmente, en la sección 13, se proporcionan las referencias para posibles consultas de los lectores interesados.

Una vez finalizado el proyecto EURAREA, siguiendo la práctica común en los proyectos europeos del 5º Programa Marco, se ha realizado una difusión muy amplia de sus resultados, con especial atención al acceso vía web.

Tanto el volumen final, en el que se integran los resultados producidos por el INE que resume el presente documento, como la teoría y el software asociados, están disponibles en la web oficial del proyecto (http://www.statistics.gov/eurarea).

También merece la pena mencionar que el último objetivo del proyecto, que era crear un foro de debate, se ha conseguido en Agosto de este año celebrando la primera conferencia internacional en Estimación en Areas Pequeñas (SAE 2005) que ha tenido lugar en Jyväskylä (Finlandia), en los días 28-31 de Agosto.

En esta conferencia han intervenido personajes internacionales en esta materia (D. Pfeffermann, J.N.K. Rao, C. Särndal, ...) junto con otras comunicaciones orales o de tipo póster presentadas por los aproximadamente 100 participantes procedentes tanto del mundo de la estadística oficial como del mundo universitario.

En concreto, el INE ha presentado junto con la Universidad Miguel Hernández de Elche un paper sobre las aplicaciones de estas técnicas de estimación en la EPA (ver el documento en inglés Small Area Estimation in the Spanish Labour Force Survey).

Los trabajos presentados recogen los resultados obtenidos en los últimos años en el campo de las Estimaciones en Areas Pequeñas en relación con:

- la investigación (desarrollos teóricos y metodológicos)
- la producción (aplicaciones al mundo real)

En la web oficial de la conferencia, http://www.stat.jyu.fi/sae2005/, es posible encontar más información y, en particular, el Proceedings de los abstracts de los trabajos presentados. Actualmente se está llevando a cabo la revisión de todas, o si no la gran mayoría, de las presentaciones para publicarlas en la revista polaca Statistics in Transition.

2 Creación de la población artificial (APES)

Una de las actividades del proyecto EURAREA en su primera fase, que forma parte de las acciones desarrolladas para el estudio de cualquiera de los cuatro temas mencionados en la Introducción, es la creación de un fichero de datos que contenga el desempleo, la renta y la composición del hogar como variables objetivo junto con una amplia colección de variables socio-demográficas como variables auxiliares.

Esta base de datos, que se describe a continuación, fue construida durante el primer año del proyecto y ha sido extensamente documentada, constituyendo la población artificial de España en el proyecto que designaremos por APES (Artificial Population EURAREA-Spain). Todas las variables en el fichero son designadas por APES+nº.

El fichero APES contiene 40 variables y 38.872.268 registros. La longitud de registro es de 90 caracteres. Para cada registro, las 35 primeras variables proceden del Censo de Población y Viviendas 1991, es decir, 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.

A continuación, en cada registro, 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.

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 por lo que esta variable ha sido imputada en APES a partir de la información recogida con la EPA en el 2º trimestre de 1991.

- Ingreso neto total anual del hogar (APES502), que se obtiene por imputación a partir de la Encuesta de Presupuestos Familiares (EPF) 1990-91. 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.

El fichero de la EPA contiene 199.231 registros (individuos) con 23 variables APES y el de la EPF contiene 21.155 registros (hogares) con 21 variables APES. Algunas de estas variables son discretas y otras son continuas, por lo que para la imputación de las variables APES501 y APES502 se han ajustado modelos generales de regresión que nos permiten predecir el valor de las mismas. Las variables discretas 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.

Los diseños muestrales de ambas encuestas seleccionan muestras independientes en las diferentes Comunidades Autónomas. En consecuencia, dos posibilidades pueden ser consideradas para la estimación de los modelos: ajustar un único modelo a la muestra nacional o ajustar 18 modelos, uno a cada muestra regional. En este caso, el profundo conocimiento de la economía y la sociedad española ha sido decisivo para preferir la estimación región a región. Sin embargo, el tamaño muestral en la EPF es menor que en la EPA y, entonces, la solución de estimar regresiones distintas en cada región para predecir la variable APES502 no siempre será posible debido a la desproporción existente entre el número de parámetros a estimar en el modelo regional y el tamaño muestral en la Comunidad Autónoma. Por todo ello, las regiones uniprovinciales no han recibido un tratamiento individual sino que han sido añadidas a alguna otra región de comportamiento similar, con excepción de la Comunidad Autónoma de Baleares debido a su carácter de archipiélago.

Con el fin de predecir el valor de APES501 para todos los individuos de la población artificial APES, se llevó a cabo un estudio para seleccionar el modelo de regresión lineal que mejor funcionaba con una variable respuesta binaria y el resultado fue la aplicación de modelos de regresión logística (logit). Una investigación similar fue realizada para la predicción de los valores de APES502 y la decisión final fue el uso de modelos tipo log-normal.

Después de estimar los modelos seleccionados con la técnica de mínimos cuadrados ponderados, se obtuvo un indicador del grado de ajuste para cada modelo. Para los modelos logísticos utilizados para imputar APES501, el siguiente porcentaje ha sido obtenido:

$$Q = \left(1 - \frac{\text{desviación del modelo}}{\text{desviaciónnula}}\right) 100$$

donde el numerador y el denominador son respectivamente la desviación del modelo elegido y la desviación del modelo nulo (aquel que contiene un único parámetro) al modelo saturado (aquel con tantos parámetros a estimar como observaciones).

Para los modelos de tipo log-normal ajustados para predecir la variable APES502, se ha utilizado como indicador el coeficiente, R^2, de determinación del modelo definido por el cociente:

$$R^2 = \frac{VE}{VT}$$

donde el numerador y el denominador representan respectivamente la variabilidad de las observaciones explicada por el modelo y la variabilidad total.

En las tablas 2.1 y 2.2 que se presentan a continuación, aparecen las variables auxiliares utilizadas en los modelos junto con el valor del indicador de la calidad del ajuste. El número de observaciones utilizadas en el ajuste del modelo es dado en la columna denominada ***n*** mientras que el número de parámetros estimados viene dado en la última columna.

Tabla 2.1. 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	**parámetros**
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				**COVARIABLES**	
APES 103	Provincia	APES 211	Condición socioeconómica	APES 203	Edad
APES 104	Estrato	APES 301	Sexo de la Persona de Referencia (PR)	APES405	Número de personas ocupadas
APES 202	Sexo	APES 303	Estudios de más alto nivel completados por PR	APES 409	Tamaño del Hogar
APES 206	Relación con la persona de referencia	APES 304	Relación con la actividad de PR		
APES 207	Estudios de más alto nivel completados	APES 306	Situación profesional de PR		
APES 208	Relación con la actividad	APES 307	Condición socioeconómica de PR		
APES 210	Situación profesional	APES 403	Tipo de Hogar		

Tabla 2.2 Factores y covariables, del fichero EPF de 1991, que aparecen en los modelos log-normales de APES502

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

3 ¿Qué se entiende por pequeña área en España?

Las encuestas realizadas por el INE han sido diseñadas para proporcionar información, periódicamente, de una gran variedad de características y para un determinado equilibrio entre coste y acuracidad, no sólo a nivel nacional sino también a otros niveles de subpoblación o dominio.

En el contexto de las encuestas por muestreo, un estimador de un dominio se denomina directo si para su construcción se utiliza solamente los datos recogidos con la muestra para ese dominio. A menudo ocurre que existe la necesidad de obtener estimaciones para ciertos dominios que no han sido tenidos en cuenta a la hora de diseñar la encuesta, bien porque no estaban disponibles los recursos necesarios bien porque dicha necesidad surgió después de haber realizado el diseño.

En tales dominios lo más probable es que el estadístico de la encuesta tenga pocas observaciones o incluso ninguna. Bajo estas circunstancias, una pequeña área es un dominio para el que no es posible obtener estimaciones directas con una adecuada precisión y suele hacer referencia al caso en que el dominio se define geográficamente.

En el marco del proyecto EURAREA, cada uno de los países participantes debía obtener estimaciones para dos niveles geográficos: la provincia (NUT3) y cualquier otro inferior a éste.

En el caso español, los niveles geográficos elegidos fueron las provincias (NUT3, 52 en toda España) y las Comarcas-EURAREA, que es la unidad territorial "ad hoc" de nivel NUT4 para la investigación del proyecto EURAREA. Esta división territorial consiste en un área geográfica intermedia entre los niveles de Provincia y Municipio que el INE utiliza como 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, siendo su tamaño medio poblacional de algo de más de 60.000 habitantes.

Esta división se consideró válida, a efectos del análisis, en sustitución de una posible división territorial tipo NUT4 (¿comarca?) que no estaba disponible en el momento de arranque del proyecto EURAREA. Actualmente, ya se encuentran muy avanzados los trabajos de definición del nivel NUT4 en España, por lo que será posible aplicar los resultados EURAREA a esta división territorial real a corto plazo.

4 Fuentes complementarias para la generación de covariables de área

A partir de la información a nivel de microdato contenida en el fichero APES se puede obtener información auxiliar agregada de la pequeña área. También, para el suministro de información auxiliar a nivel de pequeña área, se pueden considerar como fuentes complementarias a APES el Registro de Demandantes del INEM y La Renta Imponible del IRPF, ambas agregadas a nivel de Comarcas-Eurarea.

4.1 REGISTRO DE DEMANDANTES DEL INEM

El Instituto Nacional de Empleo proporcionó los totales de demandantes por municipios que, combinados adecuadamente con datos procedentes de la EPA relativos a 1991 y 1998 nos permitió obtener datos de demandantes para los niveles geográficos requeridos en el proyecto y para el año 1991.

4.2 REGISTRO ADMINISTRATIVO DE LA RENTA IMPONIBLE DEL IRPF

Cada año la Agencia Estatal de Administración Tributaria (AEAT) recoge las rentas anuales declaradas por los contribuyentes y, por primera vez ha proporcionado los totales agregados a nivel de código postal según la fuente de origen (pensiones, desempleo, actividades agrícolas, etc.). Al igual que en el caso de los totales de demandantes, se han derivado totales al nivel de la sección censal relativos al año 1991 mediante la aplicación de índices de deflación para los cruces de sección censal y código postal. Con posterioridad, la AEAT ya puede suministrar directamente el dato, a nivel de sección censal al INE, por lo que este último paso, en aplicaciones actuales, ya no será necesario.

5 Definición de los parámetros poblacionales

Los parámetros de la población artificial APES que son objeto de estimación en el proyecto EURAREA son:

- Proporción de la población que está desempleada OIT. Este parámetro corresponde a la proporción de personas en paro en la población de personas de 16 y más años. En términos del fichero APES esta proporción puede expresarse como:

$$\frac{\text{Total}\left(\text{APES503}=1\right)}{\text{Total}\left(\text{APES203}\geq 16\right)}$$

- Media del ingreso anual por unidad de consumo en los hogares. Este parámetro corresponde a:

$$\frac{1}{\text{N}}\sum_{\text{i=1}}^{\text{N}}\frac{\text{APES502}}{\text{APES505}}$$

donde N es el número total de hogares y la variable APES505 representa el número de Unidades de Consumo del Hogar según la escala OCDE modificada. Esta escala asigna los siguientes coeficientes:

  - 1 para el sustentador principal
  - 0,5 para los restantes adultos (14 ó más años)
  - 0,3 para los niños (menos de 14 años)

La suma de los miembros de hogar ponderada por estos coeficientes es lo que se denomina número de unidades de consumo del hogar.

- Proporción de hogares unipersonales. Este parámetro se corresponde con la proporción de la población de hogares con un único miembro. Más concretamente y en términos del fichero APES, esta proporción viene dada por la siguiente expresión:

$$\frac{\text{Total}\left(\text{APES504}=1\right)}{\text{Total}\left(\text{APES206}=1\right)}$$

donde la identificación del hogar se hace a través de la persona de referencia del hogar.

6 Diseños muestrales aplicados en el proyecto.

En el marco del proyecto EURAREA, todos los países participantes deben evaluar los estimadores estándar, usando su propia base de datos, y obtener estimaciones de los 3 parámetros poblacionales de interés definidos en la sección anterior.

La evaluación de los resultados consiste principalmente en la obtención de estimaciones del sesgo y del error cuadrático medio de los estimadores, mediante simulaciones de un número elevado de reiteraciones muestrales con esquemas similares a los utilizados por las encuestas oficiales en el mundo real.

Así pues han sido seleccionados dos diseños muestrales complejos similares a los comúnmente utilizados en las encuestas europeas: el de la Encuesta de Población Activa (EPA) para estimar el desempleo OIT y el de la Encuesta de Condiciones de Vida (ECV) para estimar la renta y la composición del hogar.

6.1 DISEÑO SIMILAR A LA ENCUESTA DE POBLACIÓN ACTIVA (EPA)

El tipo de muestreo utilizado, es un muestreo bietápico con estratificación de las unidades de primera etapa.

Las unidades de primera etapa son las secciones censales y están agrupadas por estratos según el tipo de municipio a que pertenecen (importancia demográfica), aplicando la siguiente clasificación:

- Estrato 1: Municipios capital de provincia.
- Estrato 2: Municipios autorrepresentados, importantes en relación con la capital.
- Estrato 3: Otros municipios autorrepresentados, importantes en relación con la capital o municipios mayores de 100.000 habitantes.
- Estrato 4: Municipios entre 50.000 y 100.000 habitantes
- Estrato 5: Municipios entre 20.000 y 50.000 habitantes
- Estrato 6: Municipios entre 10.000 y 20.000 habitantes
- Estrato 7: Municipios entre 5.000 y 10.000 habitantes
- Estrato 8: Municipios entre 2.000 y 5.000 habitantes
- Estrato 9: Municipios menores de 2.000 habitantes

Las unidades de segunda etapa están constituidas por los hogares y dentro de ellas no se realiza submuestreo alguno, recogiéndose información de todas las personas que tengan su residencia habitual en los mismos

La selección de la muestra se ha realizado de forma independiente en cada provincia y de manera que dentro de cada estrato cualquier individuo tenga la misma probabilidad de ser seleccionado, es decir, se obtienen muestras autoponderadas dentro de cada estrato.

