ESTIMACIÓN DE ÁREAS PEQUEÑAS EN LA ENCUESTA DE POBLACIÓN EN RELACIÓN CON LA ACTIVIDAD DE LA C.A. DE EUSKADI

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

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

Presentación

Eustat, consciente de la creciente demanda de estadísticas de calidad cada vez más desagregadas, constituyó en 2003 un equipo de investigación compuesto por miembros de Eustat y de la Universidad. El objetivo era trabajar en la mejora de las técnicas de estimación en diferentes operaciones estadísticas, e introducir técnicas de estimación en áreas pequeñas basadas en modelos en la producción estadística. Un resultado de este trabajo fue la aplicación del sistema de estimación en áreas pequeñas a la Estadística Industrial de periodicidad anual, editado por Eustat en un Cuaderno Técnico en 2005.

Esta metodología de estimación se ha aplicado a otra operación estadística, igualmente relevante dentro de la producción de Eustat, la Encuesta de Población en Relación con la Actividad (PRA), que facilita a los usuarios resultados trimestrales sobre el mercado de trabajo en la C.A. de Euskadi a nivel de Territorio Histórico. Al igual que en la Estadística Industrial, las estimaciones basadas en modelos dan información de las 20 comarcas estadísticas en las que está dividida la C.A. de Euskadi.

El objetivo de esta publicación es aportar material útil a todos los usuarios interesados en el conocimiento y utilización de métodos en áreas pequeñas.

Este documento tiene dos partes diferenciadas. La primera, la metodología que se ha utilizado, con algunos aspectos propios sobre los estimadores y la información auxiliar empleada; y la segunda, la presentación de los resultados comarcales correspondientes a los años 2005, 2006 y 2007.

Vitoria-Gasteiz, marzo de 2008

JOSU IRADI ARRIETA

Director General

Índice

PRESENTACIÓN 2

ÍNDICE 3

1. INTRODUCCIÓN 4

2. LA ENCUESTA DE POBLACIÓN EN RELACIÓN CON LA ACTIVIDAD (PRA) 6

2.1 DESCRIPCIÓN DE LA ENCUESTA DE POBLACIÓN EN RELACIÓN CON LA ACTIVIDAD DE LA C.A. DE EUSKADI 6

2.2 ESTIMADORES UTILIZADOS EN LA ENCUESTA DE POBLACIÓN EN RELACIÓN CON LA ACTIVIDAD DE LA C.A. DE EUSKADI 7

3. SISTEMA DE ESTIMACIÓN EN ÁREAS PEQUEÑAS EN LA PRA 11

3.1 ESTUDIO DE SIMULACIÓN 11

3.2 CÁLCULO DE ESTIMACIONES EN MUESTRAS PRA REALES 20

3.3 ESTIMACIÓN DEL ERROR CUADRÁTICO MEDIO 21

3.4 SOFTWARE EMPLEADO 27

4. ESTIMACIONES COMARCALES 2005-2007 28

4.1 DEFINICIONES 28

4.2 RESULTADOS 30

5. CONCLUSIONES 34

6. BIBLIOGRAFÍA 35

ANEXO 37

Capítulo 1

Introducción

Actualmente la estadística oficial tiene que responder a una demanda de información de calidad, cada vez más desagregada sobre los principales indicadores sociales y económicos.

Un modo de afrontar esa demanda de desagregación es aumentar los tamaños de muestra, con todos los costes que esto conlleva, y seguir aplicando los estimadores basados en el diseño utilizados actualmente en la estadística oficial.

Otra alternativa, en vías de investigación, es utilizar técnicas de estimación más complejas, asistidas y basadas en modelos.

Eustat, consciente de esa creciente demanda de estadísticas de calidad cada vez más desagregadas, constituyó hace cinco años un equipo de investigación compuesto por miembros de Eustat y miembros de la Universidad para trabajar en la mejora de las técnicas de estimación en diferentes operaciones estadísticas e introducir técnicas de estimación en áreas pequeñas basadas en modelos.

El proyecto de estimación en áreas pequeñas comenzó con un curso de formación impartido por la Universidad dirigido a personal del Instituto y de la Organización Estadística Vasca. En diversas sesiones se abordó tanto la teoría de la estimación basada en el muestreo como la estimación asistida y basada en modelos.

La primera operación que se abordó en el proyecto fue la Estadística Industrial y el resultado se plasmó en un cuaderno técnico publicado en 2005. Aquel trabajo dio origen a la publicación periódica de estimaciones anuales en las 20 comarcas estadísticas de la C.A. de Euskadi de las principales magnitudes económicas de la encuesta.

El presente documento tiene como finalidad difundir los resultados de la segunda operación abordada según esta metodología en EUSTAT, la Encuesta de población en relación con la actividad (PRA).

Un antecedente en el ámbito europeo del uso de estimación basada en modelos en estadística oficial se encuentra en la Oficina Nacional de Estadística de Reino Unido (Office for National Statistics, ONS). Allí las estimaciones basadas en modelos de datos de desempleo en “local authority” o municipalidades, a partir de su encuesta de fuerza de trabajo (Labour Force Survey, LFS), han sido aceptadas recientemente como estadística nacional. Es la primera vez que se otorga este rango en Reino Unido a estimaciones basadas en modelos (CLARKE y otros, 2007).

Otras estimaciones en áreas pequeñas obtenidas en la ONS aún tienen el rango de estadística experimental, lo que significa que aún están sujetas a posibles mejoras metodológicas.

En general se están dando pasos en el ámbito internacional en la aceptación de las estimaciones en áreas pequeñas como estadística oficial, considerada como aquella que cumple todos los requisitos del Código de Buenas Prácticas de las estadísticas oficiales. Por un lado, esto implica nuevos retos para la investigación en estos métodos, y por otro, la presentación y explicación adecuada de estos resultados a los usuarios.

En este documento se van a presentar varios aspectos. En la parte teórica, hay dos partes: en primer lugar, se van a exponer las principales características de la Encuesta de Población en Relación con la Actividad, con los estimadores de resultados y de errores que se utilizan (capítulo 2); a continuación, se va a tratar el sistema de estimación de las áreas pequeñas que se ha aplicado a la PRA (capítulo 3).

En la parte aplicada se van a comentar los resultados obtenidos a partir de la mencionada encuesta, con esta metodología, para las comarcas de la C.A. de Euskadi. Se muestran resultados para las siguientes magnitudes: tasa de actividad, tasa de paro, población ocupada y parada, siempre para el colectivo de 16 y más años (capítulo 4). Finalmente, se extraen las conclusiones del trabajo (capítulo 5) y se muestra la Bibliografía. En Anexo, se detalla la división de municipios en comarcas de la C.A. de Euskadi.

Capítulo 2

La Encuesta de Población en Relación con la Actividad (PRA)

2.1 Descripción de la Encuesta de Población en Relación con la Actividad de la C.A. de Euskadi

La Encuesta de Población en Relación con la Actividad (PRA) se puso en marcha en los años 80, con el objetivo de disponer de información rica y detallada sobre el mercado laboral que además fuese comparable internacionalmente.

Más concretamente, la finalidad de la operación es producir información estadística continua sobre la participación, o no, de la población en las actividades laborales, con especial hincapié en las de carácter económico. Esta encuesta facilita resultados trimestrales y anuales sobre el volumen y las características de los principales colectivos desde el punto de vista del mercado de trabajo: la población activa, la ocupada y la parada, con sus correspondientes tasas de actividad, de ocupación y de paro.

Esta información se obtiene para las principales características demográficas a nivel de Territorio histórico, tal y como corresponde a su diseño muestral, que se comentará más adelante.

La población de referencia de la PRA es la residente en viviendas familiares en la C.A. de Euskadi. El marco de la encuesta es el Directorio de Viviendas y el Registro Estadístico de Población de la C.A. de Euskadi. El primer año para el cual hay datos completos de la PRA es 1985. Desde entonces la encuesta ha sufrido algunos cambios en el tamaño y diseño muestral, así como en el marco de muestreo y los tratamientos de elevación.

La muestra trimestral de la encuesta es un panel rotante de viviendas. Estas viviendas permanecen durante 8 trimestres, siendo contactadas una vez cada trimestre. La forma de renovar el panel es cambiar por viviendas nuevas un octavo de las viviendas del panel en cada trimestre.

La muestra inicial (la última que se extrajo en su totalidad corresponde al primer trimestre de 2005) se compone de 12 submuestras independientes y sistemáticas, una por cada semana del trimestre. En total, se trata de 5.088 viviendas, que se distribuyen por Territorios Históricos de la siguiente forma: 1.114 para Álava, 2.196 para Bizkaia y 1.690 para Gipuzkoa.

