ESTIMACIÓN EN ÁREAS PEQUEÑAS. UNA APLICACIÓN A LA ENCUESTA INDUSTRIAL DE LA COMUNIDAD AUTÓNOMA DE EUSKADI

Iosune Azula, Patxi Garrido y Haritz Olaeta

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

ESTIMACIÓN EN ÁREAS PEQUEÑAS. UNA APLICACIÓN A LA ENCUESTA INDUSTRIAL DE LA COMUNIDAD AUTÓNOMA DE EUSKADI

Iosune Azula, Patxi Garrido y Haritz Olaeta

Toledo, junio de 2004

Indice

Capítulo

1

Introducción

Eustat ha formado un equipo investigador compuesto por miembros de diferentes departamentos de Eustat 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. Este equipo está supervisado por las profesoras Ana Fernández Militino y Lola Ugarte de la Universidad Pública de Navarra.

Este proyecto de investigación sigue en curso y en este documento describimos parte del trabajo realizado hasta el momento en la Encuesta Industrial de la Comunidad Autónoma de Euskadi. Se está trabajando en otros campos de aplicación, sobre todo en el campo de la estimación en áreas pequeñas.

Este trabajo está dividido en tres partes. Se comienza describiendo brevemente la Encuesta Industrial de Eustat (marco utilizado, diseño muestral, etc.). En el segundo apartado, se proponen estimadores alternativos al utilizado actualmente en la Encuesta Industrial. Partiendo del estimador directo de Horvitz-Thompson se proponen estimadores más complejos que hacen uso de información externa. Finalmente, se pasa al estudio de estimadores de áreas pequeñas, proponiendo un primer modelo explícito y mostrando las estimaciones que se obtienen a nivel de comarcas de la C.A. de Euskadi.

Las diferentes propuestas de estimadores que se presentan en este trabajo no son si no los que se han estudiado en la primera fase del proyecto de colaboración con las profesoras de la Universidad Pública de Navarra. El proyecto sigue en activo (actualmente se está trabajando en estimadores compuestos y en nuevos modelos de áreas pequeñas) por lo que presumiblemente los estimadores aquí propuestos serán mejorados en un futuro cercano y será entonces cuando se decida qué tipo de indicadores utilizar en la Encuesta Industrial.

Capítulo
2

Encuesta industrial

Antecedentes

Esta operación se puso en marcha en 1981, teniendo desde su creación como objetivo fundamental el conocimiento pormenorizado del entramado industrial vasco, dada su importancia tanto en términos de valor añadido como de empleo. La información básica para ello se obtiene a partir de las principales partidas de la cuenta de pérdidas y ganancias, y la consiguiente estimación, a partir de ellas, de las principales macromagnitudes.

Esta operación estadística se realiza en colaboración con el Servicio de Estadística y Análisis Sectorial del Departamento de Agricultura y Pesca, Organo Estadístico específico de dicho Departamento.

Características Técnicas

Ambitos

Universo. El ámbito poblacional se circunscribe a aquellos establecimientos cuya actividad principal, medida en términos de valor añadido generado, sea industrial.

Incluye, según la Clasificación Nacional de Actividades Económicas de 1993 (en adelante CNAE-93), las siguientes secciones:

- Sección C: Industrias extractivas

- Sección D: Industria manufacturera

- Sección E: Producción y distribución de energía eléctrica, gas y agua

Geográfico. Las unidades estadísticas que estén ubicadas en el ámbito geográfico de la C.A. de Euskadi, aun cuando su sede social o gerencia se encuentre fuera de ella.

Temporal. El período de referencia es el ejercicio económico del año natural. Excepcionalmente, de presentarse establecimientos cuya contabilidad vaya referida a períodos de tiempo que no correspondan al año natural, se referirá la información a los ejercicios que finalizan dentro de los años correspondientes.

Marco de la encuesta

El marco de la encuesta es el Directorio de Actividades Económicas de Eustat. Su utilización permite la elaboración de un muestreo probabilístico que acote los errores muestrales.

Unidad Estadística

La unidad estadística es el establecimiento definido como una unidad que ejerce, exclusiva o principalmente, una o varias actividades situada en un mismo emplazamiento geográfico.