Para ello, las unidades de primera etapa han sido seleccionadas sin reemplazamiento y con probabilidad proporcional al tamaño según el número de hogares. Dentro de cada sección seleccionada en primera etapa, se han seleccionado 20 hogares con probabilidades iguales y sin reemplazamiento.

La población artificial a utilizar en los experimentos EURAREA se ha reducido de acuerdo con todos los países participantes, en aras de conseguir un mayor equilibrio entre los tamaños a manejar por los diferentes países y por economía de recursos en el proceso de los datos.

Así pues la población objetivo considerada para estimar el desempleo está definida por todas las personas con 16 o más años en el universo español de EURAREA, es decir, perteneciente a las Comunidades Autónomas de Andalucía, Canarias, Galicia, Valenciana y Madrid (aproximadamente más de la mitad de la población artificial APES).

En la tabla 6.1.1 se incluyen los tamaños muestrales en la primera etapa:

Tabla 6.1.1 Tamaños muestrales utilizados en la primera etapa del diseño tipo EPA, por estrato y provincia

Provincias	1	2	3	4	5	6	7	8	9	Total
Alava	27				3		6			36
Albacete	15				6		3	6	6	36
Alicante	18	9		12	12	6	9	3	3	72
Almería	15				3	6	3	9		36
Avila	12						9		15	36
Badajoz	24				12	6	9	12	9	72
Baleares	30				12	12	9	9		72
Barcelona	60		30	12	18	9	6	6	3	144
Burgos	18				6		3		9	36
Cáceres	18				6	3	12	15	18	72
Cádiz	15	12	6	12	12	9	6			72
Castellón	24				15	12	3	9	9	72
Ciudad Real	12	9			12	9	15	9	6	72
Córdoba	30				12	9	12	9		72
Coruña (La)	21			12	6	12	15	6		72
Cuenca	12						6	6	12	36
Girona	15				12	12	9	12	12	72
Granada	24				12	6	12	18		72
Guadalajara	15						6		15	36
Guipúzcoa	24			6	15	15	6	6		72
Huelva	12					9	6	9		36
Huesca	12					9	6		9	36
Jaén	15	6			12	12	12	15		72
León	24	9				6	18		15	72
Lleida	12					3	3	6	12	36
Logroño	21					6	6	6	9	48
Lugo	12					6	9	9		36
Madrid	99		21	9	9		6			144
Málaga	36			6	12	6		12		72
Murcia	24	12		6	12	9	9			72
Navarra	30				3	6	6	15	12	72
Ourense	12					3	9	12		36
Oviedo	21	24		18	15	12	9	9		108
Palencia	15						9		12	36
Palmas (Las)	36			6	12	9	6	3		72
Pontevedra	12	24			9	15	9	3		72
Salamanca	18					3	3		12	36
S.Cruz Tenerife	24	12			12	9	9	6		72
Santander	24	9				12	6	12	9	72
Segovia	15						6		15	36
Sevilla	48			6	18	12	12	12		108
Soria	12						9		15	36
Tarragona	18	12			6	12	6	9	9	72
Teruel	12					3	9		12	36
Toledo	15	15					15	15	12	72
Valencia	48			6	24	9	9	6	6	108
Valladolid	24					3	3		6	36
Vizcaya	30	6		12	6	6	6	6		72
Zamora	12					6			18	36
Zaragoza	48					6	9		9	72
Ceuta	12									12
Melilla	12									12
Total	1.170	159	57	123	321	321	384	300	309	3.168

6.2 DISEÑO SIMILAR A LA ENCUESTA DE CONDICIONES DE VIDA (ECV)

Para la selección de la muestra se ha utilizado un muestreo bietápico estratificado en las unidades de primera etapa, que son las secciones censales. Las unidades de segunda etapa son los hogares.

Con este criterio se ha seleccionado una muestra independiente en cada Comunidad Autónoma.

La variable de estratificación de las secciones censales es el tamaño del municipio al que pertenece pero con ligeras diferencias respecto a la estratificación aplicada en el diseño anterior como a continuación se describe:

- Estrato 0: Municipio de Barcelona.
- Estrato 1: Resto de municipios capital de provincia.
- Estrato 2: Municipios con más de 100.000 habitantes.
- Estrato 3: Municipios entre 50.000 y 100.000 habitantes.
- Estrato 4: Municipios entre 20.000 y 49.999 habitantes.
- Estrato 5: Municipios entre 10.000 y 19.999 habitantes.
- Estrato 6: Municipios menores de 10.000 habitantes.

En la primera etapa las secciones censales han sido seleccionadas sin reemplazamiento y con probabilidades proporcionales al tamaño según el número de hogares. En la segunda etapa, 8 hogares fueron seleccionados, con probabilidades iguales y sin reemplazamiento, en cada sección censal seleccionada en la etapa anterior.

Para estimar la renta y la composición del hogar, la población investigada está constituida por todos los hogares pertenecientes al universo español en EU-RAREA, es decir, pertenecientes a las Comunidades Autónomas de Andalucía, Canarias, Galicia, Valenciana y Madrid (más de la mitad de la población APES aproximadamente).

En la tabla 6.2.1 se incluyen los tamaños muestrales utilizados en la primera etapa:

Tabla 6.2.1 Tamaños muestrales utilizados en la primera etapa del diseño tipo ECV por estrato y comunidad autónoma

Regiones	0	1	2	3	4	5	6	Total
Andalucía		60	5	18	30	24	43	180
Aragón		34				6	21	61
Asturias (Principado)		11	14	10	5	11	9	60
Baleares (Illes)		21			10	8	12	51
Canarias		26	5	4	13	8	13	69
Cantabria		17		5		7	17	46
Castilla-La Mancha		45		3	3	7	49	107
Castilla y León		15		5	6	5	41	72
Cataluña	47	8	30	15	22	16	34	172
C. Valenciana		36	6	9	29	14	28	122
Extremadura		11			9	5	33	58
Galicia		18	10	6	8	21	34	97
C. de Madrid		85	27	10	11		8	141
Murcia (Región de)		18	9	4	12	9	6	58
C. F. de Navarra		17			3	6	24	50
País Vasco		29	5	11	10	12	15	82
Rioja (La)		19				5	16	40
Ceuta y Melilla		34						34
TOTAL	47	504	111	100	171	164	403	1500

7 Los estimadores estándar

Para la estimación de cada uno de los parámetros poblacionales investigados se han probado más de 20 estimadores de pequeñas áreas mediante experimentos de simulación. Sin embargo, uno de los principales objetivos del proyecto EURAREA es evaluar la metodología estándar por lo que a continuación vamos a dar una definición completa de los estimadores para pequeñas áreas denominados estándar en el contexto de EURAREA.

Previamente vamos a mencionar algunas consideraciones sobre la notación que se va a utilizar a lo largo de todo el documento:

- Subíndices: s es utilizado para designar a las muestras.
  h=1, ,H para los estratos
  d=1,.,.D para las pequeñas áreas.
  i para las unidades investigadas.
- Tamaños: N para la población investigada y n para la muestra seleccionada. Cuando N o n lleven algún subíndice, indica el tamaño del subconjunto definido por el subíndice. Por ejemplo, n_d es el tamaño de la muestra seleccionada en la pequeña área d.
- Totales: Y ó X. Cuando Y ó X lleven subíndice, indica el total del subconjunto correspondiente al subíndice. Por ejemplo, Y_d denota el total Y en la pequeña área d.
- Medias: $\overline{Y}$ó $\overline{X}$. Cuando $\overline{Y}$ó $\overline{X}$ lleven un subíndice, indica la media del subconjunto correspondiente al subíndice. Por ejemplo $\overline{Y}_d$ denota la media $\overline{Y}$ en la pequeña área d.
- Pesos: w_i es utilizado para el peso muestral de la unidad i. También cuando lleve un subíndice indica la suma de los pesos correspondientes a las unidades muestrales pertenecientes a la subpoblación definida por el subíndice.

De una forma general los parámetros poblacionales investigados en EURAREA se pueden considerar medias poblacionales construidas sobre los valores $y_1,...,y_N$. Es decir, pueden expresarse del modo siguiente:

$$\overline{Y}_d = \frac{1}{N_d}\sum_{i=1}^{N_d} y_i$$

Entonces los estimadores estándar son los siguientes:

Estimador 1: estimador directo

$$\hat{\overline{Y}}_d^{DIRECT} = \frac{1}{\hat{N}_d} \sum_{i \in s_d} w_i \, y_i \qquad \text{donde } \hat{N}_d = \sum_{i \in s_d} w_i$$

Estimador 2: estimador general de regresión (GREG)

$$\hat{\overline{Y}}_d^{GREG} = \hat{\overline{Y}}_d^{DIRECT} + \left(\overline{X}_d - \hat{\overline{X}}_d^{DIRECT} \right)^T \hat{\beta}$$

donde $\overline{X}_d = \left(\overline{X}_{d1}, \ldots, \overline{X}_{dp} \right)^T$ es el vector columna de las medias poblacionales de las p variables auxiliares incluidas en el modelo de regresión asumido:

$$y_i = \alpha + x_i^T \beta + e_i$$

siendo $x_i = \left(x_{i1}, \ldots, x_{ip} \right)^T$ el vector columna de los valores de las variables auxiliares asociadas a la unidad i, suponiendo $E(e_i) = 0$ y $V(e_i) = \sigma_e^2$. $\hat{\beta}$ es el estimador del parámetro β obtenido por el método de mínimos-cuadrados.

Estimador 3: estimador sintético bajo el modelo A (modelo de regresión con datos individuales y efectos aleatorios de área):

$$y_i = x_i^T \beta + u_d + e_i$$

donde $u_d \sim N\left(o, \sigma_u^2\right)$ y $e_i \sim N\left(o, \sigma_e^2\right)$ son independientes.

Entonces la expresión del estimador sintético es:

$$\hat{\overline{Y}}_d^{SYNTHA} = \overline{X}_d^T \hat{\beta}$$

Estimador 4: estimador sintético bajo el modelo B (modelo de regresión con datos de área pequeña):

$$\overline{Y}_d = \overline{X}_d^T \beta + u_d \qquad \text{y} \qquad \hat{\overline{Y}}_d^{DIRECT} = \overline{Y}_d + e_d$$

donde. $u_d \sim N\left(o, \sigma_u^2\right)$ y $e_d \sim N\left(o, \sigma_e^2\right)$ independientes.

Entonces la expresión del estimador es:

$$\hat{\overline{Y}}_d^{SYNTHB} = \overline{X}_d^T \hat{\beta}$$

Estimador 5: estimador sintético bajo el modelo C (modelo logístico de regresión con datos de área pequeña):

$$\text{logit}\left(p_d\right) = \overline{X}_d^T \beta + e_d$$

donde $e_d \sim N\left(o, \sigma_e^2\right)$ y p_d representa la probabilidad del valor uno de la variable binaria objeto de estudio en la pequeña área.

Estimador 6: estimador EBLUP (Empirical Best Linear Unbiased Predictor) bajo el modelo A:

$$\hat{\overline{Y}}_d^{EBLUPA} = \hat{\gamma}_d \, \hat{\overline{Y}}_d^{GREG} + \left(1 - \hat{\gamma}_d\right) \overline{X}_d^T \, \hat{\beta}$$

donde $\hat{\gamma}_d = \dfrac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \dfrac{\hat{\sigma}_e^2}{n_d}}$

Estimador 7: estimador EBLUP bajo el modelo B:

$$\hat{\overline{Y}}_d^{EBLUPB} = \hat{\gamma}_d \, \hat{\overline{Y}}_d^{DIRECT} + \left(1 - \hat{\gamma}_d\right) \overline{X}_d^T \, \hat{\beta}$$

$$\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}$$

8 Medidas de evaluación calculadas en las simulaciones

Con el fin de valorar la idoneidad de los estimadores propuestos, K muestras independientes de cada diseño muestral han sido extraídas del universo español en EURAREA y sobre cada una de ellas se han calculado las estimaciones correspondientes.

Sea $\hat{\overline{Y}}_d(k)$ la estimación de la media poblacional $\overline{Y}_d$ obtenida con la muestra K, las medidas para valorar el comportamiento del estimador han sido las siguientes:

1. Sesgo relativo medio asociado a la pequeña área d:

$$ARB_d = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\hat{\overline{Y}}_d(k)}{\overline{Y}_d} - 1\right)100$$

2. Promedio de los sesgos relativos:

$$\overline{ARB} = \frac{1}{D}\sum_{d=1}^{D} ARB_d$$

donde D es el total de áreas pequeñas.

3. Raíz cuadrada del error cuadrático relativo medio asociado a la pequeña área d:

$$RMSE_d = \frac{100}{\overline{Y}_d}\sqrt{\frac{1}{k}\sum_{k=1}^{K}\left(\hat{\overline{Y}}_d(k) - \overline{Y}_d\right)^2}$$

4. Promedio de las raíces de los errores relativos:

$$\overline{RMSE} = \frac{1}{D}\sum_{d=1}^{D} RMSE_d$$

Durante el año 2002, después de terminar la construcción de la población artificial APES, se realizaron varios ejercicios de aproximación para valorar el consumo de tiempo y recursos al generar un número elevado de muestras, aplicar los estimadores y calcular las medidas de evaluación. Así pues, se empezó seleccionando 10.000 muestras con muestreo aleatorio simple sin estratificar y, posteriormente, con estratificación y se aplicaron en ellas más de 20 estimadores diferentes que fueron evaluados. Durante este tiempo se observó que algunos errores relativos resultaban excesivamente grandes en algunas áreas debido

a que el valor del parámetro poblacional a estimar era muy próximo a cero (sobre todo al estimar proporciones). En consecuencia al concepto de error relativo fue sustituido por el de error absoluto y las siguientes medidas comenzaron a ser calculadas:

5. Raíz cuadrada del error cuadrático medio asociado a la pequeña área d:

$$EMSE_d = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\overline{Y}}_d(k) - \overline{Y}_d\right)^2}$$

6. Promedio de las raíces de los errores:

$$\overline{EMSE} = \frac{1}{D}\sum_{D=1}^{D} EMSE_d$$

9. Teoría

En la teoría de estimación en áreas pequeñas, los modelos lineales mixtos con un factor aleatorio se usan como herramienta para la obtención de estimadores EBLUP (empirical best linear unbiased predictor) de medias o totales poblacionales. Los estimadores de los parámetros de tales modelos (coeficientes de regresión y componentes de la varianza) tienen las propiedades de eficiencia solamente en el caso de que los datos analizados provengan de diseños muestrales con probabilidades iguales de inclusión. Cuando se usa un diseño muestral con probabilidades de inclusión desiguales, los estimadores de los parámetros del modelo pierden sus propiedades óptimas. Así pues, se pueden plantear las siguientes preguntas: ¿es necesario adaptar los algoritmos de ajuste de modelos al caso de pesos muestrales desiguales (o más generalmente, a diseños muestrales complejos)?, ¿cómo hacerlo?, ¿qué diferencias habría en los estimadores de áreas pequeñas?