Este reparto de viviendas se realiza de modo proporcional a la raíz cuadrada del número de viviendas de los Territorios Históricos, para reducir las diferencias en el tamaño de población entre ellos. Para garantizar este reparto, los Territorios Históricos forman los estratos de la muestra.

Todos los trimestres se hace una renovación parcial del panel. Esta renovación consiste en extraer una muestra equivalente a 1/8 de la muestra trimestral, siguiendo las mismas pautas, muestreo sistemático de viviendas en los estratos e introducirla en el panel, para extraer las viviendas que ya han estado 8 trimestres en la encuesta.

En la encuesta hay dos tipos de unidades: una, las viviendas que forman el panel; y dos, los individuos que residen en esas viviendas. Durante el trabajo de campo, se recoge información de todos los residentes en las viviendas, normalmente a través de un informante de dicha vivienda. Los resultados de la encuesta son relativos a la población y a las viviendas o familias.

El paso de los datos muestrales a las estimaciones se realiza después de un proceso de elevación o calibración. En este proceso se calculan unos elevadores o pesos para individuos y familias, en función de las proyecciones de ambos tipos.

2.2 Estimadores utilizados en la Encuesta de Población en Relación con la Actividad de la C.A. de Euskadi

2.2.1 Definición de los estimadores y fórmulas de elevación

2.2.1.1 Totales de personas

En cada estrato h determinado por el territorio, se obtienen estimadores de calibración en dos fases:

En primer lugar, se aplica el estimador de Horvitz-Thompson , a partir del cálculo del factor de diseño o inverso de la probabilidad de selección de cada vivienda (todas las viviendas en el estrato tienen la misma probabilidad de selección). Se seleccionan todas las personas dentro de la vivienda por lo que su factor de diseño coincide con el de la vivienda. No hay falta de respuesta parcial, esto es, de personas dentro de la vivienda, así que no hay corrección por falta de respuesta.

$$\hat{X}_h = \sum_{i=1}^{v_h}\sum_{j=1}^{n_{v_i}} w_{hi} X_{hij} = \sum_{i=1}^{v_h} w_{hi} \sum_{j=1}^{n_{v_i}} X_{hij} = w_h \sum_{i=1}^{v_h}\sum_{j=1}^{n_{v_i}} X_{hij} = \frac{V_h}{v_h}\sum_{i=1}^{v_h}\sum_{j=1}^{n_{v_i}} X_{hij}$$

donde:

v_h es el número de viviendas en la muestra efectiva del estrato h

n_{v_i} es el número de personas muestreadas en la vivienda i

$w_{hi} = w_h$, factor de diseño de la vivienda i

V_h es el número de viviendas en la población del estrato h

X_{hij} es el valor de la característica a estimar de la persona j de la vivienda i del estrato h

Después, se lleva a cabo un ajuste de los pesos anteriores por calibración.

$$\hat{X}_h^* = \sum_{i=1}^{v_h} \sum_{j=1}^{n_{vi}} w_{hij}^* X_{hij}$$

con w_{hij}^* obtenidos a partir de los factores iniciales $w_{hij} = w_{hi}$ aplicando post-estratificación según la variable cruce del territorio, grupo de edad (7 grupos: <=15,16-24, 25-34,.., 55-64, >=65) y sexo.

Se utilizan las proyecciones de población correspondientes a estos grupos de edad, sexo y Territorio Histórico elaboradas por Eustat.

2.2.1.2 Totales de familias

En cada estrato h determinado por el territorio, se obtienen estimadores de calibración en dos fases:

En primer lugar, se aplica el estimador de Horvitz-Thompson, a partir del cálculo del factor de diseño ó inverso de la probabilidad de selección de cada vivienda (todas las viviendas en el estrato tienen la misma probabilidad de selección).

$$\hat{X}_h = \sum_{i=1}^{v_h} w_{hi} X_{hi} = w_h \sum_{i=1}^{v_h} X_{hi} = \frac{V_h}{v_h} \sum_{i=1}^{v_h} X_{hi}$$

donde:

v_h es el número de viviendas en la muestra efectiva del estrato h

$w_{hi} = w_h$, factor de diseño de la vivienda i

V_h es el número de viviendas en la población del estrato h

X_{hi} es el valor de la característica a estimar de la vivienda i del estrato h

A continuación, se hace un ajuste de los pesos por calibración.

$$\hat{X}_h^* = \sum_{i=1}^{v_h} w_{hi}^* X_{hi}$$

con w_{hi}^* obtenidos a partir de los factores iniciales w_{hi} aplicando un ajuste simultáneo a las distribuciones marginales de las variables siguientes: territorio y tamaño familiar (5 categorías: 1, 2, 3,…, >=5 miembros).

La información auxiliar del número de familias por territorio y por tamaño familiar se obtiene de una forma de proyección de familias. Son estimaciones obtenidas a partir de las proyecciones de población elaboradas por Eustat, y del tamaño medio familiar proveniente del Padrón Municipal de Habitantes.

Ambos ajustes se realizan con la macro Calmar de SAS (Institut National de la Statistique et des Études Économiques), aplicando el método "raking ratio" (INSEE 1993).

2.2.2 Método de estimación de los errores de muestreo

El método utilizado es el Método de Expansión de Taylor. Permite calcular estimaciones del error muestral para totales, medias y ratios en muestras con estratificación, clústers y probabilidades desiguales. El método obtiene aproximaciones lineales del estimador y calcula su varianza utilizando ésta como estimación de la varianza muestral.

La expresión para el cálculo de la varianza estimada para la media poblacional es la siguiente:

(2)

$$\overline{V}\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} \left(e_{hi\cdot} - \bar{e}_{h\cdot\cdot}\right)^2$$

Donde:

$$e_{hi\cdot} = \left(\sum_{j=1}^{m_{hi}} w_{hij} \left(y_{hij} - \hat{\bar{Y}} \right) \right) \Big/ w_{\cdots}$$

$$\bar{e}_{h\cdot\cdot} = \left(\sum_{i=1}^{n_h} e_{hi\cdot} \right) \; / \; n_h$$

y

$$w_{\cdots} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$$

Notación:

$h = 1, 2, \ldots, H$ indica el estrato con un total de H estratos.
$i = 1, 2, \ldots n_h$ indica el número de clusters en el estrato h, con un total de n_h clusters.
$j = 1, 2, \ldots, m_{hi}$ indica el número de unidad dentro del cluster i del estrato h, con un total de m_{hi} unidades

$$n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$$

es el número total de observaciones en la muestra.

w_{hij} indica el elevador de la observación j en el cluster i del estrato h

$Y_{hij} = (Y_{hij}(1), Y_{hij}(2), \ldots, Y_{hij}(P))$ son los valores observados de la variable Y en la observación j del cluster i del estrato h. (variables numéricas y categóricas).

Este cálculo se realiza con el procedimiento PROC SURVEYMEANS del paquete estadístico SAS (Sas Institute Inc. 2004).

Capítulo 3

Sistema de Estimación en Áreas Pequeñas en la PRA

3.1 Estudio de simulación

La metodología de estimación se ha establecido a partir de un estudio de simulación realizado sobre la información poblacional disponible en el último censo. Se ha analizado el rendimiento de diversos estimadores tanto clásicos como asistidos y basados en modelos para estimar el número de parados, por sexo, en 20 y en 40 comarcas de la C. A. de Euskadi.

Para ello se ha evaluado el error cuadrático medio de los estimadores calculados utilizando simulaciones a partir del Censo de Población y Viviendas (en adelante Censo ó CPV) 2001 por el mismo procedimiento de muestreo que se utiliza en la PRA a partir de 2005.

Se han obtenido por simulación 500 muestras a partir del Censo 2001 según el diseño de la PRA.

