Contribuciones a la Estadística
y a la Investigación Operativa
Tenerife, 15-19 de mayo de 2006
ISBN XXX-XXXXXX-X
páginas 1–16

Estimación en áreas pequeñas en la Encuesta Industrial de la C.A. de Euskadi

Josu Iradi[1], *Haritz Olaeta*[2]

[1]eustat@eustat.es, Eustat
[2]H_Olaeta@eustat.es, Eustat

Resumen

Eustat ha publicado por primera vez en el 2005 estimaciones comarcales de la Encuesta Industrial de los años 2002 y 2003 obtenidas mediante la aplicación de modelos de áreas pequeñas, siendo también la primera vez que se utiliza esta metodología para difundir datos en la estadística oficial de la C.A. de Euskadi. En este trabajo se resumen los modelos utilizados y se presentan las estimaciones publicadas.

Palabras Clave: Áreas Pequeñas, Modelo lineal mixto, Modelo de efectos fijos

1. Introducción

Eustat, consciente de la creciente demanda de estadísticas de calidad cada vez más desagregadas, constituyó hace dos años un equipo de investigación compuesto por miembros de distintas áreas de Eustat y miembros de la Universidad Pública de Navarra. El objetivo ha sido y es 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 de la estadística oficial.

El trabajo que aquí se presenta, hace referencia a la primera encuesta de Eustat en la que se ha definido un sistema de estimación en áreas pequeñas. Esta operación ha sido la Encuesta Industrial anual, principalmente debido a su especial relevancia dentro de las encuestas económicas de Eustat y a ser un sector clave en la economía vasca.

El contenido de este trabajo pretende ser el punto de partida de un proyecto de aplicación de estas técnicas en otras encuestas de Eustat.

2. Sistema de estimación en la Encuesta Industrial de la C.A. de Euskadi

Todas las unidades estadísticas con más de 19 empleados son autoponderadas, por lo que el interés radica principalmente en las estimaciones dentro del estrato de empleo de 1-19 empleados. En lo que sigue se describe la estimación dentro de una subclase de actividad (CNAE-93 a 5 dígitos) del total de una variable y cualquiera, así como de su correspondiente coeficiente de variación. Actualmente, para la estimación del total de la variable se utiliza el estimador indirecto de razón o estimador sintético, utilizando como información auxiliar el número de empleados de los establecimientos. Esta decisión se tomó tras realizar un estudio descriptivo exhaustivo en el que se comprobó que existe una correlación positiva fuerte entre el número de empleados de los establecimientos y la magnitud de las principales variables de la Encuesta Industrial.

El estimador indirecto de razón de una variable de interés cualquiera , cuando se dispone de una variable auxiliar , está, en el caso de la Encuesta Industrial, asistido por el modelo de regresión lineal simple heterocedástico del tipo:

$$y_{hj} = x_{hj}\beta + \epsilon_{hj} \qquad var(\epsilon_{hj}) = \sigma^2 x_{hj}, \tag{1}$$

donde h hace referencia al estrato y j a la unidad estadística. Los estratos son los Territorios Históricos (Álava, Bizkaia y Gipuzkoa) dado que en todo lo que sigue se supone el interés radica en una única subclase. El modelo lineal simple (1) contiene perturbaciones heterocedásticas, siendo la varianza función lineal creciente del número de empleados de los establecimientos industriales.

El estimador del total de la variable y en una subclase dada en el Territorio Histórico h viene dado por:

$$\hat{t}_{yh.\mathrm{SYN}} = X_h\hat{\beta} = X_h \frac{\sum_{h=1}^{H}\sum_{j=1}^{n_h} w_{hj}y_{hj}}{\sum_{h=1}^{H}\sum_{j=1}^{n_h} w_{hj}x_{hj}}, \tag{2}$$

donde $X_h = \sum_{j=1}^{N_h} x_{hj}$, w_{hj} es el peso de muestreo de la unidad j en el Territorio Histórico h, x_{hj} recoge el empleo del establecimiento j del Teritorio Histórico h y n_h es el tamaño de la muestra en el Territorio Histórico h.

El estimador de la varianza del estimador indirecto de razón se puede aproximar mediante:

$$\hat{\mathrm{var}}(\hat{t}_{yh.\mathrm{SYN}}) \approx N^2\Big(1-\frac{n}{N}\Big)\frac{1}{n}\left[\frac{X_h}{\sum_{h=1}^{H}\sum_{j=1}^{n_h} w_{hj}x_{hj}}\right]^2 \hat{\mathrm{var}}(\epsilon), \tag{3}$$

donde $\hat{\mathrm{var}}(\epsilon)$ es la varianza muestral de los residuos del modelo heterocedástico (1) con todos los datos muestrales (es decir, se calculan los residuos en toda la

CA de Euskadi, no sólo en el Territorio Histórico h) y el resto de la notación es la habitual.