En esta sección, se dan varias modificaciones del algoritmo Fisher scoring para ajustar modelos lineales mixtos con un factor aleatorio, cuando las muestras se obtienen de diseños muestrales complejos (muestreos distintos del aleatorio simple). Más concretamente, se estudia el problema de cómo introducir los pesos muestrales (inversas de probabilidades de inclusión) en los algoritmos de ajuste.

9.1. ESTIMACIÓN EBLUP EN DISEÑOS MUESTRALES COMPLEJOS

En esta sección se describe la teoría estándar de predicción empírica lineal insesgada óptima bajo modelos lineales mixtos con un factor aleatorio y con muestras obtenidas de diseños muestrales complejos.

9.1.1 El modelo y el algoritmo Fisher scoring censal

Consideremos una población con N unidades y D áreas pequeñas. Sea N_d el número de unidades en el área pequeña d. Sea Y la variable de interés tomando los valores $\boldsymbol{y} = (y_1,...,y_N)$. Supongamos que este vector poblacional es una realización de las variables $Y_1,...,Y_N$ distribuidas según el modelo

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}, \tag{9.1}$$

donde $\boldsymbol{y} = \boldsymbol{y}_{N\times 1}$, $\boldsymbol{X} = \boldsymbol{X}_{N\times p}$ es una matriz de constantes con los valores de las variables auxiliares en columnas, $r(\boldsymbol{X}) = p$, $\boldsymbol{\beta} = \boldsymbol{\beta}_{p\times 1}$ es el vector de coeficientes de regresión de las covariables o efectos fijos, $\boldsymbol{Z} = \boldsymbol{Z}_{N\times D} = diag\left(\mathbf{1}_{N_1},\ldots,\mathbf{1}_{N_D}\right)$ donde $\mathbf{1}_{N_d}$ es un vector columna de unos de tamaño N_d, $\boldsymbol{u} = \boldsymbol{u}_{D\times 1} \sim N\left(\boldsymbol{0},\sigma_u^2\boldsymbol{I}_D\right)$ es independiente de $\boldsymbol{e} = \boldsymbol{e}_{N\times 1} \sim N\left(\boldsymbol{0},\sigma_e^2 I_N\right)$, y $\boldsymbol{I}_a = diag\left(1,\ldots,1\right)_{a\times a}$. Obsérvese además que el modelo (9.1) puede escribirse alternativamente como en Prasad y Rao (1990); es decir,

$$y_{dj} = \boldsymbol{x}_{dj}\beta + u_d + e_{dj}, \quad d = 1,\ldots D,\ j = 1,\ldots, N_d, \tag{9.2}$$

donde y_{dj} es la característica de interés para la unidad j del área d y $\boldsymbol{x}_{dj}$ es la fila (d,j) de la matriz $\boldsymbol{X}$ conteniendo las correspondientes variables auxiliares. El modelo (9.2) puede interpretarse como un modelo con ordenada en el origen aleatoria.

Sea $\theta = \left(\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2\right)^t$ el vector de parámetros. La función de densidad del vector $\boldsymbol{y}$ bajo (9.1) es

$$f_\theta(\boldsymbol{y}) = c\left|\boldsymbol{V}\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)^t \boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)\right\},$$

donde $\boldsymbol{V} = \mathrm{var}(\boldsymbol{y}) = \sigma_e^2 \boldsymbol{I}_N + \sigma_u^2 \boldsymbol{Z}\boldsymbol{Z}^t = diag(\boldsymbol{V}_1,...,\boldsymbol{V}_D)$, $\boldsymbol{V}_d = \sigma_e^2 \boldsymbol{I}_{N_d} + \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}_{N_d}^t$, d=1,.D, y c es una constante. La log-densidad es

$$l(\theta) = \ln\, f_\theta(\boldsymbol{y}) = \ln c - \frac{1}{2}\ln\left|\boldsymbol{V}\right| - \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)^t \boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)$$

El vector de puntuaciones, evaluado en el punto θ, es

$$\boldsymbol{S}(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \left(\frac{\partial l(\theta)}{\partial \beta}, \frac{\partial l(\theta)}{\partial \sigma_u^2}, \frac{\partial l(\theta)}{\partial \sigma_e^2}\right) = \left(\boldsymbol{S}_\beta^t, \boldsymbol{S}_{\sigma_u^2}, \boldsymbol{S}_{\sigma_e^2}\right)^t,$$

y los estimadores de máxima verosimilitud se obtienen resolviendo la ecuación $\mathbf{0} = \boldsymbol{S}(\theta)$. El algoritmo Fisher scoring se usa frecuentemente para calcular numéricamente estos estimadores. El algoritmo comienza con unos valores iniciales (semillas) de los estimadores, $\theta^0 = \left(\boldsymbol{\beta}_0^t, \sigma_{u,0}^2, \sigma_{e,0}^2\right)$, y se actualizan en cada iteración mediante la ecuación

$$\theta^{i+1} = \theta^i + \boldsymbol{F}\left(\theta^i\right)^{-1} \boldsymbol{S}\left(\theta^i\right), \tag{9.3}$$

donde

$$\boldsymbol{F}\left(\theta\right) = -E\left[\frac{\partial \boldsymbol{S}(\theta)}{\partial \theta}\right] = \begin{pmatrix} \boldsymbol{F}_{\beta\beta} & \boldsymbol{F}_{\beta\sigma_u^2} & \boldsymbol{F}_{\beta\sigma_e^2} \\ \boldsymbol{F}_{\beta\sigma_u^2} & \boldsymbol{F}_{\sigma_u^2\sigma_u^2} & \boldsymbol{F}_{\sigma_u^2\sigma_e^2} \\ \boldsymbol{F}_{\beta\sigma_e^2} & \boldsymbol{F}_{\sigma_u^2\sigma_e^2} & \boldsymbol{F}_{\sigma_e^2\sigma_e^2} \end{pmatrix}$$

es la matriz de información de Fisher en el punto θ. Definimos

$$\boldsymbol{y}_d = \begin{pmatrix} y_{d1} \\ \vdots \\ y_{dN_d} \end{pmatrix}, \quad \boldsymbol{X}_d = \begin{pmatrix} x_{d11} & \cdots & x_{d1p} \\ \vdots & \ddots & \vdots \\ x_{dN_d 1} & \cdots & x_{dN_d p} \end{pmatrix}, \quad \text{d=1,.,D.} \tag{9.4}$$

Entonces, bajo el modelo poblacional (9.1) las puntuaciones (scores) son:

$$S_\beta = \frac{1}{\sigma_e^2}\sum_{d=1}^{D}\left(\boldsymbol{X}_d^t\varsigma_d - \frac{\gamma_d}{N_d}\boldsymbol{X}_d^t\mathbf{1}_{N_d}\mathbf{1}_{N_d}^t\varsigma_d\right),\quad S_{\sigma_u^2} = -\frac{1}{2}\sum_{d=1}^{D}\frac{1-\gamma_d}{\sigma_e^2}N_d + \frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d}{\sigma_e^2}\right)^2\varsigma_d^t\mathbf{1}_{N_d}\mathbf{1}_{N_d}^t\varsigma_d \quad \text{y}$$

$$S_{\sigma_e^2} = -\frac{1}{2}\sum_{d=1}^{D}\frac{N_d-\gamma_d}{\sigma_e^2} + \frac{1}{2\left(\sigma_e^2\right)^2}\sum_{d=1}^{D}\left[\varsigma_d^t\varsigma_d + \frac{\gamma_d\left(\gamma_d-2\right)}{N_d}\varsigma_d^t\mathbf{1}_{N_d}\mathbf{1}_{N_d}^t\varsigma_d\right], \text{ donde}$$

$\varsigma_d = \boldsymbol{y}_d - \boldsymbol{X}_d\boldsymbol{\beta}$ y $\gamma_d = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/N_d}$, d=1,..,D.

Los elementos de la matriz de información de Fisher son

$$F_{\beta\beta} = \frac{1}{\sigma_e^2}\sum_{d=1}^{D}\left(\boldsymbol{X}_d^t\boldsymbol{X}_d - \frac{\gamma_d}{N_d}\boldsymbol{X}_d^t\mathbf{1}_{N_d}\mathbf{1}_{N_d}^t\boldsymbol{X}_d\right),\quad F_{\sigma_u^2\sigma_u^2} = \frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d}{\sigma_e^2}\right)^2 N_d^2,$$

$$F_{\sigma_u^2\sigma_e^2} = \frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d}{\sigma_e^2}\right)^2 N_d,\quad F_{\sigma_e^2\sigma_e^2} = \frac{1}{2\left(\sigma_e^2\right)^2}\sum_{d=1}^{D}\left[N_d + \gamma_d\left(\gamma_d-2\right)\right] \quad \text{y} \quad F_{\beta\sigma_u^2} = F_{\beta\sigma_e^2} = 0.$$

9.1.2 Predictor empírico lineal insesgado óptimo

Sea n_d el número unidades de la población en la muestra del área d y definamos $f_d = n_d/N_d$. El predictor lineal insesgado óptimo (EBLUP) de $\overline{Y}_d$ es

$$\hat{\bar{Y}}_d^{eblup} = (1-f_d)\hat{\bar{Y}}_d^{eblupa} + f_d\left[\hat{\bar{Y}}_d + \hat{\beta}\left(\overline{\boldsymbol{X}}_d - \hat{\bar{X}}_d\right)\right],$$

donde

$$\hat{\bar{Y}}_d^{eblupa} = \overline{\boldsymbol{X}}_d\hat{\beta} + \hat{\gamma}_d^w(\hat{\bar{Y}}_d^{direct} - \hat{\bar{X}}_d^{direct}\hat{\beta}),\qquad \hat{\gamma}_d^w = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + (\hat{\sigma}_e^2/w_{d\cdot})},\qquad \overline{\boldsymbol{X}}_d = \frac{1}{N_d}\sum_{j=1}^{N_d}\boldsymbol{x}_{dj},$$

$$\hat{\bar{X}}_d = \frac{1}{n_d}\sum_{j=1}^{n_d}\boldsymbol{x}_{dj},\quad \hat{\bar{X}}_d^{direct} = \frac{1}{\hat{N}_d}\sum_{j=1}^{n_d}w_{dj}\boldsymbol{x}_{dj},\quad \hat{\bar{Y}}_d = \frac{1}{n_d}\sum_{j=1}^{n_d}y_{dj},\quad \hat{\bar{Y}}_d^{direct} = \frac{1}{\hat{N}_d}\sum_{j=1}^{n_d}w_{dj}y_{dj},\quad \hat{N}_d = \sum_{j=1}^{n_d}w_{dj}.$$

Los estimadores $\hat{\beta}$, $\hat{\sigma}_u^2$ y $\hat{\sigma}_e^2$ se pueden calcular usando una versión muestral del algoritmo Fisher scoring algoritmo introducido en (9.3).

Sea $\Omega=\{1,...,N\}$ una población finita. Sean $s\subset\Omega$ y $r=\Omega-s$ los conjuntos de unidades muestreadas y no muestreadas respectivamente. Es interesante observar que

$$\hat{\bar{Y}}_d^{eblup}=\frac{1}{N_d}\sum_{j=1}^{N_d}\hat{\bar{Y}}_{dj}^{eblup}$$

donde

$$\hat{\bar{Y}}_{dj}^{eblup}=\begin{cases}\boldsymbol{x}_{dj}\hat{\beta}+\hat{\gamma}_d^w\left(\hat{\bar{Y}}_d^{wdirect}-\hat{\bar{X}}_d^{wdirect}\hat{\beta}\right) & if \quad y_{dj}\in r\\ y_{dj} & if \quad y_{dj}\in s\end{cases}$$

de modo que $\hat{\bar{Y}}_d^{eblup}$ también se llama estimador predictivo. Por otra parte, en el proyecto EURAREA se ha usado el siguiente estimador proyectivo (EBLUPA).

$$\begin{aligned}\hat{\bar{Y}}_d^{eblupa}&=\frac{1}{N_d}\sum_{j=1}^{N_d}\left\{\boldsymbol{x}_{dj}\hat{\beta}+\hat{\gamma}_d^w\left(\hat{\bar{Y}}_d^{direct}-\hat{\bar{X}}_d^{direct}\hat{\beta}\right)\right\}\\&=(1-\hat{\gamma}_d^w)\overline{\boldsymbol{X}}_d\hat{\beta}+\hat{\gamma}_d^w\left(\hat{\bar{Y}}_d^{direct}+\left(\overline{\boldsymbol{X}}_d-\hat{\bar{X}}_d^{direct}\right)\hat{\beta}\right)=(1-\hat{\gamma}_d^w)\hat{\bar{Y}}_d^{syntha}+\hat{\gamma}_d^w\hat{\bar{Y}}_d^{greg}.\end{aligned}$$

9.1.3 Probabilidades de inclusión y pesos

En la Teoría del muestreo probabilístico un plan de muestreo (o diseño muestral) es un esquema para elegir la muestra de modo que cada subconjunto (muestra) s de unidades tiene una probabilidad p(s) de ser seleccionada. Supongamos que se extrae una muestra de tamaño n de acuerdo con un diseño muestral con probabilidades de inclusión

$$\pi_{dj}=\sum_{s\in S(d,j)}p(s),$$

donde S(d, j) es el conjunto de todas las muestras de tamaño n que contienen al individuo j del área pequeña d. Los pesos, $w_{dj}=1/\pi_{dj}$, pueden interpretarse como el número de unidades poblacionales representadas en la muestra por la unidad muestral j del área pequeña d. Consideremos también las probabilidades de inclusión $\pi_d=P(s\cap d\neq\varnothing)$, las probabilidades condicionales $\pi_{j|d}=\pi_{dj}/\pi_d$ y sus pesos correspondientes $w_d=1/\pi_d$ y $w_{j|d}=1/\pi_{j|d}$.