3.1.1 Indicadores de evaluación.

Para evaluar el sesgo y el error cuadrático medio de los estimadores a partir de 500 muestras simuladas del Censo 2001 se han calculado los siguientes indicadores:

Sesgo relativo absoluto (SRA):

$$SRA_d(\hat{y}) = \frac{1}{K}\left|\sum_{k=1}^{K}\frac{\hat{y}_d(k)-Y_d}{Y_d}\right|100 \quad \text{con d la zona ó area pequeña}$$

Y su media: $SRAM(\hat{y}) = \frac{1}{D}\sum_d SRA_d(\hat{y})$

Raíz cuadrada del error cuadrático medio relativo (REMC):

$$REMC_d(\hat{y}) = \left(\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\hat{y}_d(k)-Y_d}{Y_d}\right)^2\right)^{\frac{1}{2}}100$$

Y su media: $REMCM(\hat{y}) = \frac{1}{D}\sum_{d} EMC_d(\hat{y})$

Se han evaluado:

- Estimadores basados en el diseño
- Estimadores asistidos en modelos
- Estimadores basados en modelos

3.1.2 Estimadores basados en el diseño.

- Directo:

$$\hat{y}_d^{directo} = \frac{\sum_{j=1}^{n_d} w_j y_j}{\sum_{j=1}^{n_d} w_j} N_d$$

donde $y_j = 1$ (parado) $yj = 0$ (no parado)

N_d número de personas con edad > 16 en la zona d

n_d tamaño muestral en la zona d

d es el área pequeña (zona)

w_j peso de diseño

- Posestratificado

$$\hat{y}_d^{post} = \sum_{g} \hat{\bar{y}}_{dg} N_{dg}$$

donde $\hat{\bar{y}}_{dg}$ es la media calculada con el estimador directo anterior, $\hat{\bar{y}}_{dg} = \frac{\sum_{j \in s_{dg}} w_j y_j}{\sum_{j \in s_{dg}} w_j}$

- Sintético

$$\hat{y}_d^{\sin t} = \sum_{g} \hat{\bar{y}}_{g} N_{dg}$$

donde $\hat{\bar{y}}_g$ es la media calculada con el estimador directo anterior

- Compuesto 1 - Compuesto 4

$$\hat{y}_d^{dep} = \lambda_d \hat{y}_d^{post} + (1-\lambda_d)\hat{y}_d^{\sin t}$$

donde $0 \leq \lambda_d \leq 1$ viene dada por

$$\lambda_d = \begin{cases} 1 \text{ si } \hat{N}_d \geq \alpha N_d \\ \dfrac{\hat{N}_d}{\alpha N_d} \text{ en otro caso} \end{cases}$$

$\hat{N}_d = \sum_d w_j$ es el total poblacional estimado en cada área d y α es un parámetro. Evaluamos el estimador compuesto para distintos valores de $\alpha = 2/3, 1, 1.5$ y 2.

Se calculan los estimadores anteriores para los siguientes grupos g:

- Grupo de edad y sexo. (8 categorías)
- Grupo de edad y sexo*Nivel educativo (24 categorías)
- Grupo de edad y sexo* Relación con la actividad en CPV1996 (24 categorías)
- Grupo de edad y sexo*Nivel educativo* Relación con la actividad en CPV1996 (72 categorías)

3.1.3 Estimadores asistidos en modelos.

Estimador GREG asistido en un modelo lineal

Se modeliza la probabilidad de estar parado, por ello los valores obtenidos deberían estar en el intervalo $[0,1]$. Esto no se controla en un modelo lineal, así que si se obtiene una predicción negativa se transforma en 0 y en caso de que sea mayor que 1 se transforma a 1.

Suponemos las variables binarias y_{id} que indican si el individuo i-ésimo del área d está parado ó no.

El modelo lineal es el siguiente:

$$p_{id} = x_{id}^T \beta + \varepsilon_{id}$$

donde:

p_{id} es la probabilidad de que el individuo i-ésimo del área d esté parado

$x_{id} = (x_{id,1}, x_{id,2}, ..., x_{id,p})^T$ es el vector de p covariables, donde dichas covariables $x_{id,k}$ pueden ser las siguientes:

Grupo de edad y sexo

Nivel educativo

Relación con la actividad en CPV1996

$\varepsilon_{id} \approx N\left(0,\sigma^2\right)$ son variables aleatorias independientes.

El estimador del total en cada área viene dado por:

$$\hat{Y}_d^{GREG} = N_d\left(\frac{1}{\hat{N}_d}\sum_{i\in s_d} w_{id}\,y_{id} + \left(\overline{X}_d - \frac{1}{\hat{N}_d}\sum_{i\in s_d} w_{id}\,x_{id}\right)^T \hat{\beta}\right) = N_d\left(\overline{X}_d\,\hat{\beta} + \frac{1}{\hat{N}_d}\sum_{i\in s_d} w_{id}\left(y_{id} - x_{id}^T\hat{\beta}\right)\right)$$

donde $\overline{X}_d = \left(\overline{X}_{d1}, \overline{X}_{d2},..., \overline{X}_{dp}\right)^T$ es el vector de medias poblacional de las p covariables y $\hat{\beta} = \left(\sum_{i\in s} w_i x_i x_i^T\right)^{-1} \sum_{i\in s} w_i x_i y_i$

Nota: Un estimador GREG asistido en un modelo lineal con efecto fijo de la comarca coincide con el estimador basado en un modelo lineal con pesos y efecto fijo de la comarca. Por ello no se introduce aquí la comarca fija como variable explicativa.

Estimador GREG asistido en un modelo logit.

Suponemos las variables binarias y_{id} que indican si el individuo i-ésimo del área d está parado ó no.

El modelo logit es el siguiente:

$$\log it\left(p_{id}\right) = \log\left(\frac{p_{id}}{1-p_{id}}\right) = x_{id}^T\beta$$

donde:

p_{id} es la probabilidad de que el individuo i-ésimo del área d esté parado

$x_{id} = \left(x_{id,1}, x_{id,2},..., x_{id,p}\right)^T$ es el vector de p covariables, donde dichas covariables $x_{id,k}$ pueden ser las siguientes:

Grupo de edad y sexo

Nivel educativo

Relación con la actividad en CPV1996

El estimador del total en cada área viene dado por:

$$\hat{Y}_d^{GREG} = \sum_{i=1}^{N_d} \frac{e^{x_{id}^T \hat{\beta}}}{1+e^{x_{id}^T \hat{\beta}}} + \frac{N_d}{\hat{N}_d} \sum_{i \in s_d} w_{id} \left(y_{id} - \frac{e^{x_{id}^T \hat{\beta}}}{1+e^{x_{id}^T \hat{\beta}}} \right)$$

donde $\hat{\beta}$ se calcula con pesos w_i.

Nota: un estimador GREG asistido en un modelo logit con efecto fijo de la comarca coincide con un estimador basado en un modelo logit con pesos y efecto fijo de comarca. Por ello no se introduce la comarca fija como variable explicativa.

Estimador GREG asistido en un modelo logit mixto.

Suponemos las variables binarias y_{id} que indican si el individuo i-ésimo del área d está parado ó no.

El modelo logit con efecto aleatorio del área es el siguiente:

$$\log\left(\frac{p_{id}}{1-p_{id}} \right) = x_{id}^T \beta + u_d$$

donde:

y_{id} verifica que $y_{id} | u_d \approx Binomial(1, p_{id})$

p_{id} es la probabilidad de que el individuo i-ésimo del área d esté parado

u_d es el efecto aleatorio del área, $u_d \approx N(0, \sigma_u^2)$

$x_{id} = (x_{id,1}, x_{id,2}, ..., x_{id,p})^T$ es el vector de p covariables, donde dichas covariables $x_{id,k}$ pueden ser las siguientes:

Grupo de edad y sexo

Nivel educativo

Relación con la actividad en CPV1996

El estimador del total en cada área viene dado por:

$$\hat{Y}_d^{GREGMixto} = \sum_{i=1}^{N_d} \frac{e^{x_{id}^T \hat{\beta} + \hat{u}_d}}{1+e^{x_{id}^T \hat{\beta} + \hat{u}_d}} + \frac{N_d}{\hat{N}_d} \sum_{i \in s_d} w_{id} \left(y_{id} - \frac{e^{x_{id}^T \hat{\beta} + \hat{u}_d}}{1+e^{x_{id}^T \hat{\beta} + \hat{u}_d}} \right)$$

donde $\hat{\beta}$ se calcula sin pesos.

En algunas ocasiones el componente de varianza se estima como cero, además cuando no se estima como cero no siempre es estadísticamente significativo.

3.1.4 Estimadores basados en modelos.

3.1.4.1 Estimadores basados en un modelo lineal.

Cuando se plantea un modelo lineal no se controla que la predicción obtenida esté en el intervalo [0,1]. Se está modelizando la probabilidad de estar parado, por ello los valores obtenidos deberían estar entre cero y uno. En caso de que se obtenga una predicción negativa se transforma en 0 y en el caso de que sea mayor que 1 se transforma en 1.

***Estimador sintético basado en un modelo lineal*.**

Suponemos las variables binarias y_{id} que indican si el individuo i-ésimo del área d está parado ó no.

El modelo es el siguiente:

$$p_{id} = x_{id}^T \beta + \varepsilon_{id}$$

donde:

p_{id} es la probabilidad de que el individuo i-ésimo del área d esté parado

$x_{id} = \left(x_{id,1}, x_{id,2}, ..., x_{id,p}\right)^T$ es el vector de p covariables, donde dichas covariables $x_{id,k}$ pueden ser las siguientes:

Grupo de edad y sexo

Nivel educativo

Relación con la actividad en CPV1996

$\varepsilon_{id} \approx N\left(0, \sigma^2\right)$ son variables aleatorias independientes.