Särndal and Hidiroglou (1989) proporcionan una aproximación del sesgo del estimador sintético según la cual $E[\hat{t}_{yh.\mathrm{SYN}}] - t_{yh.\mathrm{SYN}} \approx -\sum_{j=1}^{N} \hat{\epsilon}_j$, donde $\hat{\epsilon}_j = y_j - x_j\hat{\beta}$. Por consiguiente, el estimador será aproximadamente insesgado si se verifica que $\sum_{j=1}^{N} \hat{\epsilon}_j = 0$. Esta condición no se satisface normalmente. Si el modelo no ajusta bien en el dominio de interés, la suma de residuos puede estar lejos de cero, indicando un sesgo considerable. En caso contrario, podemos esperar un sesgo limitado. Por ello, es deseable estimar el error cuadrático medio (RMSE en adelante) como medida de precisión del estimador. Viene dado por:

$$\mathrm{MSE}(\hat{t}_{yh.\mathrm{SYN}}) = \mathrm{var}(\hat{t}_{yh.\mathrm{SYN}}) + \left(\mathrm{sesgo}_{yh.\mathrm{SYN}}\right)^2, \tag{4}$$

y se estima mediante la expresión:

$$\hat{\mathrm{MSE}}(\hat{t}_{yh.\mathrm{SYN}}) = \hat{\mathrm{var}}(\hat{t}_{yh.\mathrm{SYN}}) + \Big(\sum_{j=1}^{n_h} \hat{\epsilon}_j\Big)^2, \tag{5}$$

donde $\hat{\epsilon}_j = y_j - x_j'\hat{\beta}$, para $j = 1, ..., n$ son los residuos obtenidos a partir del modelo estimado (1) con todos los datos muestrales, aunque en cada Territorio Histórico solamente se suman los específicos de ese Territorio Histórico. El coeficiente de variación se define como:

$$\hat{\mathrm{cv}}(\hat{t}_{yh.\mathrm{SYN}}) = \frac{\hat{\mathrm{RMSE}}(\hat{t}_{yh.\mathrm{SYN}})}{\hat{t}_{yh.\mathrm{SYN}}}, \qquad \hat{\mathrm{RMSE}}(\hat{t}_{yh.\mathrm{SYN}}) = \sqrt{\hat{\mathrm{MSE}}(\hat{t}_{yh.\mathrm{SYN}})}. \tag{6}$$

3. Sistema de estimación en áreas pequeñas en la Encuesta Industrial

La C.A. de Euskadi se divide en las siguientes 20 comarcas:

- **Álava**: 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.

La actividad industrial de la C.A. de Euskadi no está uniformemente repartida en las 20 comarcas estadísticas y, tanto la importancia del sector

industrial como su tamaño varía enormemente entre comarcas. De hecho, hay comarcas en las que la actividad industrial es realmente reducida, por lo que la tarea de estimación comarcal pasa forzosamente por aplicar técnicas de estimación de áreas pequeñas basadas en modelos. El aumento del tamaño muestral requerido para obtener estimaciones comarcales de calidad sería ciertamente costoso.

Los modelos de áreas pequeñas suponen la existencia de un modelo subyacente que siguen todos los datos de la población, pero que se estima con los datos de la muestra (Rao, 2003). Eustat utiliza para la obtención de estimaciones comarcales en la Encuesta Industrial dos tipos de modelos: el modelo de regresión lineal de efectos fijos y el modelo de regresión lineal con efectos fijos y aleatorios, llamado también modelo mixto (ver, por ejemplo, Rao 2003, para detalles).

3.1. Modelo lineal mixto

Se parte de una población formada por los N establecimientos de una subclase de actividad (CNAE-93 a 5 dígitos) concreta. En cada comarca d ($d = 1, ..., t$) hay N_d establecimientos en la población, de modo que $N_d = \sum_d N_d$. En dicha subclase se muestrean n establecimientos de los que n_d pertenecen a la comarca d. Se propone el siguiente modelo lineal mixto heterocedástico:

$$y_{dj} = \beta_0 + \beta_1 x_{dj} + v_d + e_{dj}, \qquad d = 1, ..., t \quad j = 1, ..., n_d \tag{7}$$

donde para el establecimiento j de la comarca d, y_{dj} es el valor que toma la variable de interés y x_{dj} es el número de empleados del establecimiento. El número total de establecimientos muestreados en la comarca d es n_d. Los efectos fijos del modelo son β_0 y β_1 . El efecto aleatorio común para todos los establecimientos de la comarca d es v_d y e_{dj} son los errores aleatorios específicos de cada establecimiento. Además, se supone que $v_d \sim N(0, \sigma_v^2)$ y $e_{dj} \sim N(0, \sigma_e^2 c_{dj}^{-1})$ son independientes. Para corregir la heterocedasticidad presente en los datos se utilizan los pesos $c_{dj} = 1/x_{dj}$. Cuando $c_{dj} = 1 \ \ \forall d, j$, este modelo es similar al propuesto por Battese *et al* (1988).