Los diseños muestrales complejos se usan frecuentemente en las encuestas de ámbito nacional para reducir costes y para tener en cuenta las características geográficas y socioeconómicas de la población estudiada. Cuando se usan diseños muestrales bietápicos, lo habitual es que las unidades de primera etapa no coincidan con las áreas pequeñas de interés. Por tal motivo, en este apartado se ilustra el cálculo de las probabilidades de inclusión de tales áreas pequeñas en

un diseño muestral bietápico con estratificación en primera etapa. Suponemos que las unidades de primera etapa son territorios completamente contenidos en alguna área pequeña y, con objeto de usar una denominación habitual, las llamaremos secciones censales. Las secciones censales se seleccionan sin reemplazamiento y probabilidades iguales. Las unidades de segunda etapa son las viviendas y 20 de ellas son seleccionadas sin reemplazamiento con un muestreo aleatorio simple. Se incluyen en la muestra todos los individuos (unidades finales) pertenecientes a una vivienda seleccionada. Esta es una versión modificada del diseño muestral de la Encuesta de Población Activa.

Sea H el número de estratos, m_h el número de secciones censales del estrato h seleccionado en la muestra, M_h el número de secciones censales del estrato h en la población, M_{hd} el número de secciones censales en el área pequeña d del estrato h en la población, N_h el número de viviendas del estrato h en la población, N_{hi} el número de viviendas en la sección censal i del estrato h en la población, N_{hd} el número de viviendas en el área pequeña d del estrato h en la población, $_{hij}$ la probabilidad de inclusión de la vivienda j de la sección censal i del estrato h, $_{hi}$ la probabilidad de inclusión de la sección censal i del estrato h, $_{j|hi}$ la probabilidad de inclusión en segunda etapa de la vivienda j de la sección censal i cuando en primera etapa la sección censal i del estrato h ha sido seleccionada, y finalmente sea $_{d}$ la probabilidad de inclusión del área pequeña d. Entonces:

$$\pi_{j|hi} = 1 - \frac{\binom{N_{hi}-1}{20}}{\binom{N_{hi}}{20}} = \frac{20}{N_{hi}}, \quad \pi_{hi} = 1 - \frac{\binom{M_h-1}{m_h}}{\binom{M_h}{m_h}} = \frac{m_h}{M_h} \quad \text{and} \quad \pi_{hij} = \pi_{j|hi}\pi_{hi} = \frac{20m_h}{M_h N_{hi}},$$

de modo que los pesos muestrales de los individuos de la vivienda j de la sección censal i del estrato h son

$$w_{hij} = \frac{1}{\pi_{hij}} = \frac{M_h N_{hi}}{20 m_h}.$$

Por otra parte, suponiendo que el tamaño poblacional es suficientemente grande, podemos aceptar en los cálculos siguientes que la selección de las unidades de primera etapa haya sido con reemplazamiento. En tal caso, se obtiene $w_d = 1/\pi_d$ con

$$\pi_d = 1 - \prod_{h=1}^{H} P(d \cap h \cap s = \varnothing) \approx \prod_{h=1}^{H} \left(1 - \frac{M_{hd}}{M_h}\right)^{m_h} = 1 - \exp\left\{\sum_{h=1}^{H} m_h \ln\left(1 - \frac{M_{hd}}{M_h}\right)\right\}.$$

9.2 VERSIONES MUESTRALES DEL ALGORITMO FISHER SCORING CENSAL

En esta sección se deducen versiones muestrales del algoritmo Fisher scoring (9.3) suponiendo que el modelo (9.2), o una modificación del mismo, también es válido para la muestra.

9.2.1 Algoritmo Fisher scoring censal

Para obtener estimadores EBLUP de áreas pequeñas, los estadísticos usualmente suponen que el modelo (9.2) es también válido para la muestra; es decir, suponen que

$$y_{dj} = \boldsymbol{x}_{dj}\boldsymbol{\beta} + u_d + e_{dj},\ \ d = 1,\ldots D,\ j = 1,\ldots,n_d\ , \qquad (9.5)$$

donde $u_d \sim iid\ N\left(\boldsymbol{0},\sigma_u^2\right)$ son independientes y $e_{dj} \sim iid\ N\left(\boldsymbol{0},\sigma_e^2\right)$. Los estimadores EBLUP se obtienen entonces ajustando el modelo (9.5).

El algoritmo Fisher scoring sin pesos usa la ecuación de actualización (9.3) con puntuaciones (scores)

$$S_\beta = \frac{1}{\sigma_e^2}\sum_{d=1}^{D}\left(\boldsymbol{X}_d^t\varsigma_d - \frac{\gamma_d}{n_d}\boldsymbol{X}_d^t\boldsymbol{1}_{n_d}\boldsymbol{1}_{n_d}^t\varsigma_d\right),\quad S_{\sigma_u^2} = -\frac{1}{2}\sum_{d=1}^{D}\frac{1-\gamma_d}{\sigma_e^2}n_d + \frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d}{\sigma_e^2}\right)^2\varsigma_d^t\boldsymbol{1}_{n_d}\boldsymbol{1}_{n_d}^t\varsigma_d\ \ ,$$

$$S_{\sigma_e^2} = -\frac{1}{2}\sum_{d=1}^{D}\frac{n_d-\gamma_d}{\sigma_e^2} + \frac{1}{2\left(\sigma_e^2\right)^2}\sum_{d=1}^{D}\left[\varsigma_d^t\varsigma_d + \frac{\gamma_d(\gamma_d-2)}{n_d}\varsigma_d^t\boldsymbol{1}_{n_d}\boldsymbol{1}_{n_d}^t\varsigma_d\right],$$

y elementos de la matriz de información de Fisher

$$F_{\beta\beta} = \frac{1}{\sigma_e^2}\sum_{d=1}^{D}\left(\boldsymbol{X}_d^t\boldsymbol{X}_d - \frac{\gamma_d}{n_d}\boldsymbol{X}_d^t\boldsymbol{1}_{n_d}\boldsymbol{1}_{n_d}^t\boldsymbol{X}_d\right),\quad F_{\sigma_u^2\sigma_u^2} = \frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d}{\sigma_e^2}\right)^2 n_d^2,\quad F_{\sigma_u^2\sigma_e^2} = \frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d}{\sigma_e^2}\right)^2 n_d\ ,$$

$$F_{\sigma_e^2\sigma_e^2} = \frac{1}{2\left(\sigma_e^2\right)^2}\sum_{d=1}^{D}\left[n_d + \gamma_d\left(\gamma_d - 2\right)\right] \quad \text{y} \quad F_{\beta\sigma_u^2} = F_{\beta\sigma_e^2} = 0\ ,$$

donde $\boldsymbol{y}_d$ y $\boldsymbol{X}_d$ se toman de (9.4) con n_d en el lugar de N_d, $\varsigma_d = \boldsymbol{y}_d - \boldsymbol{X}_d\beta$ y $\gamma_d = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/n_d}$

Al proponer el modelo (9.5) para la muestra, uno está implícitamente asumiendo que la distribución del vector muestral $(y_1,...,y_n)$ se obtiene directamente de (9.2), y que en consecuencia se ignora el mecanismo aleatorio de selección de la

muestra. Desafortunadamente, este hecho no puede justificarse con argumentos probabilísticos (véase, por ejemplo, la sección 2.6.2 de Vaillant et al. (2000)). En consecuencia, es necesario hacer una investigación más profunda para aclarar cuando el modelo (9.5) puede usarse para obtener del estimador EBLUP de áreas pequeñas asociado al modelo (9.2).

9.2.2 Algoritmo Fisher scoring con pesos en las unidades

Un procedimiento alternativo consiste en usar la muestra para construir una población artificial repitiendo la unidad (d,j) w_{dj} veces. Para esa población artificial proponemos, por analogía con (9.2), el modelo

$$y_{djk} = \boldsymbol{x}_{djk}\boldsymbol{\beta} + u_d + e_{djk}, \quad d = 1,\dots D,\ j = 1,\dots,n_d,\ k = 1,...,w_{dj}, \tag{9.6}$$

donde $u_d \sim iid\ N\left(\boldsymbol{0},\sigma_u^2\right)$ son independientes y $e_{djk} \sim iid\ N\left(\boldsymbol{0},\sigma_e^2\right)$. Los estimadores EBLUP se obtienen ajustando el modelo (9.6). En (9.6), tomamos en el lugar de w_{dj} el entero más cercano a $1/\pi_{dj}$. Nótese que en las encuestas realizadas por los institutos nacionales de estadística $1/\pi_{dj}$ es usualmente mayor que 100, de modo que esta última aproximación es admisible.

También es posible considerar el siguiente modelo para la muestra.

$$y_{dj} = \boldsymbol{x}_{dj}\boldsymbol{\beta} + u_d + e_{dj}, \quad d = 1,\dots D,\ j = 1,\dots,n_d, \tag{9.7}$$

donde $u_d \sim iid\ N\left(\boldsymbol{0},\sigma_u^2\right)$ y $e_{dj} \sim iid\ N\left(\boldsymbol{0},w_{dj}^{-1}\sigma_e^2\right)$ son independientes. Los estimadores EBLUP pueden obtenerse ajustando el modelo (9.7).

En la sección 4.5 de Morales y Molina (2002) se demuestra que si σ_u^2 y σ_e^2 son conocidos, entonces los estimadores/predictores de máxima verosimilitud $\hat{\beta}$ y $\hat{u}_d$, d=1,...,D, en los modelos (9.6) y (9.7) coinciden. Bajo las mismas hipótesis, Los correspondientes estimadores BLUP también coinciden. Recuérdese que Henderson (1975) demostró que el estimador BLUP de $\boldsymbol{\ell}^t\boldsymbol{\beta} + u_d$ es $\boldsymbol{\ell}^t\hat{\beta} + \hat{u}_d$ en modelos lineales mixtos del tipo (9.6) o (9.7) con σ_u^2 y σ_e^2 conocidos.

Usando argumentos de consistencia, si se admite que la población artificial con el modelo (9.6) es una aproximación razonablemente buena de la población real con el modelo (9.1), entonces se propone ajustar el modelo (9.7) para estimar/predecir β y u_d, d=1,...,D, del modelo (9.1).

El Algoritmo Fisher scoring con pesos en las unidades usa la ecuación de actualización (1.3) con puntuaciones (scores)

$$S_\beta = \frac{1}{\sigma_e^2}\sum_{d=1}^{D}\left(\boldsymbol{X}_d^t\boldsymbol{W}_d\varsigma_d - \frac{\gamma_d^w}{w_{d\cdot}}\boldsymbol{X}_d^t\boldsymbol{w}_{n_d}\boldsymbol{w}_{n_d}^t\varsigma_d\right), \quad S_{\sigma_u^2} = -\frac{1}{2}\sum_{d=1}^{D}\frac{1-\gamma_d^w}{\sigma_e^2}w_{d\cdot} + \frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d^w}{\sigma_e^2}\right)^2\varsigma_d^t\boldsymbol{w}_{n_d}\boldsymbol{w}_{n_d}^t\varsigma_d\ ,$$

$$S_{\sigma_e^2} = -\frac{1}{2}\sum_{d=1}^{D}\frac{n_d-\gamma_d^w}{\sigma_e^2} + \frac{1}{2\left(\sigma_e^2\right)^2}\sum_{d=1}^{D}\left[\varsigma_d^t\boldsymbol{W}_d\varsigma_d + \frac{\gamma_d^w\left(\gamma_d^w-2\right)}{\omega_{d\cdot}}\varsigma_d^t\boldsymbol{w}_{n_d}\boldsymbol{w}_{n_d}^t\varsigma_d\right]$$

y elementos de la matriz de información de Fisher

$$F_{\beta\beta} = \frac{1}{\sigma_e^2}\sum_{d=1}^{D}\left(\boldsymbol{X}_d^t\boldsymbol{W}_d\boldsymbol{X}_d - \frac{\gamma_d^w}{w_{d\cdot}}\boldsymbol{X}_d^t\boldsymbol{w}_{n_d}\boldsymbol{w}_{n_d}^t\boldsymbol{X}_d\right), \quad F_{\sigma_u^2\sigma_u^2} = \frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d^w}{\sigma_e^2}\right)^2 w_{d\cdot}^2\ , \quad F_{\sigma_u^2\sigma_e^2} = \frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d^w}{\sigma_e^2}\right)^2 w_{d\cdot}\ ,$$

$$F_{\sigma_e^2\sigma_e^2} = \frac{1}{2\left(\sigma_e^2\right)^2}\sum_{d=1}^{D}\left[n_d + \gamma_d^w\left(\gamma_d^w-2\right)\right], \quad F_{\beta\sigma_u^2} = F_{\beta\sigma_e^2} = 0\ ,$$

donde $\boldsymbol{y}_d$ y $\boldsymbol{X}_d$ se toman de (9.4) con n_d en el lugar de N_d, $\varsigma_d = \boldsymbol{y}_d - \boldsymbol{X}_d\boldsymbol{\beta}$,

$$\boldsymbol{W}_d = diag\left(w_{d1},\ldots,w_{dn_d}\right)_{n_d\times n_d}, \quad \boldsymbol{w}_{n_d} = \left(w_{d1},\ldots,w_{dn_d}\right)_{n_d\times 1}^t, \quad w_{d\cdot} = \mathbf{1}_{n_d}^t\boldsymbol{w}_{n_d} = \sum_{j=1}^{n_d}w_{dj} \quad \text{y} \quad \gamma_d^w = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/w_{d\cdot}}\ ,$$

d=1,.,D.