El estimador (proyectivo) del total en cada área viene dado por:

$$\hat{P}_d = X_d \hat{\beta}$$

donde $X_d = \left(X_{d1}, X_{d2}, ..., X_{dp}\right)^T$ es el vector del total poblacional de las p covariables, y

$$\hat{\beta} = \left(\sum_{i \in s} x_i x_i^T \right)^{-1} \sum_{i \in s} x_i y_i$$

***Estimador sintético basado en un modelo lineal con pesos*.**

El estimador sintético basado en un modelo lineal con pesos coincide con el anterior, la diferencia está en el estimador de $\hat{\beta}$. Este último se calcula con pesos w_i.

$$\hat{\beta} = \left(\sum_{i \in s} w_i x_i x_i^T \right)^{-1} \sum_{i \in s} w_i x_i y_i$$

***Estimador basado en un modelo lineal con efecto fijo del área*.**

Suponemos las variables binarias y_{id} que indican si el individuo i-ésimo del área d está parado ó no.

El modelo es el siguiente:

$$p_{id} = x_{id}^T \beta + \varepsilon_{id}$$

donde:

p_{id} es la probabilidad de que el individuo i-ésimo del área d esté parado

$x_{id} = \left(x_{id,1}, x_{id,2}, ..., x_{id,p} \right)^T$ es el vector de p covariables, donde $x_{id,1}$ es el área que se incluye como efecto fijo y además puede incluir cualquiera de las covariables ya mencionadas en el apartado anterior.

El estimador (proyectivo) del total en cada área viene dado por:

$$\hat{P}_d = \sum_{i=1}^{N_d} \hat{p}_{id} = X_d \hat{\beta}$$

donde $X_d = \left(X_{d1}, X_{d2}, ..., X_{dp} \right)^T$ es el vector del total poblacional de las p covariables, y

$$\hat{\beta} = \left(\sum_{i \in s} x_i x_i^T \right)^{-1} \sum_{i \in s} x_i y_i$$

Estimador basado en un modelo lineal con pesos y efecto fijo del área.

El estimador basado en un modelo lineal con pesos y efecto fijo del área coincide con el anterior, la diferencia está en el estimador de $\hat{\beta}$. Este último se calcula con pesos w_i.

$$\hat{\beta} = \left(\sum_{i \in s} w_i x_i x_i^T \right)^{-1} \sum_{i \in s} w_i x_i y_i$$

Estimador sintético basado en un modelo lineal con efecto aleatorio del área.

Se realizaron varias simulaciones y en algunas de ellas resultó imposible ajustar un modelo lineal con efecto aleatorio del área. Luego se ha desechado la posibilidad de utilizar este modelo.

3.1.4.2 Estimadores basados en un modelo logit.

***Estimador sintético basado en un modelo logit*.**

Suponemos las variables binarias y_{id} que indican si el individuo i-ésimo del área d está parado ó no.

El modelo logit es el siguiente:

$$\log it(p_{id}) = \log\left(\frac{p_{id}}{1 - p_{id}} \right) = x_{id}^T \beta$$

donde:

p_{id} es la probabilidad de que el individuo i-ésimo del área d esté parado

$x_{id} = (x_{id,1}, x_{id,2}, ..., x_{id,p})^T$ es el vector de p covariables, donde dichas covariables $x_{id,k}$ pueden ser las siguientes:

Grupo de edad y sexo

Nivel educativo

Relación con la actividad en CPV1996

El estimador del total en cada área viene dado por:

$$\hat{P}_d = \sum_{i=1}^{N_d} \frac{e^{x_{id}^T \hat{\beta}}}{1 + e^{x_{id}^T \hat{\beta}}}$$

***Estimador sintético basado en un modelo logit con pesos*.**

El estimador sintético basado en un modelo logit con pesos coincide con el anterior, la diferencia está en el estimador de $\hat{\beta}$. Este último se calcula con pesos w_i.

Estimador basado en un modelo logit con efecto fijo del área.

Suponemos las variables binarias y_{id} que indican si el individuo i-ésimo del área d está parado ó no.

El modelo logit es el siguiente:

$$\log it(p_{id}) = \log\left(\frac{p_{id}}{1-p_{id}}\right) = x_{id}^T \beta$$

donde:

p_{id} es la probabilidad de que el individuo i-ésimo del área d esté parado

$x_{id} = (x_{id,1}, x_{id,2}, ..., x_{id,p})^T$ es el vector de p covariables, donde $x_{id,1}$ es el Área que se incluye como efecto fijo y además puede incluir cualesquiera de las covariables mencionadas anteriormente.

El estimador del total en cada área d viene dado por:

$$\hat{P}_d = \sum_{i=1}^{N_d} \frac{e^{x_{id}^T \hat{\beta}}}{1+e^{x_{id}^T \hat{\beta}}}$$

***Estimador basado en un modelo logit con pesos y efecto fijo del área*.**

El estimador basado en un modelo logit con pesos y efecto fijo del área coincide con el anterior, la diferencia está en el estimador de $\hat{\beta}$. Este último se calcula con pesos w_i.

Estimador basado en un modelo logit con efecto aleatorio del área (EB Logit Mixto).

Suponemos las variables binarias y_{id} que indican si el individuo i-ésimo del área d está parado ó no.

El modelo logit con efecto aleatorio del área es el siguiente:

$$\log it(p_{id}) = \log\left(\frac{p_{id}}{1-p_{id}}\right) = x_{id}^T \beta + u_d$$

donde:

y_{id} verifica que $y_{id}|u_d \approx Binomial(1, p_{id})$

p_{id} es la probabilidad de que el individuo i-ésimo del área d esté parado

u_d es el efecto aleatorio del área, $u_d \approx N(0, \sigma_u^2)$

$x_{id} = (x_{id,1}, x_{id,2}, ..., x_{id,p})^T$ es el vector de p covariables, donde dichas covariables $x_{id,k}$ pueden ser las siguientes:

Grupo de edad y sexo

Nivel educativo

Relación con la actividad en CPV1996

El estimador (EB Empírico-Bayesiano) del total en cada área es:

$$\hat{P}_d = \sum_{i=1}^{N_d} \frac{e^{x_{id}^T \hat{\beta} + \hat{u}_d}}{1 + e^{x_{id}^T \hat{\beta} + \hat{u}_d}}$$

3.1.5 Conclusiones

Una vez realizado el estudio con todos los estimadores mencionados, el estimador más adecuado por comarcas ha resultado ser, en términos de los indicadores de evaluación descritos, el compuesto 4, que incluye como variables auxiliares el sexo y edad, el nivel educativo y la relación con la actividad en el período anterior.

Los resultados obtenidos para 40 y para 20 comarcas no difieren considerablemente, ya que las zonas con estimaciones más sensibles permanecen en ambas clasificaciones.

3.2 Cálculo de estimaciones en muestras PRA reales

El estimador seleccionado, compuesto 4, se ha aplicado en muestras reales de la PRA de todos los trimestres.

Para ello fue necesario tener la información auxiliar para cada una de las unidades de la muestra. El grupo de variables edad y sexo y el nivel educativo estaban disponibles para prácticamente la totalidad de la muestra. Para obtener la relación con la actividad en el período anterior, disponible en este caso en el Censos de Población y Viviendas (en adelante Censo) 2001, era necesaria la fusión de los registros de la PRA con los del Censo. En este paso, no se consiguió identificar correctamente aproximadamente al 12% de la muestra.

Otra información precisada para la estimación de las comarcas son las proyecciones poblacionales a ese nivel según las variables auxiliares utilizadas:

grupo de edad y sexo, nivel educativo y relación con la actividad en el período anterior. Los tamaños poblacionales se calcularon a partir del Censo 2001 y se calibraron a las proyecciones de población por grupo de edad, sexo y territorio histórico calculadas trimestralmente por Eustat.

Las estimaciones por comarcas obtenidas para cada uno de los estimadores, nuevamente se calibraron a las estimaciones proporcionadas por el estimador directo por territorio histórico. De esta manera, se obtienen las estimaciones calibrados los datos ya publicados trimestralmente. Además, con el fin de obtener estimaciones más estables a lo largo del tiempo, se han calculado medias móviles cada 4 trimestres.

3.3 Estimación del error cuadrático medio

3.3.1 Procedimientos para el cálculo del error cuadrático medio (MSE)

De igual manera se ha realizado un estudio de simulación para analizar el comportamiento de tres métodos de estimación de los correspondientes errores cuadráticos medios, el método de linealización de la varianza y los métodos de remuestreo siguientes: el método jacknife y el método bootstrap. Se estudia el comportamiento del estimador del MSE utilizando los tres métodos. Para ello se ha analizado si dichos estimadores del MSE proporcionan resultados similares a los obtenidos en el estudio de simulación.