El modelo superpoblacional correspondiente al modelo (7) escrito en forma matricial se expresa como:

$$Y = X\beta + Zv + \epsilon \qquad v \sim N(0, \sigma_v^2 I_t), \qquad \epsilon \sim N(0, \sigma_\epsilon^2 C^{-1}), \tag{8}$$

donde $C = \text{diag}(c_{dj})$ $(d = 1, ..., t)$, es la matriz de pesos del modelo y j es el establecimiento $(j = 1, ..., N_d)$. El vector $Y = (Y_1', ...Y_t')$ es el vector $(N \times 1)$ cuyas componentes Y_d' son los valores de la variable de interés para cada comarca, $\beta = (\beta_0, \beta_1)'$ es el vector de coeficientes del modelo, X es la matriz de diseño $(N \times 2)$ formada por una columna de unos asociada a la ordenada en el origen y otra columna asociada a la variable auxiliar que es, en este caso,

el número de empleados de cada establecimiento. La matriz $Z = \text{diag}(1_{N_d})$, $d = 1, ..., t$, es la matriz de diseño $(N \times t)$ diagonal por bloques asociada a los efectos aleatorios. Es decir, para cada comarca d, la matriz Z tiene una columna asociada de unos definida por el vector $1_{N_d} = (1, ..., 1)'$ de dimensión N_d. Los efectos aleatorios $v = (v_1, ..., v_t)'$, son comunes a los N_d elementos de la misma comarca y $\epsilon = (\epsilon_1', ..., \epsilon_t')'$ es el vector de errores aleatorios, donde $\epsilon_d = (\epsilon_{d_1}, ..., \epsilon_{d_{N_d}})'$.

Es conveniente diferenciar en el modelo la parte muestreada y la no muestreada del siguiente modo:

$$\begin{pmatrix} Y_s \\ Y_r \end{pmatrix} = \begin{pmatrix} X_s \\ X_r \end{pmatrix} \beta + \begin{pmatrix} Z_s \\ Z_r \end{pmatrix} v + \begin{pmatrix} \epsilon_s \\ \epsilon_r \end{pmatrix}, \tag{9}$$

donde los subíndices s y r denotan los establecimientos muestreados y no muestreados respectivamente. Entonces el modelo muestral puede escribirse como:

$$Y_s = X_s\beta + Z_s v + \epsilon_s \qquad v \sim N(0, \sigma_v^2 I_t), \qquad \epsilon_s \sim N(0_s, \sigma_e^2 C_s^{-1}), \tag{10}$$

donde $C_s = \text{diag}(c_{dj} = 1/x_{dj})$, $d = 1, ..., t_s$, $j = 1, ..., n_d$ y t_s es el número total de comarcas donde se ha muestreado. La matriz de varianzas y covarianzas de Y_s puede expresarse como $\text{var}(Y_s) = V_s = Z_s\sigma_v^2 Z_s' + \sigma_e^2 C_s^{-1} = \text{diag}(V_1, ..., V_{t_s})$ donde $V_d = \sigma_e^2 C_d^{-1} + \sigma_v^2 1_{n_d} 1_{n_d}'$ y $C_d = \text{diag}(c_{d_1}, ..., c_{d_{n_d}})_{n_d \times n_d} = \text{diag}(c_{n_d})$.

Cuando la fracción de muestreo por comarca $f_d = n_d/N_d$ es significativa, se recomienda utilizar la versión predictiva para obtener la predicción del total o de la media de la comarca d. Esta versión consiste en diferenciar la parte muestreada de la no muestreada. Así, la predicción de la parte muestreada es la misma muestra, mientras que la no muestreada se predice con el estimador de tipo proyectivo.

Para obtener la versión predictiva se descompone el total $\sum_{j \in N_d} y_{dj} = \sum_{j \in d_r} y_{dj} + \sum_{j \in d_s} y_{dj}$, donde d_s indica la muestra en la comarca d y d_r el resto de los establecimientos no pertenecientes a la muestra de la comarca d . A continuación se descompone la media poblacional:

$$\bar{Y}_d = \frac{\sum_{j \in d_r} y_{dj} + \sum_{j \in d_s} y_{dj}}{N_d} = \frac{(N_d - n_d)\bar{Y}_{dr} + n_d \bar{y}_{ds}}{N_d} = (1 - f_d)\bar{Y}_{dr} + f_d \bar{y}_{ds}. \tag{11}$$

El estimador predictivo de la media del área d, para todo $d = 1, ..., t$ viene dado por:

$$\hat{\bar{y}}_d = \hat{\bar{Y}}_d = (1 - f_d)\bar{Y}_{dr} + f_d \bar{y}_{ds} = (1 - f_d)\hat{\bar{y}}_{dr}^* + f_d \bar{y}_{ds} \tag{12}$$

Sustituyendo $\hat{\bar{y}}^*_{dr}$ por su expresión $\bar{X}'_{dr}\hat{\beta} + \hat{v}_d$ donde $\bar{X}'_{dr} = \left(1, \bar{x}_{dr}\right)$ y $\bar{x}_{dr} = \frac{\sum_{j \in d_r} x_{dj}}{N_d - n_d}$, entonces se tiene que:

$$\hat{\bar{y}}_d = (1 - f_d)\left[\bar{X}_{d(p_r)}\hat{\beta} + \hat{\gamma}_{dc}(\bar{y}_{dc} - \bar{x}'_{dc}\hat{\beta}\right] + f_d\bar{y}_{ds}, \qquad \hat{\gamma} = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_d} \tag{13}$$

que conduce a la versión predictiva del total:

$$\hat{t}_d = X_{d(p_r)}\hat{\beta} + (N_d - n_d)\hat{\gamma}_{dc}\left[\bar{y}_{dc} - \bar{x}_{dc}\hat{\beta}\right] + \sum_{j=1}^{n_d} y_{dj}, \qquad d = 1, ..., t \tag{14}$$

donde $\hat{\beta} = \tilde{\beta}_c\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)$ ha sido evaluada con las estimaciones de los componentes de varianza con $\tilde{\beta} = (X'_s V_s^{-1} X_s)^{-1} X'_s V_s^{-1} Y_s$ y $X_{d(p_r)}$ es el total de empleados en la comarca d para todos los establecimientos no muestreados.