Diseño muestral y extrapolación

Se realiza un muestreo probabilístico en dos fases: una primera en la que se seleccionan con probabilidad "uno" todas las unidades que tengan más de 19 empleados; en la segunda fase, se realiza un muestreo aleatorio estratificado donde las variables de estratificación son:

- **Territorio Histórico**

  - Araba

  - Bizkaia

  - Gipuzkoa

- **Actividad**: Clasificación Nacional de Actividades Económicas (CNAE-93) a nivel de subclase, es decir, a 5 dígitos. Posteriormente para su difusión se utiliza la clasificación normalizada de EUSTAT A84. La clasificación A84 es una desagregación de la A60 (CNAE-93 a 2 dígitos) en función de la estructura económica de la C.A. de Euskadi.

El tamaño de la muestra seleccionada es de 3.000 unidades estadísticas, aproximadamente.

Previamente a la extrapolación, se post-estratifican los establecimientos muestrales, según los tres Territorios Históricos (Araba, Bizkaia, Gipuzkoa) , subclase de la CNAE-93 y 5 tamaños de establecimientos, que son:

1. Entre 1 y 19 empleados
2. Entre 20 y 49 empleados
3. Entre 50 y 99 empleados
4. Entre 100 y 499 empleados
5. Mayores o iguales a 500 empleados.

El paso de datos muestrales a los poblacionales se realiza a través de una matriz de elevadores por cada estrato. La variable utilizada para la obtención de los elevadores ha sido el número de ocupados de los establecimientos industriales. El uso de esta variable está justificado en que es la más correlacionada con las principales variables económicas que intenta medir la encuesta.

En el presente trabajo se utilizan los datos muestrales correspondientes a la Encuesta Industrial del año 2000.

Capítulo
3

Estimación de sectores A84 por TH

Introducción

Actualmente en la Encuesta Industrial se utiliza un método de estimación indirecta utilizando como información auxiliar el empleo para aquellos establecimientos con 20 o menos empleados (para establecimientos mayores la encuesta es censal).

En lo que sigue, mostraremos las estimaciones obtenidas utilizando diferentes métodos de estimación para el valor añadido bruto a coste de factores de las empresas de menos de 20 empleados en el sector 9 de la clasificación A84 (Minerales no metálicos) de Eustat. Este sector ha sido escogido al azar para ilustrar en el presente trabajo los diferentes estimadores.

La información muestral de la que disponemos (la correspondiente a la Encuesta Industrial del año 2000) se resume en la Tabla 1. Dado que para estratos de empleo superiores la encuesta es censal, en este trabajo haremos únicamente referencia al estrato de empleo de 1 a 19 empleados. La encuesta está diseñada para obtener estimaciones de los sectores A84 por Territorio Histórico para todos los estratos de empleo. Por consiguiente, los coeficientes de variación aquí presentados son sustancialmente mayores que los correspondientes a las estimaciones publicadas (el estrato de empleo de 1-19 empleados supone aproximadamente el 23% del VABcf y el 31% del empleo total industrial).

Tabla1. Información muestral de establecimientos de menos de 20 empleados del sector 9.- Minerales no metálicos

Código	CNAE-93	TH	VABcf	Empleo muestral	Empleo poblacional
6903	14210	1	235	5	
6994	14210	1	1129	11	
ARABA			**1364**	**16**	**66**
6109	14111	20	480	12	
6502	14111	20	631	19	
GIPUZKOA			**1111**	**31**	**185**
6996	14210	48	408	4	
6997	14210	48	408	4	
6999	14210	48	408	4	
BIZKAIA			**1224**	**12**	**191**
C.A. EUSKADI					**442**

Estimación directa

En el muestreo clásico o basado en el diseño no necesitamos hipótesis específicas sobre la distribución de la población de interés. Únicamente requerimos conocer la probabilidad de extraer una muestra cualquiera, o equivalentemente, la probabilidad de inclusión de un elemento o la fracción de muestreo.

Probablemente el estimador directo más utilizado en los Institutos de Estadística es el estimador de Horvitz-Thompson.