Se evaluaron los 3 métodos para el estimador compuesto 4 y los estimadores postestratificado y sintético.

El método de linealización de la varianza consiste en la aplicación de un desarrollo en serie de Taylor.

Los métodos de remuestreo se basan en la evaluación de los estadísticos en remuestras ó submuestras obtenidas a partir de los datos originales, y mediante esos valores se obtienen estimadores de las medidas de exactitud ó de la distribución muestral del estadístico.

En el caso del método jacknife se dispone de tantas submuestras como clusters tenga la muestra, ya que se obtienen por sucesivas eliminaciones de clusters en la muestra original. Para cada submuestra se definen nuevos pesos y se calcula el estimador (postestratificado, sintético ó compuesto). Luego se obtiene la varianza y sesgo de los estimadores como se detalla posteriormente.

En el método bootstrap, las submuestras se obtienen mediante muestreo aleatorio simple, pero se ha de determinar cuántas son necesarias. De forma análoga, para cada submuestra se definen nuevos pesos y se calcula cada estimador. Con

dichas estimaciones se obtiene el error cuadrático medio como se detalla posteriormente.

En lo que sigue se utilizan los siguientes índices:

- h es el número de estrato, donde $h = 1,2,..,H$
- i es el cluster i-ésimo en el estrato h, donde $i = 1,2,..,n_h$
- j es la unidad j-ésima del cluster i en el estrato h, donde $j = 1,2,..,m_{hi}$

En la PRA, el estrato h es el territorio histórico ó la provincia, el cluster i es la vivienda y j es la persona j-ésima de la vivienda.

3.3.1.1 Método de linealización de la varianza.

El método de linealización ó método delta consiste en aplicar un desarrollo en serie de Taylor.

Definimos para cada dominio D las siguientes variables indicadoras:

$$I_D(h,i,j) = \begin{cases} 1 & \text{si } (h,i,j) \text{ está en } D \\ 0 & \text{en otro caso} \end{cases}$$

$$z_{hij} = y_{hij} I_D(h,i,j) = \begin{cases} y_{hij} & \text{si } (h,i,j) \text{ está en } D \\ 0 & \text{en otro caso} \end{cases}$$

$$v_{hij} = w_{hij} I_D(h,i,j) = \begin{cases} w_{hij} & \text{si } (h,i,j) \text{ está en } D \\ 0 & \text{en otro caso} \end{cases}$$

El estimador de la media en un dominio D viene dado por la expresión:

$$\hat{\bar{Y}}_D = \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij} z_{hij} \right) \Big/ v..., \text{ donde } v... = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij}$$

La varianza linealizada del estimador de la media en un dominio D viene dada por:

$$Var_L\left(\hat{\bar{Y}}_D\right) = \sum_{h=1}^{H} Var_h\left(\hat{\bar{Y}}_D\right) \quad \text{donde} \quad Var_h\left(\hat{\bar{Y}}_D\right) = \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(U_{hi.} - \bar{U}_{h..}\right)^2$$

$$U_{hi.} = \frac{1}{v...}\sum_{j=1}^{m_{hi}} v_{hij}\left(z_{hij} - \hat{\bar{Y}}_D\right) \quad \text{y} \quad \overline{U}_{h..} = \frac{1}{n_h}\sum_{i=1}^{n_h} U_{hi.}$$

3.3.1.2 Método jacknife para la estimación de la varianza y del sesgo.

Para aplicar el método jacknife al esquema de muestreo utilizado en la PRA se ha de eliminar un cluster (vivienda) cada vez. Definimos unos nuevos pesos, dados por:

$$w_{j(hi)} = \begin{cases} w_{hij} & \text{si la unidad j no está en el estrato h} \\ 0 & \text{si la unidad j está en el cluster i del estrato h} \\ \dfrac{n_h}{n_h - 1} w_{hij} & \text{si la unidad j está en el estrato h pero no en el cluster i} \end{cases}$$

Sea :

- $\hat{\theta}$ el estimador compuesto 4 obtenido con los datos de una simulación utilizando pesos w_{hij}
- $\hat{\theta}_{(hi)}$ el estimador compuesto 4 obtenido con los datos de la submuestra resultante de eliminar el cluster i (vivienda) del estrato h (th) de dicha simulación y utilizando los pesos $w_{j(hi)}$

El estimador jacknife de la varianza en el estrato h viene dado por:

$$Var_{JK(h)}\left(\hat{\theta}\right) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} \left[\hat{\theta}_{(hi)} - \hat{\theta}_{(h.)}\right]^2$$

donde:

$$\hat{\theta}_{(h.)} = \frac{1}{n_h}\sum_{i=1}^{n_h} \hat{\theta}_{(hi)}$$

El estimador jacknife del sesgo de un estimador en el estrato h viene dado por:

$$Sesgo_{JK(h)}\left(\hat{\theta}\right) = \left(n_h - 1\right)\left(\hat{\theta}_{(h.)} - \hat{\theta}\right)$$

El estimador jacknife del MSE en el estrato h:

$$\hat{MSE}_{JK(h)}\left(\hat{\theta}\right)=\hat{Var}_{JK(h)}\left(\hat{\theta}\right)+\hat{Sesgo}^2_{JK(h)}\left(\hat{\theta}\right)$$

Como los estratos son independientes el MSE del estimador viene dado por:

$$\hat{MSE}_{JK}\left(\hat{\theta}\right)=\sum_{h=1}^{H}\left[\frac{n_h-1}{n_h}\sum_{i=1}^{n_h}\left[\hat{\theta}_{(hi)}-\hat{\theta}_{(h.)}\right]^2+\left(\left(n_h-1\right)\left(\hat{\theta}_{(h.)}-\hat{\theta}\right)\right)^2\right]$$

Se ha calculado también el MSE directamente como sigue:

$$\hat{MSE}_{JK}\left(\hat{\theta}\right)=\sum_{h=1}^{H}\left[\frac{n_h-1}{n_h}\sum_{i=1}^{n_h}\left[\hat{\theta}_{(hi)}-\hat{\theta}\right]^2\right]$$

No se han observado diferencias apreciables.

3.3.1.3 Estimación bootstrap del error cuadrático medio.

A continuación se describen los pasos a dar para construir la versión del bootstrap reescalado en un muestreo simple estratificado propuesta por Rao y Wu (1988).

1. Fijado un estrato h, tenemos una muestra con n_h clusters. Extraemos una submuestra con n_h-1 clusters mediante muestreo aleatorio simple con reemplazamiento de la muestra del estrato h. Repetimos este proceso de forma independiente en cada estrato.

2. Para cada submuestra r (r=1,2,..R) construimos un nuevo peso: $w_{hij}\left(r\right)=w_{hij}\frac{n_h}{n_h-1}m_i\left(r\right)$ donde $m_i\left(r\right)$ es el número de veces que el cluster i es seleccionado en la submuestra, y calculamos $\hat{\theta}^*_r$ utilizando los nuevos pesos $w_{hij}\left(r\right)$.

3. Repetimos los pasos 1 y 2, R veces.

4. Para obtener el estimador bootstrap del error cuadrático medio realizamos: $\hat{MSE}_B\left(\hat{\theta}\right)=\frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\theta}^*_r-\hat{\theta}\right)^2$ donde:

- $\hat{\theta}$ es el estimador compuesto 4 obtenido con los datos de una simulación utilizando pesos w_{hij} .

- $\hat{\theta}_r^*$ es el estimador compuesto 4 obtenido con los datos de la submuestra r utilizando los pesos $w_{hij}(r)$.

Una de las cuestiones a decidir es el tamaño de R para que el método funcione correctamente. Para ello se realizan 5 simulaciones con los estimadores postestratificado, sintético y compuesto 4 utilizando como variables auxiliares el grupo de edad y sexo, nivel educativo, y relación con la actividad en el período anterior. Se consideran distintos valores de R, en concreto R toma los valores 200, 500, 1000 y 2000. Se observa que no existen diferencias de comportamiento según los tamaños de R. A la vista de los resultados se opta por utilizar R=200.

3.3.2 Comparación de los resultados obtenidos con los de la simulación.

Se estudia el comportamiento del estimador del MSE utilizando los tres métodos. Para ello se analiza si dichos estimadores del MSE proporcionan resultados similares a los obtenidos en el estudio de simulación.