Morales, Molina y Santamaría (2002) han obtenido algunos resultados computacionales en los que se muestra que los estimadores EBLUP obtenidos ajustando el modelo (9.7) son más eficientes (en sesgo y error cuadrático medio relativo) que los estimadores EBLUP obtenidos del modelo muestral (9.1). Si los pesos individuales son todos iguales a uno, este algoritmo coincide con el algoritmo Fisher scoring sin pesos.

9.2.3 Algoritmo Fisher scoring con pesos en las unidades y en las áreas

En la sección 9.2.2 se propone implícitamente calcular la verosimilitud poblacional (censal) a partir de la verosimilitud muestral y obtener estimadores consistentes a partir de esta última verosimilitud. Esta forma de proceder es ampliamente aceptada por los estadísticos cuando se ajustan modelos de regresión lineal estándar (con perturbaciones aleatorias en un solo nivel) a los datos muestrales. En tal caso los valores de la variable estudiada en las unidades elementales de la población finita se consideran independientes, de modo que la verosimilitud censal es una suma que puede ser estimada consistentemente ponderando las observaciones. Pfeffermann y otros (1998) dan las siguientes razones por las cuales los modelos multinivel son distintos de los modelos de un nivel respecto de la ponderación de observaciones.

1. Los valores observados en las unidades de una población finita no son independientes en tales modelos y por tanto la log-verosimilitud censal no

es una suma simple a lo largo de la población. Ello implica que no puede ser estimada por el método de ponderar las observaciones muestrales.

2. Las probabilidades de inclusión de las unidades muestrales últimas no proporcionan suficiente información para efectuar una corrección de sesgos apropiada, al contrario de lo que ocurría en el caso de modelos de regresión de un nivel.

Con el objeto de reducir el sesgo de los estimadores procedentes del modelo (9.5), Pfeffermann y otros (1998) sugieren reproducir el método de estimación en la población con las observaciones muestrales, mediante la introducción de pesos a dos niveles. Ellos proponen reemplazar cada suma poblacional sobre las unidades j dentro del área d, por las sumas muestrales con valores ponderados por $w_{j|d}$, y cada suma poblacional sobre las áreas pequeñas d, por la correspondiente suma con elementos ponderados por $w_{d.}$ Para aplicar la sugerencia de Pfeffermann, reemplazamos en las expresiones censales de las puntuaciones (scores) y elementos de la matriz de Fisher las sumas $\sum_{j=1}^{N_d} b_{dj}$ y $\sum_{d=1}^{D} a_d$ por $\sum_{j=1}^{n_d} w_{j|d} b_{dj}$ i $\sum_{d=1}^{D} w_d a_d$ respectivamente. También los tamaños N_d se reemplazan por los correspondientes estimadores $\hat{N}_d = \sum_{j=1}^{n_d} w_{j|d}$. Pfeffermann y otros (1998) justifican su propuesta argumentando que las sumas poblacionales son estimadas de forma insesgada y consistente (con respecto a la distribución del muestreo) por las correspondientes sumas muestrales ponderadas.

Sean $\boldsymbol{y}_d$ y $\boldsymbol{X}_d$ el vector y la matriz definidos en (9.4), pero para las unidades muestrales; es decir, de tamaños n_d y $n_d \times p$ respectivamente. Definamos además $\varsigma_d = \boldsymbol{y}_d - \boldsymbol{X}_d \beta$, $\gamma_{|d} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2 / \hat{N}_d}$, $\boldsymbol{W}_{|d} = diag\left(w_{1|d}, w_{2|d}, \ldots, w_{n_d|d}\right)$ y $\boldsymbol{w}_{|d} = \left(w_{1|d}, w_{2|d}, \ldots, w_{n_d|d}\right)^t$, d=1,.,D. Los estimadores muestrales de las puntuaciones y elementos de la matriz de Fisher del modelo (9.2), con pesos muestrales en dos niveles, son

$$S_\beta = \frac{1}{\sigma_e^2} \sum_{d=1}^{D} w_d \left(\boldsymbol{X}_d^t \boldsymbol{W}_{|d} \varsigma_d - \frac{\gamma_d}{\hat{N}_d} \boldsymbol{X}_d^t \boldsymbol{w}_{|d} \boldsymbol{w}_{|d}^t \varsigma_d \right), \quad S_{\sigma_u^2} = -\frac{1}{2} \sum_{d=1}^{D} w_d \frac{1-\gamma_d}{\sigma_e^2} \hat{N}_d + \frac{1}{2} \sum_{d=1}^{D} w_d \left(\frac{1-\gamma_d}{\sigma_e^2} \right)^2 \varsigma_d^t \boldsymbol{w}_{|d} \boldsymbol{w}_{|d}^t \varsigma_d,$$

$$S_{\sigma_e^2} = -\frac{1}{2} \sum_{d=1}^{D} w_d \frac{\hat{N}_d - \gamma_d}{\sigma_e^2} + \frac{1}{2\left(\sigma_e^2\right)^2} \sum_{d=1}^{D} w_d \left[\varsigma_d^t \boldsymbol{W}_{|d} \varsigma_d + \frac{\gamma_d \left(\gamma_d - 2\right)}{\hat{N}_d} \varsigma_d^t \boldsymbol{w}_{|d} \boldsymbol{w}_{|d}^t \varsigma_d \right],$$

$$F_{\beta\beta} = \frac{1}{\sigma_e^2} \sum_{d=1}^{D} w_d \left(\boldsymbol{X}_d^t \boldsymbol{W}_{|d} \boldsymbol{X}_d - \frac{\gamma_d}{\hat{N}_d} \boldsymbol{X}_d^t \boldsymbol{w}_{|d} \boldsymbol{w}_{|d}^t \boldsymbol{X}_d \right), \quad F_{\sigma_u^2 \sigma_u^2} = \frac{1}{2} \sum_{d=1}^{D} w_d \left(\frac{1-\gamma_d}{\sigma_e^2} \right)^2 \hat{N}_d^2,$$

$$F_{\sigma_u^2 \sigma_e^2} = \frac{1}{2} \sum_{d=1}^{D} w_d \left(\frac{1-\gamma_d}{\sigma_e^2} \right)^2 \hat{N}_d, \quad F_{\sigma_e^2 \sigma_e^2} = \frac{1}{2\left(\sigma_e^2\right)^2} \sum_{d=1}^{D} w_d \left[\hat{N}_d + \gamma_{|d} \left(\gamma_{|d} - 2 \right) \right] \text{ y } F_{\beta \sigma_u^2} = F_{\beta \sigma_e^2} = 0.$$

En las expresiones anteriores, las sumas $\sum_{d=1}^{D}$ no se evalúan de hecho sobre todas las áreas d=1,.,D, sino solamente sobre las que están representadas en la muestra. El algoritmo Fisher scoring con pesos en las unidades y en las áreas se obtiene sustituyendo las puntuaciones y elementos de la matriz de Fisher en (9.3) por sus correspondientes estimadores. Si los pesos individuales y de áreas son todos iguales a uno, este algoritmo coincide con el de Fisher scoring sin pesos. Sin embargo, si solamente son iguales a uno los pesos de las áreas pequeñas, este algoritmo difiere del algoritmo Fisher scoring con pesos en las unidades.

9.2.4 Semillas para los algoritmos Fisher scoring

Se aconseja usar los siguientes valores iniciales de los estimadores.

$$\beta_0 = \left(\sum_{d=1}^{D} \boldsymbol{X}_d^t \boldsymbol{W}_d \boldsymbol{X}_d - \sum_{d=1}^{D} \tfrac{1}{w_{d.}} \boldsymbol{X}_d^t \boldsymbol{w}_{n_d} \boldsymbol{w}_{n_d}^t \boldsymbol{X}_d \right)^{-1} \left(\sum_{d=1}^{D} \boldsymbol{X}_d^t \boldsymbol{W}_d \boldsymbol{y}_d - \sum_{d=1}^{D} \tfrac{1}{w_{d.}} \boldsymbol{X}_d^t \boldsymbol{w}_{n_d} \boldsymbol{w}_{n_d}^t \boldsymbol{y}_d \right),$$

$$\sigma_{u,0}^2 = \frac{1}{D} \left[\sum_{d=1}^{D} \frac{1}{w_{d.}^2} \boldsymbol{y}_d^t \boldsymbol{w}_{n_d} \boldsymbol{w}_{n_d}^t \boldsymbol{y}_d - 2\boldsymbol{\beta}_0^t \sum_{d=1}^{D} \frac{1}{w_{d.}^2} \boldsymbol{X}_d^t \boldsymbol{w}_{n_d} \boldsymbol{w}_{n_d}^t \boldsymbol{y}_d + \boldsymbol{\beta}_0^t \left(\sum_{d=1}^{D} \frac{1}{w_{d.}^2} \boldsymbol{X}_d^t \boldsymbol{w}_{n_d} \boldsymbol{w}_{n_d}^t \boldsymbol{X}_d \right) \boldsymbol{\beta}_0 \right],$$

$$\sigma_{e,0}^2 = \frac{1}{n - r(\boldsymbol{X})} \left[\sum_{d=1}^{D} \boldsymbol{y}_d^t \boldsymbol{W}_d \boldsymbol{y}_d - 2\boldsymbol{\beta}_0^t \sum_{d=1}^{D} \boldsymbol{X}_d^t \boldsymbol{W}_d \boldsymbol{y}_d + \boldsymbol{\beta}_0^t \left(\sum_{d=1}^{D} \boldsymbol{X}_d^t \boldsymbol{W}_d \boldsymbol{X}_d \right) \boldsymbol{\beta}_0 \right].$$

10 Software

En el proyecto EURAREA el equipo de investigación español ha desarrollado software en C++ y en SAS/IML. Las simulaciones en 2002 se realizaron en C++ y las simulaciones en 2003 en SAS/IML. La selección de muestras se ha realizado siempre con C++ ya que éste es un lenguaje de programación de "bajo nivel", que es más flexible, permite una mejor gestión de la memoria y produce una ganancia significativa en la velocidad de los algoritmos. Sin embargo, SAS/IML es un lenguaje de programación de "alto nivel" en el sentido de que tiene una batería de procedimientos estadísticos incorporados que pueden ser utilizados directamente y de forma sencilla en la fase de estimación, y además los Institutos Nacionales de Estadística usan generalmente SAS/IML. Por esta razón, en el proyecto EURAREA, el software de estimación en áreas pequeñas debía ser desarrollado en SAS/IML y no en C++.

Las principales razones por las cuales se ha producido software en C++ y se han realizado simulaciones con el citado lenguaje de programación han sido: el tamaño de los ficheros que contienen el universo artificial para las simulaciones (2,4 Gb en el caso español), la dificultad de extraer muestras con diseños muestrales complejos, y la necesidad de realizar los experimentos de simulación con un número alto de replicaciones. En ese sentido, las simulaciones del año 2002 se realizaron con 10.000 replicaciones extrayendo una muestra completa en cada replicación. Excepto en el caso de los algoritmos Fisher scoring descritos en la sección 9, el software no se desarrollo en forma cerrada y lista para usar. La estrategia de programación no estuvo dirigida a la producción de software, sino a obtener la ganancia máxima de eficiencia en los experimentos de simulación. Esto significa, por ejemplo, que dentro de cada replicación cada vez que se calcula una cantidad los cálculos parciales realizados no se vuelven a repetir. Las simulaciones del año 2003 se hicieron con 500 replicaciones y el proceso se estructuró en dos partes. En la primera parte, se extrajeron 500 muestras al azar usando C++. En este caso, se produjeron subrutinas para diferentes diseños muestrales. En la segunda parte se usó SAS/IML para tratar secuencialmente las 500 muestras.

Información adicional sobre la implementación de los algoritmos Fisher scoring, tanto en el caso de utilizar pesos de unidades como en el caso de usar pesos de unidades y de áreas, puede encontrarse en los documentos en inglés mencionados en la sección anterior.

11 Simulaciones realizadas en el 2002

Los experimentos de simulación realizados para valorar los estimadores descansan en la base de seleccionar muestras independientes, calcular las estimaciones del parámetro en las pequeñas áreas y compararlas con los valores poblacionales conocidos.

Dada la cantidad de operaciones a realizar en las simulaciones efectuadas para estimar cada uno de los tres parámetros investigados, se decidió comenzar con el diseño muestral más sencillo aunque respetando los tamaños muestrales descritos en la sección 6. Así pues, las primeras muestras independientes extraídas del universo español en EURAREA correspondían a un esquema de muestreo con probabilidades iguales y, mas adelante se obtuvieron muestras estratificadas con probabilidades iguales en cada estrato.

Primero se hizo una prueba con los registros en la Comunidad Valenciana y se seleccionaron 10.000 muestras sin estratificar y otras 10.000 muestras estratificadas. Todas ellas fueron posteriormente procesadas para evaluar los estimadores a nivel provincial y comarcal, así cuando se trataba de estimadores basados o asistidos por modelos, el modelo se ajustaba con los datos muestrales obtenidos en la comunidad autónoma.

Más adelante se extendió el trabajo a todo el universo español en EURAREA y el número de muestras procesadas fue 2.000, tanto en el caso sin estratificar como estratificando, y las estimaciones se calcularon sólo a nivel provincial y, consecuentemente, los modelos fueron ajustados con los datos muestrales de todo el universo.

También se ha querido analizar la forma más conveniente de ajustar los modelos, es decir, si resulta más beneficioso estimar un modelo para cada región o ajustar un modelo a todo el universo investigado. El caso de la estimación de la renta del hogar parece el indicado para realizar la comparación de los resultados obtenidos después de aplicar ambas formas de ajuste. En el caso de los estimadores estándar, los mejores resultados se han obtenido cuando la estimación de los parámetros del modelo se basan únicamente en los datos muestrales de la comunidad autónoma.

En la siguiente tabla se enumeran los estimadores evaluados en este caso.