El estimador de la media por Territorio Histórico viene dado por:

$$\hat{\bar{y}}_h = \frac{1}{N_h}\sum_{d \in h} N_d \hat{\bar{y}}_d = \frac{1}{N_h}\sum_{d \in h} \hat{t}_d, \tag{15}$$

donde $d \in h$ indica que la suma se efectúa en todas las comarcas del estrato h (en este caso $h = 1, 2, 3$) son los Territorios Históricos) y $N_h = \sum_{d \in h} N_d$ es el total poblacional del Territorio Histórico h.

El estimador del total por Territorio Histórico viene dado por:

$$\hat{t}_h = \sum_{d \in h} \hat{t}_d. \tag{16}$$

El estimador de la media para la C.A. de Euskadi viene dado por:

$$\hat{\bar{y}} = \frac{1}{N}\sum_{h=1}^{3} N_h \hat{\bar{y}}_h \qquad\qquad N = \sum_{h=1}^{3} N_h. \tag{17}$$

El estimador del total para la C.A. de Euskadi viene dado por:

$$\hat{t} = \sum_{h=1}^{3} N_h \hat{\bar{y}}_h = \sum_{h=1}^{3} \hat{t}_h. \tag{18}$$

En la versión predictiva, el estimador del error cuadrático medio del predictor (13), válido cuando los estimadores de los componentes de varianza

se obtienen por el método REML (máxima verosimilitud restringida) o por el método de los momentos (ver, por ejemplo, Searle *et al*, 1992), viene dado por:

$$\hat{\mathrm{MSE}}\big[\hat{\bar{y}}\big] = g_{1d}(\hat{\sigma}^2) + g_{2d}(\hat{\sigma}^2) + g_{3d}(\hat{\sigma}^2) + g_{4d}(\hat{\sigma}^2) \tag{19}$$

donde:

$$\begin{array}{rcl} g_{1d}(\hat{\sigma}^2) & = & (1-f_d)^2(1-\hat{\gamma}_{dc})\hat{\sigma}^2 \\ g_{2d}(\hat{\sigma}^2) & = & (1-f_d)^2\big[\big(\hat{X}_{d(p_r)} - \hat{\gamma}_{dc}\bar{x}_{dc}\big)'\hat{\Phi}_s\big(\bar{X}_{d(p_r)} - \hat{\gamma}_{dc}\bar{x}_{dc}\big)\big] \\ g_{3d}(\hat{\sigma}^2) & = & (1-f_d)^2 c_d^{-1}\big(\hat{\sigma}_v^2 + \hat{\sigma}_e^2/c_{d.}\big)^{-3} \\ & & \big[\hat{\sigma}_e^4 \hat{\mathrm{var}}(\hat{\sigma}_v^2) + \hat{\sigma}_v^4 \hat{\mathrm{var}}(\hat{\sigma}_e^2) - 2\hat{\sigma}_e^2\hat{\sigma}_v^2 - 2\hat{\sigma}_e^2\hat{\sigma}_v^2 \hat{\mathrm{cov}}(\hat{\sigma}_e^2, \hat{\sigma}_v^2)\big] \\ g_{4d}(\hat{\sigma}^2) & = & \hat{\sigma}_e^2 N_d^{-1} \sum_{j\in p_r} c_{dj}^{-1}, \end{array} \tag{20}$$

son las contribuciones al MSE de la estimación de los efectos aleatorios, los efectos fijos, los componentes de varianza y los pesos del modelo. $\hat{\Phi}_s = \mathrm{var}(\tilde{\beta}) = (X_s' V_s^{-1} X_s)^{-1}$. Además, p_r representa el dominio d de establecimientos censales no pertenecientes a la muestra y $c_{dj} = 1/x_{dj}$, donde x_{dj} es el empleo censal de los establecimientos no muestreados. Si se hacen agregaciones para conseguir un tamaño mínimo antes de proceder a las estimaciones, se considera como población no muestreada la de la subclase en la que se hace la proyección, y no la de la agregación que suele ser superior.