El indicador de evaluación utilizado en la simulación para evaluar el error cuadrático medio de los estimadores viene dado por:

$$REMCR_d(\hat{y}_d) = \left(\frac{1}{K} \sum_{k=1}^{K} \left(\frac{\hat{y}_d(k) - Y_d}{Y_d} \right)^2 \right)^{\frac{1}{2}} 100$$

Este indicador REMCR (raíz cuadrada del error cuadrático medio relativo) se ha obtenido con 500 muestras simuladas extraídas del Censo 2001. A partir de ese indicador se define un nuevo indicador que se aproxima a la raíz cuadrada del error cuadrático medio $RMSE\_S_d(\hat{y}_d)$ (raíz cuadrada del error cuadrático medio obtenido por simulación). Viene dado por:

$$RMSE\_S_d(\hat{y}_d) = \frac{Y_d}{100} RECMR_d(\hat{y}_d)$$

Este error es una aproximación de Monte Carlo al verdadero error.

Para evaluar el comportamiento de cualesquiera de los métodos de estimación del MSE, calculamos el MSE del estimador utilizando las mismas 500 muestras obtenidas por simulación.

3.3.3 Aplicación de los métodos de estimación del MSE a muestras PRA reales. Calibración.

Trabajando con cada una de las muestras reales de la PRA obtenemos las estimaciones por comarca para cada uno de los estimadores. Estas estimaciones son

calibradas a las estimaciones proporcionadas por el estimador directo por territorio histórico que difunde Eustat trimestralmente.

Los métodos de cálculo de los MSE de las estimaciones calibradas son:

- **bootstrap**: para obtener la estimación bootstrap del error cuadrático medio se aplica $\hat{MSE}_B\left(\hat{y}_d\right)=\frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{y}^*_{d(r)}-\hat{y}_d\right)^2$ donde:
  - $\hat{y}_d$ es el estimador obtenido con los datos de la muestra de un trimestre concreto, calibrando las estimaciones a los valores que proporcionaría el estimador directo por th.
  - $\hat{y}^*_{d(r)}$ es el estimador obtenido con los datos de la submuestra r con pesos $w_{hij}\left(r\right)$, calibrando las estimaciones a los valores que proporcionaría el estimador directo por th si utilizase los datos de esa submuestra r.
- **jacknife**: para obtener la estimación jacknife del error cuadrático medio se aplica $\hat{MSE}_{JK}\left(\hat{y}_d\right)=\sum_{h=1}^{H}\left[\frac{n_h-1}{n_h}\sum_{i=1}^{n_h}\left[\hat{y}_{d(hi)}-\hat{y}_{d(h.)}\right]^2+\left(\left(n_h-1\right)\left(\hat{y}_{d(h.)}-\hat{y}_d\right)\right)^2\right]$ donde:
  - $\hat{y}_d$ es el estimador obtenido con los datos de la muestra de un trimestre concreto, calibrando las estimaciones a los valores que proporcionaría el estimador directo por th.
  - $\hat{y}^*_{d(r)}$ es el estimador obtenido con los datos de la submuestra resultante de eliminar el cluster i (vivienda) del estrato h (th) con pesos $w_{j(hi)}$, calibrando las estimaciones a los valores que proporcionaría el estimador directo por th si utilizase los datos de esa submuestra r.
  - $\hat{y}_{d(h.)}=\frac{1}{n_h}\sum_{i=1}^{n_h}\hat{y}_{d(hi)}$
- **Linealización**: Se les asigna el MSE de las estimaciones sin calibrar.

3.3.4 Conclusiones

Los estudios de simulación realizados muestran que el estimador Boostrap es el estimador más adecuado para el cálculo del error cuadrático medio del estimador compuesto 4.

3.4 Software empleado

Para el estudio de esta metodología y la aplicación de los estimadores indicados anteriormente se ha utilizado la programación informática basada en SAS. Se ha elaborado programas específicos diseñado a modo de macro informática, que ejecuta las diferentes tareas descritas: elaboración de estimaciones (número de parados, ocupados, tasa de actividad, ocupación y paro) por comarcas y el cálculo de los errores cuadráticos medios para los diferentes métodos.

La macro proporciona estimaciones calculadas mediante el estimador compuesto (el parámetro alpha es un parámetro de entrada) y al ser éste combinación de un estimador postestratificado y un sintético, también proporciona las estimaciones calculadas mediante éstos.

Otros parámetros de entrada de esta macro son: la variable a estimar, las variables auxiliares a utilizar, la opción de calibración de las estimaciones comarcales a las territoriales obtenidas mediante la estimación directa de la encuesta, el método de estimación de los errores cuadráticos medios (y en el caso del bootstrap, el valor de R).

El programa también ofrece la posibilidad de obtener medias móviles de varios trimestres consecutivos. Además se calculan los errores cuadráticos medios y sus correspondientes coeficientes de variación (cociente de la raíz cuadrada del error cuadrático medio y la estimación), con los tres métodos mencionados.

Esta macro se aplica trimestralmente a las muestras de la PRA y se obtienen, con los parámetros determinados como mejores en el estudio de simulación, las estimaciones de las magnitudes mencionadas y sus errores cuadráticos medios.

Capítulo 4

Estimaciones comarcales 2005-2007

4.1 Definiciones

A continuación se presentan las estimaciones obtenidas utilizando el sistema de estimación antes expuesto en la Encuesta de Población en Relación con la Actividad (PRA), para los años 2005-2007.

Las magnitudes, siempre referidas a la población de 16 y más años, escogidas para su publicación han sido las siguientes, de las cuales se adjunta su definición en la encuesta:

- La tasa de actividad: proporción que forma parte de la población activa. Suele expresarse en porcentajes. En la PRA están calculadas sobre la población de 16 y más años.

$$Tasa\_Actividad = \frac{\sum activos}{\sum población \quad 16 \quad y \quad más} \times 100$$

- La tasa de paro: proporción de activos que se encuentra en paro. Suele expresarse en porcentajes. En la PRA están calculadas sobre la población de 16 y más años.

$$Tasa\_Paro = \frac{\sum parados}{\sum activos} \times 100$$

- La población ocupada: son todas aquellas personas que tienen un empleo remunerado o ejercen una actividad independiente y se encuentran trabajando, o bien, mantienen un vínculo formal con su empleo, estuvieron ausentes del trabajo por vacaciones, enfermedad, conflicto laboral, incidentes técnicos, etc.

- La población parada. Son todas aquellas personas que no tienen empleo asalariado o empleo independiente y están actualmente buscando empleo y disponibles para trabajar.

Desde 2002, siguiendo el Reglamento de la Comisión Europea 1897/2000, se consideran paradas aquellas personas que además de cumplir las anteriores condiciones, han realizado durante las cuatro semanas anteriores alguna de las gestiones de búsqueda de empleo de las consideradas activas en dicho reglamento. No se considera búsqueda activa el trámite de renovar la demanda de trabajo ("sellar la cartilla") o de contactar por motivo de los cursos de formación con la oficina pública de empleo.

Junto con las estimaciones se ofrecen en las tablas los coeficientes de variación (CV) de las mismas.

La división oficial en comarcas de la C.A. de Euskadi es la siguiente:

Alava: Valles Alaveses, Llanada Alavesa, Montaña Alavesa, Rioja Alavesa, Estribaciones del Gorbea y Cantábrica Alavesa:

Bizkaia: Arratia-Nervión, Gran Bilbao, Duranguesado, Encartaciones, Gernika-Bermeo, Markina-Ondarroa y Plentzia-Mungia

Gipuzkoa: Bajo Bidasoa, Bajo Deba, Alto Deba, Donostia-San Sebastián, Goierri, Tolosa y Urola Costa

(Ver Anexo, con la relación de comarcas y municipios)

Seguidamente, se pasa a presentar los resultados más destacables.

4.2 Resultados

Tabla 1. Tasa de actividad de la población de 16 y más años por Territorio Histórico y comarca. Estimación y Coeficientes de Variación (en porcentajes).