Tabla 11.1. Estimadores utilizados en la estimación de la renta

Estimadores	Comentarios
1 Directo (con N_d conocido -Horvitz-Thompson)	
2 Directo (con N_d estimado)	Estimador estándar 1 (DIRECT).
3 Post-estratificado	Variables cualitativas A, B o C.
4 Sintético básico	Variables cualitativas A, B o C.
5 Sintético de regresión	APES409 como covariable.
6 Sintético GREG	APES409.
7 GREG1	APES409. Estimador estándar 2 (GREG).
8 Versión BLUP (Best Linear Unbiased Predictor) del estimador GREG1	APES409.
9 EBLUE1	APES409. Estimador estándar 6 (EBLUPA).
9s Sintético del EBLUE1	APES409. Estimador estándar 3 (SYNTHA).
10 Versión EBLUP (Empirical BLUP) del EBLUE1	APES409.
11 Compuesto dependiente del tamaño muestral (SSD1)	Variables cualitativas A, B o C. Resulta de combinar los estimadores 3 y 4.
12 Compuesto dependiente del tamaño muestral (SSD2)	APES409. Resulta de combinar los estimadores 7 y 6.
13 Compuesto dependiente del tamaño muestral (SSD3)	Variables cualitativas A, B o C. Resulta de combinar los estimadores 2 y 4.
14 GREG2	APES409 y APES412.Estimador estándar 2 (GREG).
15 Versión BLUP del estimador GREG2	APES409 y APES412.
16 GREG3	Variables cualitativas A, B o C, junto con APES409 y APES412. Estimador estándar 2 (GREG).
17 Versión BLUP del estimador GREG3	Variables cualitativas A, B o C, junto con APES409 y APES412.
18 Fay-Herriot	Renta imponible del IRPF como covariable. Resulta de combinar los estimadores 2 y 18s. Estimador estándar 7 (EBLUPB).
18s Sintético del Fay-Herriot	Renta imponible del IRPF como covariable. Estimador estándar 4 (SYNTHB).

Donde las variables auxiliares mencionadas en la tabla son:

- Variable cualitativa A. Es una variable derivada de la variable APES403 (Tipo de hogar) que agrupa a los hogares en 6 clases según el hogar sea unipersonal, esté formado por sólo dos adultos, por un adulto y uno o más niños, por dos adultos y uno o más niños, por tres adultos y uno o más niños, y el resto de los hogares.
- Variable cualitativa B. Es una variable derivada de las variables de individuo APES208 (Relación con la actividad) y la APES211 (Condición socioeconómica) en los componentes del hogar. De esta manera se agrupan los hogares en 4 clases según ningún miembro del hogar sea ocupado, algún miembro del hogar sea ocupado pero ninguno trabaje en determinadas actividades (agrarias, militares, .), ídem pero sólo uno trabaje en dichas actividades y, por último, ídem pero 2 o más trabajan en las mencionadas actividades.

- Variable cualitativa C. Es una variable derivada de la variable APES211 (Estudios de más alto nivel completados) que agrupa a los hogares en 4 clases según todos los adultos del hogar hayan completado la educación secundaria, sólo el 50% o más, menos del 50% pero al menos uno, y ninguno.
- APES409. Tamaño del hogar.
- APES412. Superficie útil de la vivienda en metros cuadrados.

En las tablas 11.2 y 11.3 se presentan los resultados obtenidos para los estimadores estándar ($\mathrm{ARE}_{\mathrm{d}} = \mathrm{ARB}_{\mathrm{d}} + 100$).

Tabla 11.2. Estimadores provinciales en la Comunidad Valenciana cuando sólo se usan datos muestrales de la región para el ajuste de los modelos.

	Alicante		Castellón		Valencia		Media	
Estimador	ARE_1	$RMSE_1$	ARE_2	$RMSE_2$	ARE_3	$RMSE_3$	$\overline{ARE}$	$\overline{RMSE}$
Directo	99.986	1.987	99.976	3.105	100.034	1.563	99.999	2.218
GREG1	99.989	1.982	99.981	3.113	100.039	1.563	100.003	2.219
GREG2	99.989	1.982	99.981	3.113	100.039	1.563	100.003	2.219
EBLUPA	99.988	1.982	99.980	3.113	100.039	1.563	100.002	2.219
SYNTHB	94.001	6.147	99.182	1.635	105.931	6.121	99.705	4.634
EBLUPB	99.986	1.987	99.976	3.105	100.034	1.563	99.999	2.218

Tabla 11.3. Estimadores provinciales en la Comunidad Valenciana cuando se usan datos muestrales de todo el universo para el ajuste de los modelos.

	Alicante		Castellón		Valencia		Media	
Estimador	ARE_1	$RMSE_1$	ARE_2	$RMSE_2$	ARE_3	$RMSE_3$	$\overline{ARE}$	$\overline{RMSE}$
Directo	99.923	1.929	100.060	3.173	100.060	1.588	100.014	2.230
GREG1	99.933	1.901	100.070	3.188	100.057	1.598	100.020	2.229
GREG2	99.933	1.901	100.070	3.188	100.057	1.598	100.020	2.229
EBLUPA	99.932	1.901	100.070	3.188	100.057	1.598	1.598	2.229
SYNTHB	98.223	1.906	103.636	3.708	110.688	10.716	104.182	5.443
EBLUPB	99.923	1.929	100.060	3.173	100.060	1.588	100.014	2.230

En el siguiente gráfico se representan los valores numéricos de $\overline{\text{RMSE}}$:

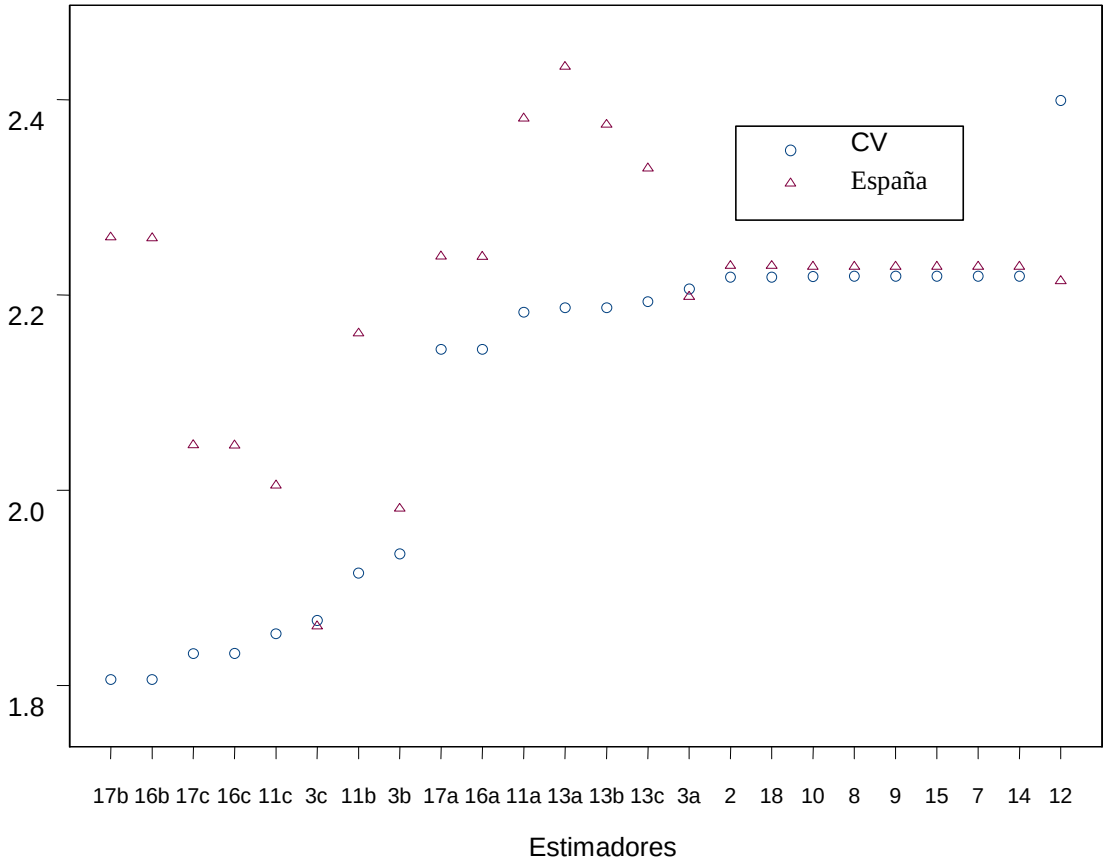

Gráfico11.1 $\overline{\text{RMSE}}$ de los estimadores para la estimación de la renta utilizando sólo los datos de la Comunidad Valenciana (CV) y los de todo el universo (España).

Por último, reseñar que en el año 2002 la definición del parámetro estimado relativo al desempleo OIT era la proporción de la población económicamente activa que estaba desempleada. Sin embargo, esta definición fue modificada para la realización de las simulaciones del año 2003 dado que la población económicamente activa es, en general, una cantidad desconocida y la definición dada en la sección 5 fue adoptada.

12 Simulaciones realizadas en el 2003

En noviembre del año 2002 los participantes en el proyecto EURAREA celebraron una reunión a la que asistió el profesor Tim Holt que, como experto en el tema de estimadores para áreas pequeñas, sugirió la necesidad de que todos los participantes realizaran las simulaciones para probar los estimadores estándar bajo condiciones similares de trabajo para así validar la comparación de resultados entre los diferentes países.

Desafortunadamente el proceso de homogeneización de las simulaciones también significa limitaciones en cuanto al número de cuestiones a investigar, ya que no todos los países tienen los mismos medios disponibles. Por lo tanto, inevitablemente, las simulaciones de los estimadores estándar han obviado el análisis de algunos temas que, fuera del contexto del proyecto, seguramente son de interés general.

12.1 EVALUACIÓN DE LOS ESTIMADORES ESTÁNDAR

Después de discutir largamente sobre la forma de implementar la "simulación estándar" en cada país participante para obtener la máxima comparabilidad en los resultados, se tomaron las siguientes decisiones:

- Se adoptaron las definiciones de los parámetros poblacionales dadas en la sección 5.
- Todos los países utilizaron el mismo conjunto de variables auxiliares en los modelos, con pequeñas variaciones en unos pocos casos.
- En cada país, la selección de las muestras fue lo más parecida posible al método de selección actualmente aplicado para la estimación de los parámetros investigados.
- Los modelos fueron ajustados con los datos de la muestra completa.
- Para obtener resultados comparables, la Oficina de Estadística del Reino Unido (ONS), como coordinadora del proyecto, instó a todos los países a utilizar el software elaborado por su equipo en lenguaje SAS y que permite calcular los estimadores estándar y sus errores cuadráticos medios estimados.
- El número de reiteraciones fue 500 en cada experimento de simulación y en cada país.

En consecuencia 500 muestras similares a la EPA y otras 500 muestras similares a la ECV (ver sección 6) fueron extraídas, de forma independiente, del universo español en EURAREA. Basándose en las primeras se obtuvieron estimaciones del desempleo OIT y a partir de las segundas se estimó la renta y la composición del hogar. Los modelos fueron ajustados usando la muestra completa y los estimadores estándar fueron aplicados para obtener estimaciones provinciales (NUT3) y para las comarcas-EURAREA (NUT4 provisionales). Además el software de la ONS se utilizó ignorando los pesos muestrales al ajustar los modelos.

En los siguientes gráficos se muestran resultados relativos al RMSE_d para estimaciones provinciales y de comarcas-EURAREA.

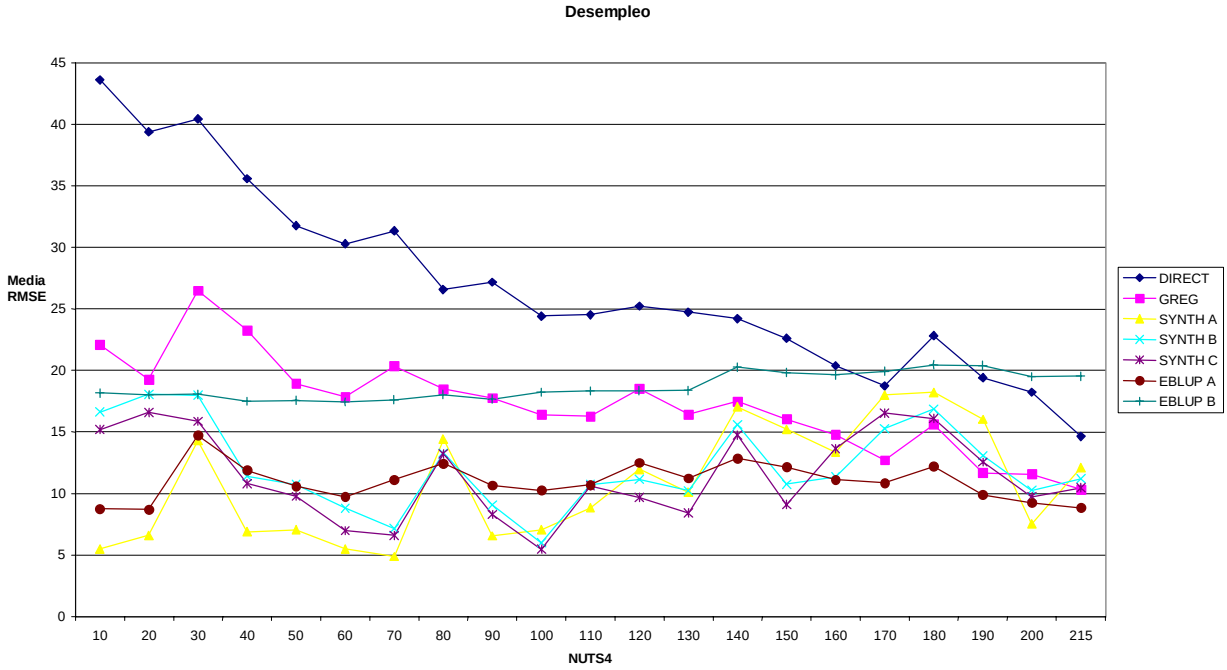

Gráfico 12.1. Valores medios del RMSE, por grupos de comarcas-EURAREA tomados de 10 en 10, ordenados por tamaños muestrales crecientes.