El MSE del predictor del total para cada comarca se estima multiplicando el estimador del MSE de la media por el cuadrado del tamaño poblacional de la comarca, N_d^2 . En efecto,

$$\hat{\mathrm{MSE}}\big[\hat{t}_d\big] = N_d^2\big[g_{1d}(\hat{\sigma}^2) + g_{2d}(\hat{\sigma}^2) + g_{3d}(\hat{\sigma}^2) + g_{4d}(\hat{\sigma}^2\big], \tag{21}$$

y, por lo tanto:

$$\hat{\mathrm{RMSE}}\big[\hat{t}_d\big] = \sqrt{\hat{\mathrm{MSE}}\big[\hat{t}_d\big]} \quad \hat{\mathrm{cv}}\big[\hat{t}_d\big] = \frac{\hat{\mathrm{RMSE}}\big[\hat{t}_d\big]}{\hat{t}_d} \tag{22}$$

3.2. Modelo lineal de efectos fijos

El modelo superpoblacional de efectos fijos viene dado por:

$$Y = X\beta + \epsilon \qquad \epsilon \sim N(0, \sigma_e^2 C^{-1}), \tag{23}$$

donde $C = \mathrm{diag}(c_{dj})$, es la matriz de pesos del modelo, d representa la copmarca $(d = 1, ..., t)$ y j es el establecimiento $(j = 1, ..., N_d)$. El vector $Y = (Y_1', ..., Y_t')$ es el vector $(N \times 1)$ cuyas componentes Y_d' son los valores observados de la variable de interés para cada comarca d, β es el único coeficiente fijo del modelo, X es el vector columna $(N \times 1)$ de la variable auxiliar, es decir del empleo, y $\epsilon' = (\epsilon_1', ..., \epsilon_t')$ donde $\epsilon_d' = (\epsilon_{d_1}, ..., \epsilon_{d_{N_d}})$ es el vector de errores aleatorios.

De forma similar a la descomposición realizada en el modelo mixto, se puede separar la parte muestreada y no muestreada del siguiente modo:

$$\begin{pmatrix} Y_s \\ Y_r \end{pmatrix} = \begin{pmatrix} X_s \\ X_r \end{pmatrix} \beta + \begin{pmatrix} \epsilon_s \\ \epsilon_r \end{pmatrix}, \tag{24}$$

donde los subíndices s y r denotan los establecimientos muestreados y no muestreados respectivamente. Entonces el modelo muestral de efectos fijos puede escribir como

$$Y_s = X_s\beta + \epsilon_s \qquad \epsilon_s \sim N(0_s, \sigma_e^2 C_s^{-1}), \tag{25}$$

donde $C_s = \text{diag}(c_{dj})$, $d = 1, ..., t_s$, $j = 1, ..., n_d$ y t_s es el número total de comarcas donde se ha muestreado. De forma extendida, el modelo (25) se expresa como:

$$y_{dj} = \beta x_{dj} + e_{dj} \qquad d = 1, ..., t \quad j = 1, ..., n_d \tag{26}$$

donde para el establecimiento j de la comarca d , y_{dj} es el valor que toma la variable de interés y x_{dj} es el número de empleados del establecimiento. El número total de establecimientos muestreados en la comarca d es n_d, β es el único efecto fijo del modelo y $e_{dj} \sim N(0, \sigma_e^2 c_{dj}^{-1})$ son los errores aleatorios. Además, $e_{dj} \sim N(0, \sigma_e^2 c_{dj}^{-1})$. Para corregir la heterocedasticidad presente en los datos se utilizan los pesos $c_{dj} = 1/x_{dj}$.

Si N_d es el número total de unidades del área d , la media poblacional del área d viene dada por

$$\bar{Y}_d = \frac{1}{N_d} \sum_{d=1}^{N_d} y_{dj} = f_d \bar{y}_{ds} + (1 - f_d)\bar{y}_{dr}, \tag{27}$$

donde $f_d = n_d/N_d$, $\bar{y}_{ds}$ es la media muestral de las unidades muestreadas y $\bar{y}_{dr}$ es la media de las no muestreadas. Dado que el segundo término de (27) no se ha observado, se sustituye por su valor estimado. Un estimador de (27) obtenido de manera similar al dado en (11) viene dado por:

$$\hat{\bar{y}}_d^F = f_d \bar{y}_{ds} + (1 - f_d)\bar{X}_{d(p_r)}\hat{\beta} \tag{28}$$

donde $\bar{X}_{d(p_r)} = \sum_{j \in d_r} x_{dj}/(N_d - n_d)$ es la media poblacional del número de empleados no muestreados en el área d.

El estimador de β viene dado por:

$$\hat{\beta} = \left(X_s' C_s X_s\right)^{-1} \left(X_s' C_s Y_s d\right) = \sum_{d=1}^{t_s} \sum_{j=1}^{n_d} y_{dj} / \sum_{d=1}^{t_s} \sum_{j=1}^{n_d} x_{dj} \tag{29}$$

que es el estimador por mínimos cuadrados generalizados de β y

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 \left(X_s' C_s X_s\right)^{-1} = \sigma^2 / \sum_{d=1}^{t_s} \sum_{j=1}^{n_d} x_{dj} \tag{30}$$

esa su matriz de varianzas y covarianzas.

El estimador del total para la comarca d se obtiene como:

$$\hat{t}_d^F = \sum_{j=1}^{n_d} y_{dj} + X_{d(p_r)} \hat{\beta}, \tag{31}$$

y el estimador de la media por Territorio Histórico viene dado por:

$$\hat{\bar{y}}_h^F = \frac{1}{N_h} \sum_{d \in h} N_d \hat{\bar{y}}_d^F, \tag{32}$$

donde $d \in h$ indica que la suma se efectúa en todas las áreas del estrato h (en este caso $h = 1, 2, 3$ son los Territorios Históricos) y $N_h = \sum_{d \in h} N_d$ es el total poblacional del Territorio Histórico h.