Fuente: EUSTAT. Encuesta de Población en Relación con la Actividad (PRA)

	2005		2006		2007	
	Estimación	CV	Estimación	CV	Estimación	CV
C.A. de Euskadi	**54,8**	**0,35**	**54,6**	**0,37**	**54,6**	**0,35**
Alava	**57,0**	**0,68**	**57,6**	**0,68**	**56,8**	**0,66**
Valles Alaveses	45,0	1,79	45,6	1,89	54,0	1,70
Llanada Alavesa	58,6	0,92	59,2	0,88	58,2	0,87
Montaña Alavesa	41,0	2,32	35,0	2,84	42,2	3,39
Rioja Alavesa	54,7	2,15	55,0	2,36	50,5	2,20
Estribaciones del Gorbea	51,4	1,41	49,7	2,02	51,7	2,71
Cantábrica Alavesa	48,8	1,34	49,6	1,47	49,8	1,49
Bizkaia	**53,1**	**0,54**	**52,6**	**0,55**	**52,9**	**0,56**
Arratia-Nervión	43,6	1,91	44,5	1,94	47,5	1,86
Gran Bilbao	53,1	0,74	52,6	0,74	52,8	0,76
Duranguesado	55,6	1,13	54,2	1,07	54,7	1,43
Encartaciones	44,8	1,55	43,7	1,92	46,6	1,84
Gernika-Bermeo	49,0	1,28	48,8	1,37	49,1	1,25
Markina-Ondarroa	45,7	1,54	46,6	1,38	47,9	1,81
Plentzia-Mungia	59,1	1,18	57,6	1,47	58,0	1,35
Gipuzkoa	**56,7**	**0,58**	**56,6**	**0,58**	**56,7**	**0,59**
Bajo Bidasoa	55,9	1,13	55,7	1,09	57,5	0,98
Bajo Deba	49,4	1,30	49,3	1,31	49,2	1,17
Alto Deba	57,6	1,11	57,0	1,15	53,1	1,14
Donostialdea	58,8	0,83	58,4	0,88	59,0	0,88
Goierri	52,4	1,35	52,8	1,38	51,7	1,24
Tolosa	52,5	1,16	54,4	1,22	55,2	1,36
Urola Costa	55,8	1,12	57,3	1,07	57,4	1,08

Tabla 2. Tasa de paro de la población de 16 y más años por Territorio Histórico y comarca. Estimación y Coeficientes de Variación (en porcentajes).

Fuente: EUSTAT. Encuesta de Población en Relación con la Actividad (PRA)

	2005		2006		2007	
	Estimación	CV	Estimación	CV	Estimación	CV
C.A. de Euskadi	**5,7**	**2,30**	**4,1**	**2,63**	**3,3**	**2,90**
Alava	**3,0**	**5,70**	**3,5**	**5,39**	**2,3**	**6,90**
Valles Alaveses	2,8	8,70	3,4	9,06	2,1	9,10
Llanada Alavesa	3,0	7,60	3,5	7,26	2,2	9,50
Montaña Alavesa	3,1	10,00	5,8	31,37	2,9	18,60
Rioja Alavesa	2,3	10,10	3,0	16,82	2,9	20,50
Estribaciones del Gorbea	3,0	7,50	3,1	13,38	1,9	11,60
Cantábrica Alavesa	3,6	9,90	4,1	10,86	2,5	14,20
Bizkaia	**7,4**	**2,90**	**5,0**	**3,55**	**4,0**	**4,00**
Arratia-Nervión	7,2	14,50	6,5	20,38	3,5	17,80
Gran Bilbao	7,8	4,00	5,2	4,86	4,2	5,40
Duranguesado	5,6	7,30	3,4	8,77	2,6	9,20
Encartaciones	5,4	8,10	4,6	17,51	3,7	18,50
Gernika-Bermeo	7,5	10,90	4,1	12,00	4,0	13,70
Markina-Ondarroa	4,8	8,70	3,2	11,54	2,4	9,40
Plentzia-Mungia	5,8	7,90	4,6	12,05	3,6	13,50
Gipuzkoa	**4,2**	**4,10**	**2,9**	**5,05**	**2,6**	**5,30**
Bajo Bidasoa	4,4	7,70	2,8	9,44	3,0	11,00
Bajo Deba	4,2	9,40	3,3	12,95	2,6	11,80
Alto Deba	3,1	8,10	2,2	8,78	2,0	9,10
Donostialdea	4,7	6,10	3,2	7,58	2,7	8,30
Goierri	3,7	9,00	2,5	10,18	2,6	12,00
Tolosa	4,1	9,70	2,4	8,73	2,2	13,00
Urola Costa	3,8	7,80	2,9	9,40	2,3	10,40

Tabla 3. Población de 16 y más años ocupada por Territorio Histórico y comarca. Estimación (en miles) y Coeficientes de Variación (en porcentajes).

Fuente: EUSTAT. Encuesta de Población en Relación con la Actividad (PRA)

	2005		2006		2007	
	Estimación	CV	Estimación	CV	Estimación	CV
C.A. de Euskadi	**941,2**	**0,78**	**954,2**	**0,78**	**964,8**	**0,77**
Alava	**141,1**	**1,53**	**142,4**	**1,54**	**143,5**	**1,55**
Valles Alaveses	1,9	2,24	1,9	2,76	2,4	1,99
Llanada Alavesa	116,1	1,16	117,5	1,17	117,8	1,18
Montaña Alavesa	1,2	3,05	1,0	4,19	1,2	3,55
Rioja Alavesa	4,8	3,22	4,8	3,44	4,5	2,57
Estribaciones del Gorbea	3,0	1,92	3,0	2,45	3,1	2,68
Cantábrica Alavesa	13,8	1,68	14,1	1,80	14,5	1,89
Bizkaia	**484,7**	**1,17**	**491,9**	**1,17**	**500,0**	**1,16**
Arratia-Nervión	7,6	2,24	7,8	2,41	8,6	2,05
Gran Bilbao	375,2	0,95	382,0	0,94	387,0	0,94
Duranguesado	41,4	1,36	41,2	1,28	41,9	1,63
Encartaciones	11,2	2,03	11,0	2,44	11,9	2,26
Gernika-Bermeo	17,7	1,64	18,3	1,82	18,5	1,61
Markina-Ondarroa	10,0	2,05	10,4	2,19	10,8	2,36
Plentzia-Mungia	21,4	1,59	21,1	1,75	21,5	1,59
Gipuzkoa	**315,5**	**1,28**	**319,8**	**1,28**	**321,3**	**1,26**
Bajo Bidasoa	33,0	1,36	33,4	1,23	34,4	1,28
Bajo Deba	22,7	1,51	22,9	1,46	23,0	1,27
Alto Deba	30,3	1,35	30,2	1,37	28,2	1,29
Donostialdea	152,2	1,11	153,5	1,12	155,9	1,07
Goierri	27,8	1,76	28,3	1,66	27,7	1,46
Tolosa	19,2	1,62	20,2	1,61	20,5	1,53
Urola Costa	30,1	1,37	31,2	1,41	31,4	1,49

Tabla 4. Población de 16 y más años parada por Territorio Histórico y comarca. Estimación (en miles) y Coeficientes de Variación (en porcentajes).

Fuente: EUSTAT. Encuesta de Población en Relación con la Actividad (PRA)

	2005		2006		2007	
	Estimación	CV	Estimación	CV	Estimación	CV
C.A. de Euskadi	**57,0**	**3,01**	**40,5**	**3,53**	**32,5**	**3,88**
Alava	**4,4**	**7,76**	**5,2**	**7,17**	**3,4**	**9,03**
Valles Alaveses	0,1	7,83	0,1	8,82	0,1	8,97
Llanada Alavesa	3,6	8,08	4,3	7,48	2,7	9,40
Montaña Alavesa	0,0	9,86	0,1	28,88	0,0	17,93
Rioja Alavesa	0,1	8,44	0,1	18,12	0,1	20,89
Estribaciones del Gorbea	0,1	7,45	0,1	12,25	0,1	11,51
Cantábrica Alavesa	0,5	10,36	0,6	12,23	0,4	15,11
Bizkaia	**38,6**	**3,86**	**25,7**	**4,75**	**20,7**	**5,17**
Arratia-Nervión	0,6	15,22	0,5	20,36	0,3	17,83
Gran Bilbao	31,6	4,01	21,0	5,02	17,0	5,49
Duranguesado	2,5	8,66	1,4	9,68	1,1	8,59
Encartaciones	0,6	8,79	0,5	17,59	0,5	20,58
Gernika-Bermeo	1,4	11,60	0,8	12,17	0,8	15,39
Markina-Ondarroa	0,5	9,05	0,3	10,86	0,3	8,15
Plentzia-Mungia	1,3	8,93	1,0	14,25	0,8	14,81
Gipuzkoa	**14,0**	**5,53**	**9,6**	**6,72**	**8,5**	**7,01**
Bajo Bidasoa	1,5	9,10	1,0	10,34	1,1	11,58
Bajo Deba	1,0	10,48	0,8	13,80	0,6	13,25
Alto Deba	1,0	8,75	0,7	9,46	0,6	9,30
Donostialdea	7,4	6,65	5,0	8,03	4,3	8,63
Goierri	1,1	9,91	0,7	11,04	0,7	13,51
Tolosa	0,8	10,25	0,5	8,80	0,5	14,73
Urola Costa	1,2	8,90	0,9	10,46	0,7	11,70

Capítulo 5

Conclusiones

La cada vez mayor demanda de información desagregada y la necesidad de no recargar a los informantes hacen que los métodos de estimación basados en modelos estén progresivamente siendo más utilizados en la estadística oficial.