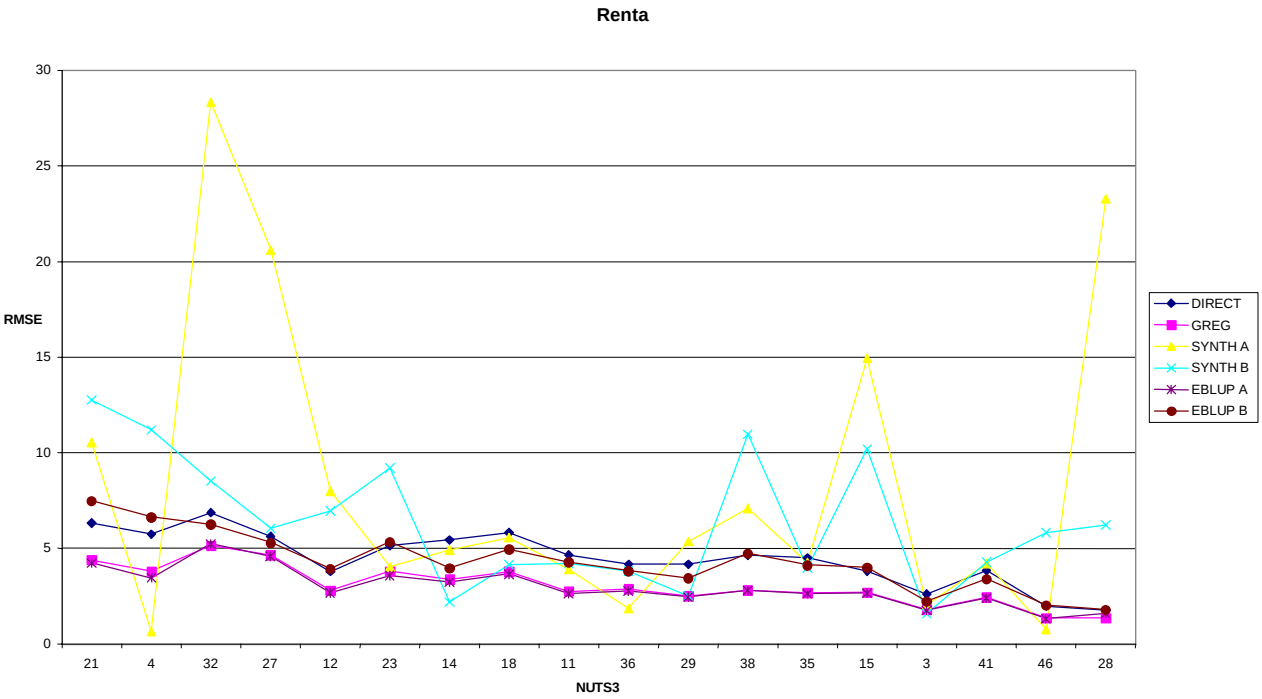

Gráfico 12.2. RMSE de las provincias ordenadas por el tamaño muestral creciente.

Claramente en ambos gráficos se aprecia que los errores muestrales decrecen al aumentar los tamaños muestrales, especialmente en los casos del estimador directo y del GREG. Para las comarcas-EURAREA, las áreas más pequeñas consideradas en estos experimentos, el comportamiento de los estimadores estándar es similar a medida que el tamaño de la muestra en el área crece pero si el área tiene pocas observaciones en la muestra, los estimadores sintéticos se comportan erráticamente comparado con los otros estimadores. En general, se aprecia el mejor comportamiento en los estimadores GREG y EBLUPA.

12.2 ESTIMACIÓN DEL DESEMPLEO OIT BAJO EL DISEÑO MUESTRAL DE LA EPA MODIFICADO

En todos los experimentos mencionados hasta ahora, el cálculo de los estimadores asistidos o basados en modelos se ha realizado sin tener en cuenta los pesos muestrales para el ajuste de los modelos.

Sin embargo, dentro del tema de diseños complejos, hay dos puntos a investigar en relación con los pesos muestrales que hacen necesaria la realización de nuevos experimentos de simulación. Las dos cuestiones a investigar son:

- El comportamiento de los estimadores cuando los modelos son ajustados con pesos muestrales.
- El desarrollo, uso y evaluación de un sistema de pesos muestrales a dos niveles, es decir, en el que tanto las unidades muestrales como las áreas pequeñas tengan asignados pesos muestrales.

Con el fin de alcanzar estos objetivos, enfocamos la investigación a la estimación del desempleo OIT. Por otra parte, es de todos sabido que si los pesos muestrales son iguales, su uso o no uso en el ajuste de los modelos es irrelevante. Así pues también decidimos aumentar la variabilidad de los pesos muestrales del diseño tipo EPA utilizado hasta ahora para obtener la estimación del desempleo.

Por todo ello, el diseño tipo EPA descrito en la sección 6.1 fue modificado. Entonces, bajo esta nueva perspectiva, las unidades muestrales en la primera etapa (secciones censales) fueron seleccionadas con probabilidades iguales en lugar de proporcionales al tamaño. De esta manera, la probabilidad de selección de un individuo depende de la sección censal a la que pertenece y no únicamente del estrato, resultando los pesos muestrales de los individuos mucho más heterogéneos que antes.

Para el cálculo de los estimadores provinciales y comarcales, se seleccionaron 500 muestras independientes con este nuevo diseño, y todos los datos de la muestra fueron utilizados para la estimación de los modelos. En el proceso del ajuste de los modelos, los pesos muestrales intervinieron de diferentes maneras como se describe a continuación:

- Los estimadores DIRECT, EBLUPB, SYNTHB y SYNTHC han sido calculados como en el caso de los estimadores estándar, es decir, sin uso de los pesos muestrales individuales para estimar los parámetros del modelo.
- El estimador GREG se aproximó de dos formas: ajustando el modelo con pesos (GREGW) y sin pesos (GREG), siendo este último método igual al utilizado en las simulaciones estándar.
- Los estimadores EBLUPA Y SYNTHA se han calculado para 3 métodos diferentes de estimación del modelo A:
  - Método 1: como en las simulaciones estándar (EBLUPA1 y SYNTHA1)
  - Método 2: utilizando los pesos muestrales individuales (EBLUPA2 y SYNTHA2)
  - Método 3:utilizando el sistema de pesos a dos niveles desarrollado por la UMH (EBLUPA3 y SYNTHA3)

Para las provincias, en general, el mejor comportamiento corresponde a los estimadores GREG Y EBLUPA mientras los estimadores sintéticos tienen una conducta errática y los valores del RMSE muy elevados. En relación con el sistema de pesos muestrales utilizado, si concentramos nuestra atención en los diferentes estimadores EBLUPA calculados, el método 2 proporciona resultados ligeramente mejores.

Para las comarcas-EURAREA, el comportamiento de los estimadores sintéticos es considerablemente mejor que el de los otros estimadores ya que los tamaños muestrales en éstas son, generalmente, muy pequeños. Al igual que en las simulaciones estándar, los errores muestrales disminuyen a medida que los tamaños muestrales crecen, tanto en el caso de la provincia como en el de la comarcas-EURAREA. Por otra parte, no existen diferencias significativas entre el comportamiento del estimador GREG y GREGW excepto, levemente, a nivel de las comarcas-EURAREA.

En los siguientes gráficos están representados los resultados:

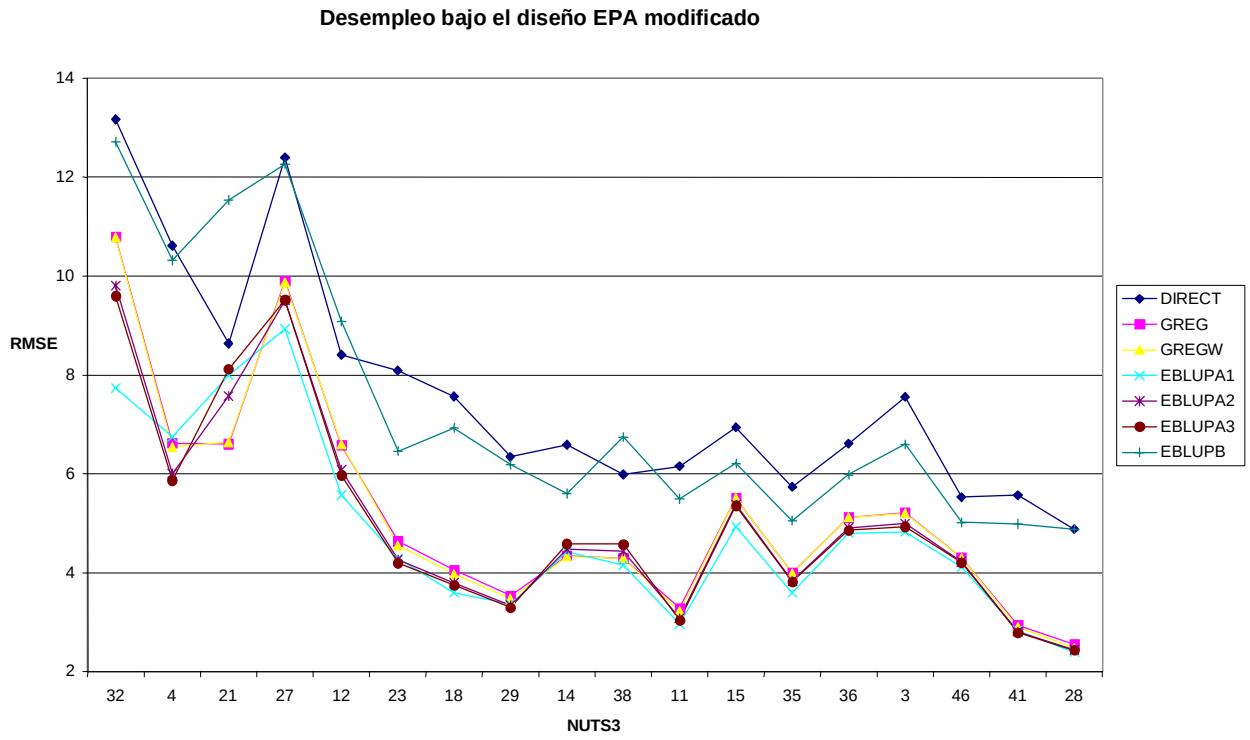

Gráfico 12.3. RMSE de las provincias ordenadas por tamaños muestrales crecientes.

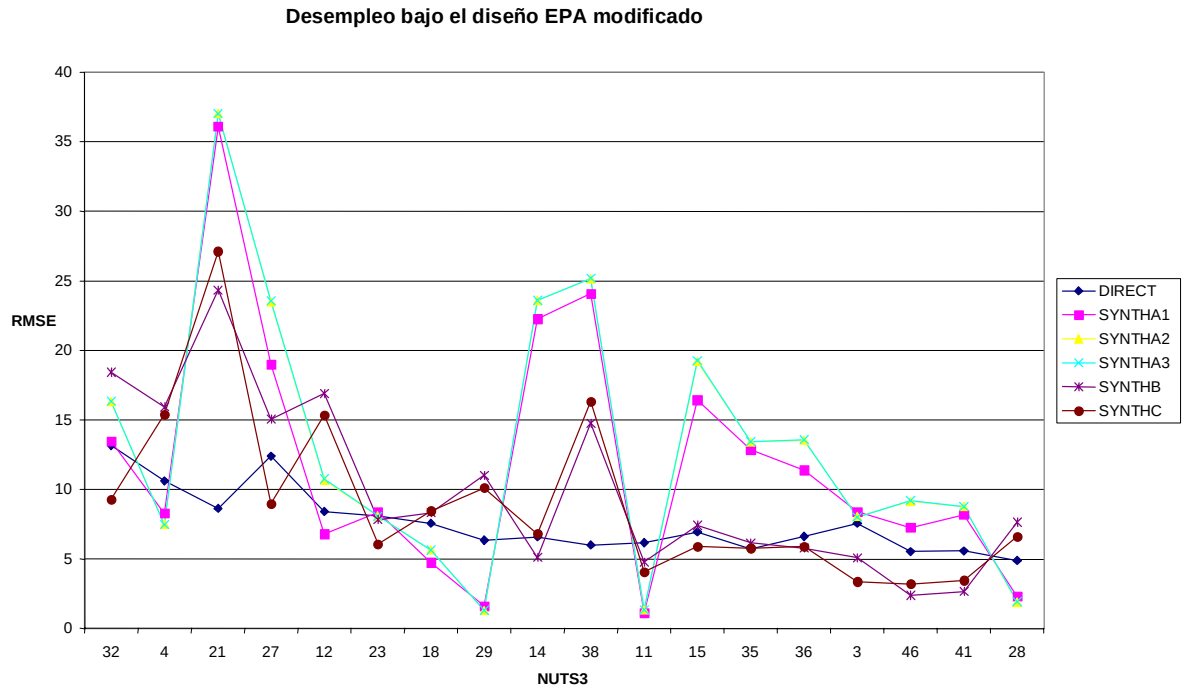

Gráfico 12.4 RMSE de las provincias ordenadas por tamaños muestrales crecientes

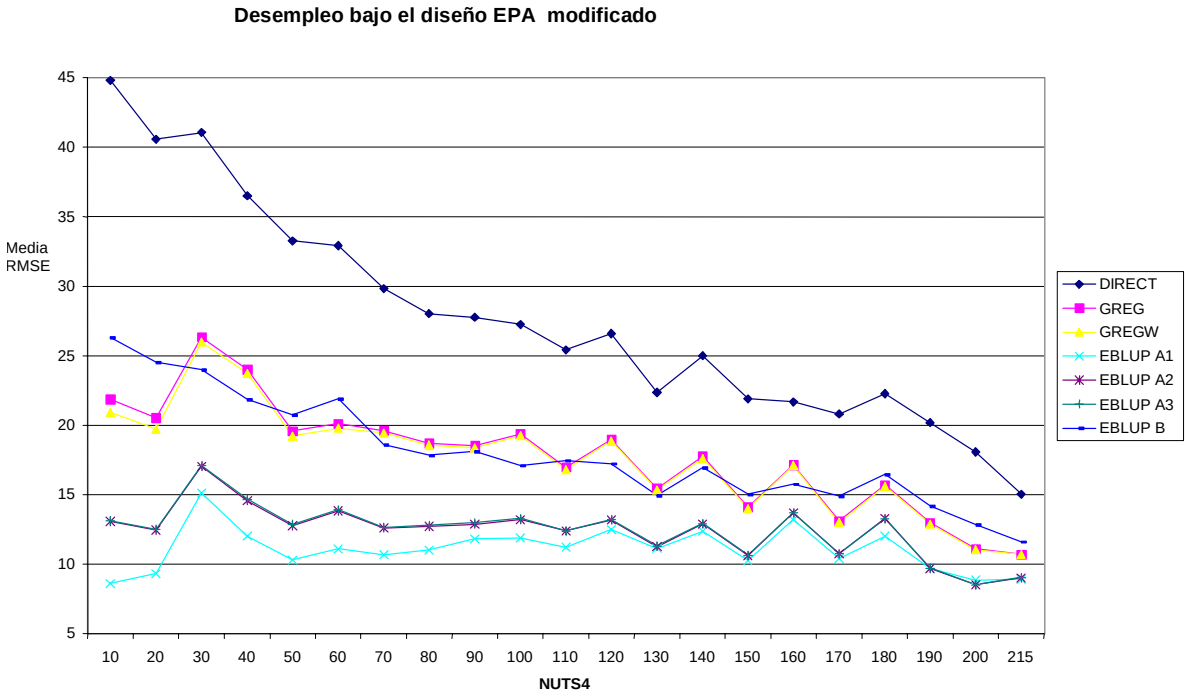

Grafico 12.5. Valores medios, por grupos de comarcas-EURAREA tomados de 10 en 10, de RMSE ordenados por tamaños muestrales crecientes.