El estimador del total por Territorio Histórico viene dado por:

$$\hat{t}_h^F = \sum_{d \in h} N_d \hat{\bar{y}}_d^F = \sum_{d \in h} \hat{t}_d^F \tag{33}$$

donde $d \in h$ indica que la suma se efectúa en todas las áreas del estrato h y $N_h = \sum_{d \in h} N_d$ es el total poblacional del Territorio Histórico h.

El estimador de la media y del total para la C.A. de Euskadi vienen dados respectivamente por:

$$\hat{y}^F = \frac{1}{N} \sum_{h=1}^{3} N_h \hat{y}_h^F, \qquad \hat{t}^F = \sum_{h=1}^{3} \hat{t}_h^F. \tag{34}$$

Los errores cuadráticos medios de los estimadores de la media por comarcas, en su versión predictiva, vienen dados por:

$$\hat{\mathrm{MSE}}\left[\hat{\bar{y}}_d^F\right] = E\left[\left(\hat{\bar{Y}}_d^F - \bar{Y}_d\right)^2\right] = \left(1 - f_d\right)^2 \left[\bar{X}_{d(p_r)} \hat{\mathrm{var}}(\hat{\beta}) \bar{X}_{d(p_r)}\right] + \frac{\sigma^2}{N_d^2} \sum_{j \in p_r} x_{dj}, \tag{35}$$

donde p_r es el empleo censal no muestreado. Si se hacen agregaciones para conseguir un tamaño mínimo antes de proceder a las estimaciones, se considera como población no muestreada la de la subclase en la que se hace la proyección, y no la de la agregación que suele ser superior.

El MSE para el total de la comarca se estima mediante:

$$\hat{\mathrm{MSE}}\left[\hat{t}_d^F\right] = N_d^2\left(1-f_d\right)^2\left[\bar{X}'_{d(p_r)}\hat{\mathrm{var}}(\hat{\beta})\bar{X}_{d(p_r)}\right] + \sigma^2\sum_{j\in p_r} x_{dj}. \qquad (36)$$

3.3. Totales por sector

Los modelos presentados hasta ahora permiten obtener las predicciones de los totales por comarcas, Territorios Históricos y C.A. de Euskadi para cada subclase, así como sus errores estándar (RMSE) y coeficientes de variación. Para el cálculo de los totales por sector, por ejemplo para la clasificación propia de Estat A84, simplemente se realiza una agregación de totales obtenidos por comarcas para cada una de las subclases que lo conforman, agregación que puede hacerse por comarcas, Territorios Históricos y C.A. de Euskadi. Los errores estándar del sector se obtienen como raíz cuadrada de la suma de los MSE de cada subclase. Dentro del mismo sector puede ocurrir que unas subclases se hayan estimado por modelos mixtos y otras por fijos. Para obtener los resultados por un segundo nivel de agregación, por ejemplo para la clasificación A31 se agregan los totales de los sectores A84 que correspondan. El cálculo de errores estándar se hace también suponiendo que los sectores son independientes y por tanto para calcular el RMSE se calcula la raíz cuadrada de la suma de los MSE de los sectores correspondientes.

3.4. Proceso de calibración

La calibración permite obtener exactamente los mismos totales que los proporcionados por el estimador de la Encuesta Industrial a nivel de Territorio Histórico y C.A. de Euskadi por A84 u otro nivel de agregación. Si llamamos t_d a la estimación del total obtenida bajo los modelos en la comarca d por un sector concreto, t_h a la estimación por modelos obtenida por Territorio Histórico y C_h a la estimación del total por Territorio Histórico obtenida por la Encuesta Industrial, entonces la nueva estimación calibrada por comarcas para cada sector vendrá dada por:

$$\tilde{t}_d = t_d\frac{C_h}{t_h}. \qquad (37)$$

De este modo el total calibrado por modelos $\tilde{t}_h$ para cada Territorio Histórico en cada sector coincide con el total estimado C_h para cada Territorio Histórico con el estimador de la Encuesta Industrial para ese mismo sector, ya que:

$$\tilde{t}_h = \sum_d \tilde{t}_d = \sum_d t_d\frac{C_h}{t_h} = \frac{C_h}{t_h}\sum_d t_d = C_h. \qquad (38)$$

3.5. Plan de estimación en la Encuesta Industrial

Eustat ha programado una aplicación informática ad hoc en SAS para la introducción del cálculo de estimaciones de áreas pequeñas en la producción estadística. Se trata de un programa específico para la Encuesta Industrial pero que fácilmente puede adaptarse a otro tipo de encuestas económicas. Son diversas las decisiones que se han tomado y que se han programado:

- Tanto el modelo lineal mixto como el modelo lineal de efectos fijos se calculan en la versión predictiva. Ello es debido a que las poblaciones de algunas subclases son muy pequeñas, con lo que la fracción de muestreo no es despreciable.
- La versión predictiva separa la predicción en dos partes. La obtenida por el modelo para los establecimientos no muestreados y la observada para los establecimientos muestreados. Esta forma es especialmente útil en la Encuesta Industrial debido a que en algunas subclases hay presencia de establecimientos atípicos, es decir, de establecimientos cuyo comportamiento es muy diferente del resto, y que pueden distorsionar considerablemente las estimaciones. Estos establecimientos se clasifican como muestreados no válidos y al utilizar la versión predictiva no se tienen en cuenta en los procedimientos de estimación, pero sí en la predicción final. Los establecimientos clasificados como no válidos se suman al igual que el resto de la muestra, a la predicción de los no observados para obtener la predicción del total.
- Se ha considerado necesario establecer un número mínimo de establecimientos para proceder al cálculo de los modelos mixtos o fijos. Si no se dispone de este número mínimo de establecimientos, se procede a hacer agregaciones de CNAEs con un dígito menos. Primero se estima el modelo mixto a ese nivel de agregación y si $\sigma_v^2 = 0$ ó $\sigma_e^2 = 0$ entonces se estima el modelo de efectos fijos. Este número mínimo se ha fijado en la actualidad (puede ser variado) en 5 establecimientos.
- Se ha tomado la decisión de utilizar el modelo de efectos fijos cuando el modelo mixto no es válido debido a que se considera prioritaria la decisión de no agrupar subclases. Es por ello, que se prefiere un modelo de efectos fijos con 5 dígitos por ejemplo, a un modelo mixto a 3 dígitos.
- Cuando se realiza una agregación, ésta permite estimar los coeficientes del modelo pero las predicciones se hacen de forma particularizada a la subclase considerada.
- El uso de la variable auxiliar 'número de empleados' introduce heterocedasticidad en los modelos, ya que habitualmente la variable respuesta

tiene mayor variabilidad a medida que aumenta el número de empleados. Por ello, todos los modelos de efectos fijos y mixtos consideran que la varianza del error es proporcional al número de empleados.

- En cada subclase los totales por Territorio Histórico y por C.A. de Euskadi se obtienen de manera agregada a partir de las estimaciones por comarcas. Los totales por sector A84 se obtienen agregando las predicciones obtenidas a nivel de subclase. o mismo sucede para los totales por Territorios Históricos y C.A. de Euskadi. Se procede de igual modo para otros tipos de agregaciones.

- En cada subclase, para calcular las raíces cuadradas de los errores cuadráticos medios de las predicciones a nivel de Territorio Histórico, se aplican fórmulas específicas, ya que no se obtiene como raíz cuadrada de la suma de los errores cuadráticos medios de las predicciones por comarcas. Ello es debido a que en cada subclase las estimaciones por comarcas no son independientes en ninguno de los modelos.

- Sin embargo, una vez hechas las estimaciones de los RMSE por subclases para cada Territorio Histórico, el cálculo de las estimaciones por C.A. de Euskadi es directo, ya que ahora se cumple la hipótesis de independencia. A partir de los cálculos por subclases los RMSE de la variable A84 o de cualquier otra agrupación se obtienen de forma directa, es decir, calculando la raíz cuadrada de la suma de los MSE de las subclases que forman cada sector.

4. Aplicación de técnicas de estimación en áreas pequeñas a la Encuesta Industrial de la C.A. de Euskadi. 2002-2003

A continuación se presentan las estimaciones obtenidas utilizando el sistema de estimación antes expuesto en la Encuesta Industrial de la C.A. de Euskadi correspondientes a los años 2002 y 2003. Las macromagnitudes escogidas para su publicación han sido: valor añadido a coste de factores, ventas netas, costes de personal, excedente bruto de explotación, inversión y resultado antes de impuestos.

Se han ofrecido, también, las estimaciones del Empleo de las comarcas dado que ésta ha sido la variable exógena utilizada en los modelos de áreas pequeñas utilizados en la Encuesta Industrial de la C.A. de Euskadi.

En lo que sigue únicamente se detallan las estimaciones publicadas para las variables 'valor añadido bruto a coste de factores' y 'ventas netas'.

	2002	cv	2003	cv	Δ 03/02
C.A. de Euskadi	**13008214**	**0,01**	**13371649**	**0,01**	**2,8**
Alava	**2676614**	**0,00**	**2747477**	**0,00**	**2,6**
Valles Alaveses	89743	0,03	106061	0,02	18,2
Llanada Alavesa	1708733	0,01	1763392	0,01	3,2
Montaña Alavesa	17455	0,19	14211	0,16	-18,6
Rioja Alavesa	243155	0,02	270594	0,04	11,3
Estribaciones del Gorbea	216196	0,02	210776	0,01	-2,5
Cantábrica Alavesa	401332	0,01	382443	0,01	-4,7
Bizkaia	**5371448**	**0,01**	**5561854**	**0,01**	**3,5**
Arratia-Nervión	252658	0,02	259622	0,01	2,8
Gran Bilbao	3259555	0,01	3395279	0,01	4,2
Duranguesado	1109460	0,01	113048	0,01	0,3
Encartaciones	154460	0,05	155544	0,04	0,7
Gernika-Bermeo	180951	0,03	199627	0,02	10,3
Markina-Ondarroa	149595	0,03	167664	0,02	12,1
Plentzia-Mungia	264769	0,01	271070	0,01	2,4
Gipuzkoa	**4960152**	**0,01**	**5062318**	**0,01**	**2,1**
Bajo Bidasoa	218659	0,02	231525	0,02	5,9
Bajo Deba	482592	0,02	482471	0,02	0,0
Alto Deba	996600	0,01	1013516	0,00	1,7
Donostia-San Sebastián	1585031	0,01	682139	0,01	0,4
Goierri	655356	0,01	682139	0,01	4,1
Tolosa	422004	0,02	442768	0,01	4,9
Urola Costa	599910	0,01	618822	0,01	3,2