La obtención de estimaciones de las magnitudes relacionadas con la actividad en áreas pequeñas como son las comarcas, que presentamos aquí, es un paso adelante en la aplicación de las nuevas metodologías de estimación basadas en modelos en el Instituto.

Los resultados presentados en este documento ofrecen una calidad aceptable en términos de precisión. La mayoría de los coeficientes de variación (CV) obtenidos en las estimaciones no supera el 15% y sólo alguno sobrepasa el 20%.

Eusta, a partir de ahora, puede ofrecer estimaciones comarcales a partir de una encuesta coyuntural con lo que ello supone de aumento de la eficiencia de la operación.

Las estimaciones podrán ser mejoradas en la medida en que una mejor información auxiliar esté disponible. La disponibilidad de una información auxiliar adecuada es fundamental en los modelos y, por ello, es importante contar con unos marcos adecuados y tener acceso a la información de los ficheros administrativos.

Eustat pretende seguir avanzando en el estudio y aplicación de la metodología de estimación basada en modelos para poder ofrecer cada vez información más desagregada y de calidad.

Capítulo 6

Bibliografía

CLARKE, PHILIP; MCGRATH, KEVIN; HUKUM, CHANDRA AND TZAVIDIS, NIKOS (2007)

Developments in Small Area Estimation in UK with focus on current research activities. IASS Satellite Meeting on Small Area Estimation

EUSTAT (2005)

Informe sobre el Cálculo de Errores de Muestreo. Encuesta de Población en Relación con la Actividad. (PRA). http://www.eustat.es/document/datos/Calculo_errores_PRA_c.pdf

EUSTAT (2007)

Proyecto Técnico de la Operación Encuesta de Población en Relación con la Actividad. (PRA).

GHOSH, M. AND RAO, J.N.K.,(1994)

Small Area Estimation: An Appraisal. Statistical Science, 9, 55-93.

GHOSH, N. AND SÄRNDAL, C.E. (2001)

Lecture Notes for Estimation for Population Domains and Small Areas. Statistics Finland ., vol. 48.

INSEE INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES (1993)

"La macro Calmar, Redressement d'un échantillon par calage sur marges", Document nº F9310 25/11/1993, Olivier Sautory. Série des documents de travail de la Direction des Statistiques Démographiques et Sociales.. Insee - La macro SAS Calmar.

QUENOUILLE, M. (1949)

Approximate tests of correlation in time series. J. Roy Statist. Soc. Ser. B, 11, 18-84.

QUENOUILLE, M. (1956)

Notes on bias in estimation. Biometrika, 43 pp. 353-360.

RAO, J.N.K. AND WU, C.F.J. (1988)

Resampling Inference with Complex Survey Data. Journal of the American Statistical Association , 83, 231-241

SÄRNDAL, C.E. SWENSSON, B. AND WRETMAN J. (1992)

Model Assisted Survey Sampling. Springer-Verlag

SAS INSTITUTE INC., "SAS/STAT® 9. (2004)

"User's Guide". Copyright © 2004, Cary, NC, USA. ISBN

TUKEY, J. (1958)

Bias and confidence in not quite large samples. Abstract, Ann. Math. Statist.., 29, 614

WOODRUFF, R.S., (1971)

A Simple Method for Approximating the Variance of a Complicated Estimate. Journal of The American Statistical Association. 66(334), 411-414

Anexo

ALAVA/ARABA

Arabako Ibarrak / Valles Alaveses: Añana, Armiñón, Berantevilla, Kuartango, Lantarón, Ribera Alta, Ribera Baja/Erribera Beitia, Valdegovía/Gaubea, Zambrana

Arabako Lautada / Llanada Alavesa: Alegría-Dulantzi, Arrazua-Ubarrundia Asparrena, Barrundia, Elburgo/Burgelu, Iruña Oka/Iruña de Oca, Iruraiz-Gauna, Salvatierra/Agurain, San Millán/Donemiliaga, Vitoria-Gasteiz, Zalduondo

Arabako Mendialdea / Montaña Alavesa: Arraia-Maeztu, Bernedo, Campezo/Kanpezu, Harana/Valle de Arana, Lagrán, Peñacerrada-Urizaharra

Errioxa Arabarra / Rioja Alavesa: Baños de Ebro/Mañueta, Elciego, Elvillar/Bilar, Kripan, Labastida/Bastida, Laguardia, Lanciego/Lantziego, Lapuebla de Labarca, Leza, Moreda de Álava, Navaridas, Oyón-Oion, Samaniego, Villabuena de Alava/Eskuernaga, Yécora/Iekora

Gorbeia Inguruak / Estribaciones del Gorbea: Aramaio, Legutiano, Urkabustaiz, Zigoitia, Zuia

Kantauri Arabarra / Cantábrica Alavesa: Amurrio, Artziniega, Ayala/Aiara, Laudio/Llodio, Okondo

BIZKAIA

Arratia Nerbioi / Arratia-Nervión: Arakaldo, Arantzazu, Areatza, Arrankudiaga, Artea, Dima, Igorre, Orozko, Otxandio, Ubide, Ugao-Miraballes, Urduña-Orduña, Zeanuri, Zeberio

Bilbo Handia / Gran Bilbao: Abanto y Ciérvana-Abanto Zierbena, Alonsotegi, Arrigorriaga, Barakaldo, Basauri, Berango, Bilbao, Derio, Erandio, Etxebarri, Galdakao, Getxo, Larrabetzu, Leioa, Lezama, Loiu, Muskiz, Ortuella, Portugalete, Santurtzi, Sestao, Sondika, Valle de Trápaga-Trapagaran, Zamudio, Zaratamo, Zierbena

Durangaldea / Duranguesado: Abadiño, Amorebieta-Etxano, Atxondo, Bedia, Berriz, Durango, Elorrio, Ermua, Garai, Iurreta, Izurtza, Lemoa, Mallabia, Mañaria, Zaldibar

Enkartazioak / Encartaciones: Artzentales, Balmaseda, Galdames, Gordexola, Güeñes, Karrantza Harana/Valle de Carranza, Lanestosa, Sopuerta, Trucios-Turtzioz, Zalla

Gernika-Bermeo: Ajangiz, Arratzu, Bermeo, Busturia, Ea, Elantxobe, Ereño, Errigoiti, Forua, Gautegiz Arteaga, Gernika-Lumo, Ibarrangelu, Kortezubi, Mendata, Morga, Mundaka, Murueta, Muxika, Nabarniz, Sukarrieta

Markina-Ondarroa: Amoroto, Aulesti, Berriatua, Etxebarria, Gizaburuaga, Ispaster, Lekeitio, Markina-Xemein, Mendexa, Munitibar-Arbatzegi Gerrikaitz-, Ondarroa, Ziortza-Bolibar

Plentzia-Mungia: Arrieta, Bakio, Barrika, Fruiz, Gamiz-Fika, Gatika, Gorliz, Laukiz, Lemoiz, Maruri-Jatabe, Meñaka, Mungia, Plentzia, Sopelana, Urduliz

GIPUZKOA

Bidasoa Beherea / Bajo Bidasoa: Hondarribia, Irun

Deba Beherea / Bajo Deba: Deba, Eibar, Elgoibar, Mendaro, Mutriku, Soraluze-Placencia de las Armas

Deba Garaia / Alto Deba: Antzuola, Aretxabaleta, Arrasate/Mondragón, Bergara, Elgeta, Eskoriatza, Leintz-Gatzaga, Oñati

Donostialdea / Donostia-San Sebastián: Andoain, Astigarraga, Donostia-San Sebastián, Errenteria, Hernani, Lasarte-Oria, Lezo, Oiartzun, Pasaia, Urnieta, Usurbil

Goierri: Altzaga, Arama, Ataun, Beasain, Ezkio-Itsaso, Gabiria, Gaintza, Idiazabal, Itsasondo, Lazkao, Legazpi, Mutiloa, Olaberria, Ordizia, Ormaiztegi, Segura, Urretxu, Zaldibia, Zegama, Zerain, Zumarraga

Tolosaldea / Tolosa: Abaltzisketa, Aduna, Albiztur, Alegia, Alkiza, Altzo, Amezketa, Anoeta, Asteasu, Baliarrain, Belauntza, Berastegi, Berrobi, Bidegoian, Elduain, Gaztelu, Hernialde, Ibarra, Ikaztegieta, Irura, Larraul, Leaburu, Legorreta, Lizartza, Orendain, Orexa, Tolosa, Villabona, Zizurkil

Urola-Kostaldea / Urola Costa: Aia, Aizarnazabal, Azkoitia, Azpeitia, Beizama, Errezil, Getaria, Orio, Zarautz, Zestoa, Zumaia