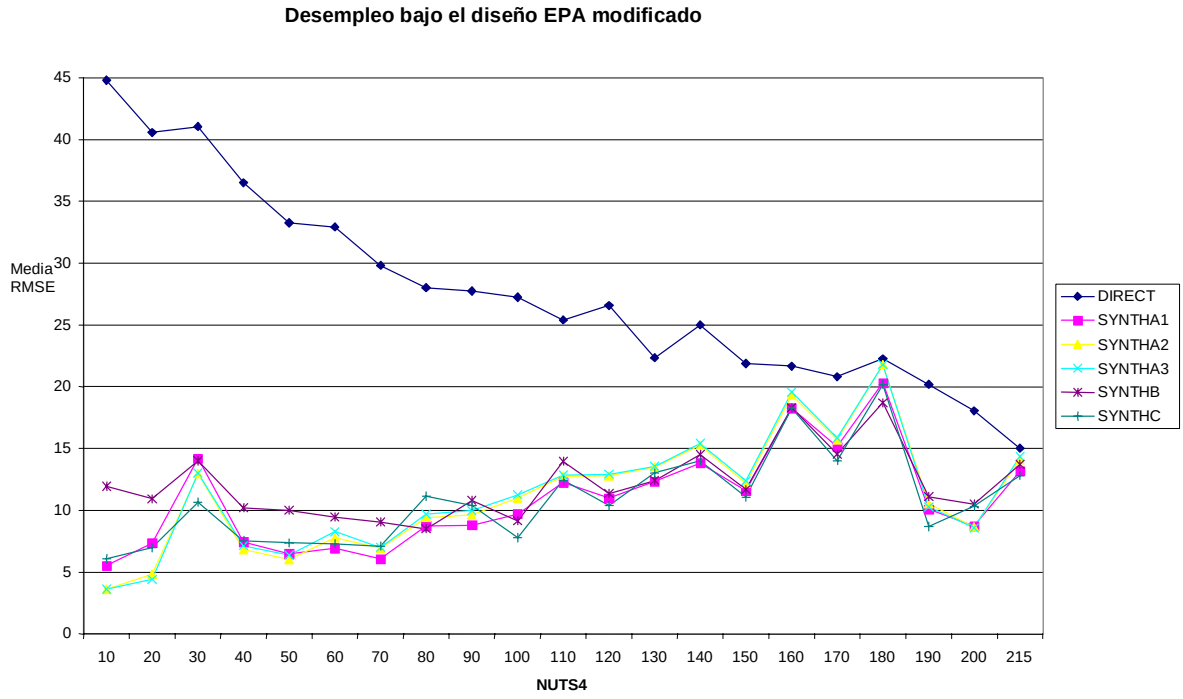

Gráfico 12.6. Valores medios del RMSE, por grupos de comarcas-EURAREA tomados de 10 en 10, ordenados por tamaños muestrales crecientes.

12.3 ESTIMACIÓN DE LA RENTA CON USO DE LA RENTA IMPONIBLE DEL IRPF COMO VARIABLE EXPLICATIVA EN LOS MODELOS.

En los últimos años, el sistema estadístico se está viendo fortalecido por la colaboración de la AEAT para el uso de fuentes fiscales con fines estadísticos. En particular, en el contexto del proyecto EURAREA, la AEAT ha proporcionado los datos agregados a nivel de pequeñas áreas geográficas, preservando la confidencialidad del dato tributario.

En las simulaciones estándar realizadas previamente esta información auxiliar no ha sido utilizada ya que no está disponible en ningún otro de los países participantes en el proyecto. Sin embargo, una vez finalizado el análisis comparativo de los resultados estándar obtenidos por todos los participantes, se acordó que cada país podía realizar cualquier trabajo extra que considerara necesario para alcanzar sus propios objetivos.

En nuestro caso, el interés estaba dirigido a cubrir esta laguna y comenzamos a trabajar de nuevo con la estimación de la renta y las 500 muestras de tipo ECV seleccionadas durante las simulaciones standard.

Para derivar las estimaciones del ingreso, la renta imponible del IRPF agregada a nivel de pequeña área (provincia o comarcas-EURAREA) ha sido utilizada como covariable en los modelos considerados los cuales, a su vez, han sido ajustados en base a toda la muestra con los siguientes procedimientos.

- Método 1: como en las simulaciones estándar
- Método 2: utilizando los pesos muestrales.

Para analizar el impacto de la información auxiliar en los estimadores para áreas pequeñas, vamos a comparar los resultados del Método 1 con los de las simulaciones estándar. Obsérvese que ambos experimentos están basados en el mismo conjunto de muestras (las 500 muestras tipo ECV) y los estimadores han sido calculados con los mismos métodos (sin uso de los pesos para ajustar los modelos) sin embargo, las covariables utilizadas son diferentes. Las variables auxiliares utilizadas en las simulaciones no-estándar (Método 1) son más realistas en el sentido de que el INE puede disponer de esta información en el mundo real. A este conjunto de covariables le vamos a denominar $A_2 = \{X_1,..,X_6\}$ donde:

X_1 = APES409 (Tamaño del hogar)

X_2 = Variable cualitativa B (Condición socioeconómica del hogar derivada de variables directas)

X_3 - X_6 = renta imponible del IRPF según diferentes orígenes de renta (total, pensiones, desempleo y agrarias)

Por otra parte, al conjunto de variables auxiliares utilizadas en las simulaciones estándar le vamos a denominar $A_1 = \{\xi_1,..,\xi_6\}$ donde:

ξ_1 = APES405 (Número de ocupados en el hogar)

ξ_2 = APES409 (Tamaño del hogar)

ξ_3 = APES412 (Superficie útil de la vivienda m^2)

ξ_4 = Suma de las edades de los miembros masculinos

ξ_5 = Suma de las edades de los miembros femeninos

ξ_6 = Variable derivada de la variable cualitativa C (1 si ninguno de los adultos en el hogar ha completado la enseñanza secundaria, y 0 en caso contrario)

son menos realista en el sentido establecido anteriormente.

En las tablas que vienen a continuación, se resumen los valores medios de RMSE y EMSE:

Tabla 12.1. Valores medios del sesgo relativo ARB sobre las provincias como áreas pequeñas

Experimento	DIRECTO	GREG	EBLUPA3	EBLUPB
A_2 / con IRPF	-0,097	-0,070	-0,244	-0,121
A_1 / sin IRPF	-0,097	-0,013	0,084	-0,226

Tabla 12.2. Valores medios del error relativo RMSE sobre las provincias como áreas pequeñas

Experimento	DIRECTO	GREG	EBLUPA3	EBLUPB
A_2 / con IRPF	4,489	3,205	3,036	4,147
A_1 / sin IRPF	4,489	3,049	2,982	4,323

Tabla 12.3. Valores medios del sesgo relativo ARB sobre las comarcas-EURAREA como áreas pequeñas

Experimento	DIRECTO	GREG	EBLUPA3	EBLUPB
A_2 / con IRPF	-0,082	0,068	0,472	0,492
A_1 / sin IRPF	-0,082	0,046	0,762	-0,008

Tabla 12.4. Valores medios del error relativo RMSE sobre las comarcas-EURAREA como áreas pequeñas

Experimento	DIRECTO	GREG	EBLUPA3	EBLUPB
A_2 / con IRPF	12,347	9,711	6,610	8,571
A_1 / sin IRPF	12,347	8,941	7,692	7,806

A nivel de provincia, nuestra conclusión es que el uso de la renta imponible del IRPF es recomendable, sobre todo en los estimadores EBLUPB asistidos por modelos de área, ya que se observa una reducción tanto en el sesgo como en el error relativo. A nivel de comarcas-EURAREA, seguramente debido al incremento del número de áreas pequeñas, el uso de la renta imponible del IRPF es recomendable sobre todo en los estimadores EBLUPA asistidos por modelos de hogar con un factor aleatorio de área, ya que se observa una reducción tanto en el sesgo como en el error relativo.

De todas maneras, es difícil establecer conclusiones definitivas pues no hay que olvidar que la variable investigada, el ingreso en el hogar, es una variable imputada, ya que no se recoge en el Censo de Población de 1991 y, en consecuencia, existe una fuente de error potencial que no está bajo total control en los experimentos analizados.

12.4 ESTIMACIÓN DE LA RENTA CON FALTA DE RESPUESTA.

Con el fin de extender la investigación del efecto de los pesos muestrales, incluimos en el tema de trabajo del equipo español de EURAREA un estudio sobre el impacto del uso de diseños muestrales informativos (los pesos dependen de la variable objetivo) en los estimadores asistidos o basados en modelos.

Para ello, en cada una de las 500 muestras tipo ECV, se introdujo un mecanismo de falta de respuesta correlado con la renta del hogar. De esta manera, la probabilidad de responder de los hogares decrecía según su renta aumentaba. Después de generar la falta de respuesta en cada muestra según este mecanismo, los pesos muestrales finales dependían obviamente de la renta del hogar.

Para derivar las estimaciones del ingreso, la renta imponible del IRPF agregada a nivel de pequeña área (provincia o comarcas-EURAREA) ha sido utilizada como covariable en los modelos considerados los cuales, a su vez, han sido ajustados en base a toda la muestra con los siguientes procedimientos.

- Método 1: como en las simulaciones estándar
- Método 2: utilizando los pesos muestrales.

Es decir, se usaron las mismas covariables que antes pero diferentes pesos.

Para estudiar el efecto de los pesos informativos vamos a comparar los resultados de estos dos experimentos. En las tablas que vienen a continuación, se resumen los valores medios de ARB y RMSE:

Tabla 12.5. Valores medios del sesgo relativo ARB sobre las provincias como áreas pequeñas

Experimento	DIRECTO	GREG	EBLUPA3	EBLUPB
A_2 / sin pesos muestrales	-0,083	-0,063	-4,131	-1,382
A_2 / con pesos muestrales	-0,083	-0,061	0,054	0,806

Tabla 12.6.Valores medios del error relativo RMSE sobre las provincias como áreas pequeñas

Experimento	DIRECTO	GREG	EBLUPA3	EBLUPB
A_2 / sin pesos muestrales	5,031	3,708	4,983	4,707
A_2 / con pesos muestrales	5,031	3,692	3,397	5,394

Tabla 12.7. Valores medios del sesgo relativo ARB sobre las comarcas-EURAREA como áreas pequeñas

Experimento	DIRECTO	GREG	EBLUPA3	EBLUPB
A_2 / sin pesos muestrales	-0,254	-0,127	-2,754	-1,586
A_2 / con pesos muestrales	-0,254	0,086	0,868	2,116

Tabla 12.8. Valores medios del error relativo RMSE sobre las comarcas-EURAREA como áreas pequeñas

Experimento	DIRECTO	GREG	EBLUPA3	EBLUPB
A_2 / sin pesos muestrales	14,077	11,106	7,023	9,401
A_2 / con pesos muestrales	14,077	11,183	7,375	7,350

En general, nuestra conclusión es que el uso de los pesos muestrales informativos reduce el sesgo de los estimadores, tanto a nivel de provincia como a nivel de comarcas-EURAREA. Por otra parte, en relación al error relativo, esta reducción no es tan significativa excepto en el caso del estimador EBLUPB, basado en un modelo de área, aplicado para áreas pequeñas tipo comarcas-EURAREA.

13 Referencias

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under selection model. Biometrics, **31**, 423-427.

Morales D. y Molina I. (2002). Small area mixed linear models for normal variables. No publicado.

Morales D., Molina I. y Santamaría L. (2002). A comparative study of small area estimators with applications to surveys on income and living conditions. No publicado.

Pfefferman, D., Skinner, C.J., Holmes, D.J., Goldstein, H. y Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society, Series B, **60**, 23-40.

Prasad, N.G.N. y Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association, **85**, 163-171.

Saralegui, J. y Herrador, M. (2003). El problema de la estimación en áreas pequeñas para la estadística oficial. Recientes progresos en España. 27 Congreso Nacional de Estadística e Investigación Operativa.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). Finite Population Sampling and Inference. A Prediction Approach. John Wiley. New York.

Relación de participantes del Grupo de Trabajo en el Proyecto EURAREA

INSTITUTO NACIONAL DE ESTADÍSTICA

Jorge Saralegui (Coordinador)
Montserrat Herrador
Carlos Pérez
Florentina Alvarez
Ramiro López
Francisco Hernández

UNIVERSIDAD MIGUEL HERNÁNDEZ (UMH) DE ELCHE

Domingo Morales
Isabel Molina
Yolanda Marhuenda
Laureano Santamaría
Dolores Esteban
Angel Sánchez