Cuadro 1: Valor añadido a coste de factores de la industria y coeficiente de variación (cv) por territorio histórico y comarca (miles de euros

	2002	cv	2003	cv	Δ 03/02
C.A. de Euskadi 42393031	**0,01**	**43768410**	**0,01**	**3,2**	
Alava	**8336340**	**0,01**	**8614962**	**0,01**	**3,3**
Valles Alaveses	321016	0,02	390287	0,02	21,6
Llanada Alavesa	5521407	0,01	5564479	0,01	0,8
Montaña Alavesa	46122	0,17	40494	0,17	-12,2
Rioja Alavesa	603353	0,02	702555	0,03	16,4
Estribaciones del Gorbea	705855	0,02	763485	0,01	8,2
Cantábrica Alavesa	1138586	0,01	1153661	0,01	1,3
Bizkaia	**19341307**	**0,01**	**20078450**	**0,01**	**3,8**
Arratia-Nervión	743317	0,03	822265	0,02	10,6
Gran Bilbao	12733700	0,01	13268830	0,01	4,2
Duranguesado	3518902	0,02	3562602	0,01	1,2
Encartaciones	422018	0,05	417302	0,03	-1,1
Gernika-Bermeo	673681	0,03	727858	0,02	8,0
Markina-Ondarroa	495976	0,03	524603	0,03	5,8
Plentzia-Mungia	753713	0,03	754989	0,02	0,2
Gipuzkoa	**14715384**	**0,01**	**15074998**	**0,01**	**2,4**
Bajo Bidasoa	581630	0,03	630369	0,02	8,4
Bajo Deba	1273758	0,03	1238176	0,04	-2,8
Alto Deba	3134142	0,01	3202536	0,01	2,2
Donostia-San Sebastián	4600028	0,01	4557629	0,01	-0,9
Goierri	2039115	0,01	2224916	0,01	9,1
Tolosa	1239249	0,04	1315686	0,02	6,2
Urola Costa	1847462	0,02	1905687	0,03	3,2

Cuadro 2: Ventas netas de la industria y coeficiente de variación (cv) por territorio histórico y comarca (miles de euros

5. Conclusiones

Los modelos de áreas pequeñas descritos en este documento permiten obtener estimaciones de totales de las diferentes macromagnitudes de la Encuesta Industrial a nivel de comarcas. Los resultados obtenidos son altamente satisfactorios con carácter general, dado que los coeficientes de variación estimados son ciertamente moderados en aquellas comarcas en las que la representación muestral y el tamaño poblacional es relativamente elevado. En comarcas donde esto no ocurre, los coeficientes de variación estimados son mayores.

Eustat ofrecerá a partir del presente año estimaciones comarcales de la Encuesta Industrial a un nivel de desagregación mayor o igual que el presentado en este documento con carácter anual. Esto permitirá disponer, en un futuro cercano, de series temporales de estimaciones comarcales, que permitirán realizar análisis coyunturales más completos que los actuales, facilitando un mayor conocimiento de la evolución de las principales macromagnitudes económicas dentro de la C.A. de Euskadi. Toda esta información podrá ser de gran utilidad en la confección de las diferentes políticas territoriales.

Eustat sigue investigando la metodología de los modelos de estimación en áreas pequeñas con objeto de dar más adelante un paso más y poder ofrecer estimaciones de calidad a un nivel de desagregación aún mayor. Se está también trabajando en la utilización de esta metodología en otras encuestas de Eustat, no sólo en encuestas relacionadas con el ámbito industrial, si no también en encuestas donde todas o parte de las variables a estimar son discretas

6. Contactar

Haritz Olaeta Goiriena

Dirección:

Área de Metodología
Eustat
Donostia-San Sebastián 1 01005 - Vitoria-Gasteiz
Spain

Teléfono: (+34) 945017522

Fax: (+34) 945017501

e-mail: H_Olaeta@eustat.es

www: http://www.eustat.es/

7. Bibliografía

[1] Battese, G.E., Harter, R.M and Fuller, W.A. (1988). An Error-Components Model for Prediction of Country Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association* 83, 28-36.

[2] Rao, J.N.K. (2003). *Small Area Estimation.* Wiley Series in Survey Methodology.

[3]] Särndal, C.E., Hidirogloy, M.A. (1989). Small Domain Estimation: A Conditional Analysis. *Journal of the American Statistical Association* 84, 266-275.

[4] Searle, S.R.., Casella, G. and McCullogh, C.E. (1992). *Variance Components.* Wiley Series in Probability and Statistics.