Estimador de Horvitz-Thompson:

El estimador de Horvitz-Thompson para el total en la población de estudio de y, T_y, con el diseño muestral π se define como:

$$\hat{T}_{y,s,HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} w_i y_i ,$$

donde s es una muestra $(y_1,\ldots,y_n)$ de tamaño n extraída de la población U formada por los elementos $(y_1,\ldots,y_N)$ con N conocida[1], π_i es la probabilidad de inclusión del elemento y_i en la muestra s bajo el diseño y w_i es el peso muestral de y_i (en caso de muestreo aleatorio simple sin reposición, $\pi_i = \frac{n}{N}$ y $w_i = \frac{N}{n}$).

La varianza del estimador $\hat{T}_{y,HT}$ viene dada por:

$$\mathrm{var}(\hat{T}_{y,HT}) = \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j},$$

donde π_{ij} es la probabilidad conjunta de inclusión de los elementos y_i e y_j en la muestra s bajo el diseño π (en caso de muestreo aleatorio simple sin reposición $\pi_{ij} = \frac{n}{N} \frac{n-1}{N-1}$). Dado que el estimador de Horvitz-Thompson es un estimador insesgado, el error cuadrático medio coincide exactamente con la varianza.

Un posible estimador insesgado de la varianza y por lo tanto del error cuadrático medio es el siguiente:

[1] En lo que sigue se omitirá el subíndice s por comodidad.

$$\hat{\text{var}}(\hat{T}_{y,HT}) = \sum_{i\in s}(1-\pi_i)\frac{\hat{y}_i^2}{\pi_i^2} + \sum_{i\in s}\sum_{i\neq j}(\frac{1}{\pi_i\pi_j} - \frac{1}{\pi_{ij}})y_i y_j \ .$$

Dado que nuestro interés es obtener un estimador no sólo para el total de la población si no para los diferentes dominios que la componen (sectorización A84 por Territorio Histórico), el estimador de Horvitz-Thompson para un dominio cualquiera d_j viene obviamente dado por:

$$\hat{T}_{yd_j,HT} = \sum_{k=1}^{n_{d_j}} w_k y_k = \sum_{k=1}^{n_{d_j}} \frac{N_{d_j}}{n_{d_j}} y_k \ ,$$

donde la segunda igualdad es únicamente cierta cuando el muestreo es aleatorio simple sin reposición en el dominio, siendo N_{d_j} y n_{d_j} el tamaño poblacional y el tamaño muestral del dominio d_j .

No es difícil comprobar que cuando el muestreo es aleatorio simple sin reposición, la varianza estimada para el dominio d_j adquiere la siguiente forma más sencilla:

$$\hat{\text{var}}(\hat{T}_{yd_j,HT}) = N_{d_j}^2 (1-\frac{n_{d_j}}{N_{d_j}})\frac{1}{n_{d_j}}\text{var}_{s_{d_j}}(y_k)$$

donde $\text{var}_{s_{d_j}}(y_k)$ es la cuasi-varianza muestral calculada con las observaciones de la muestra que caen en el dominio:

$$\text{var}_{s_{d_j}}(y_k) = \frac{n_{d_j}}{n_{d_j}-1}\frac{\sum_{i=1}^{n_{d_j}}(y_i - \bar{y}_{d_j})^2}{n_{d_j}} .$$

Los resultados obtenidos para las empresas de 19 o menos empleados del sector 9 de la clasificación A84 se muestran en la Tabla 2.

Tabla 2. Estimaciones de Horvitz-Thompson para establecimientos de menos de 20 empleados del sector 9.- Minerales no metálicos

Dominio	n_{d_j}	N_{d_j}	$\hat{\text{var}}(\hat{T}_{yd_j,HT})$	$\hat{ee}(\hat{T}_{yd_j,HT})$	$\hat{T}_{yd_j,HT}$	$c.v$
Araba	2	9	12587967	3547.95	**6138**	**0.578**
Bizkaia	3	24	0	0	**9792**	**0**
Gipuzkoa	2	24	3009732	1734.86	**13332**	**0.13**
C.A.Euskadi	7	57	34143103.75	5843.21	**30120.43**	**0.194**

El estimador Horvitz-Thompson es un estimador directo. No hace uso de ningún tipo de información auxiliar (del mismo dominio o de otros dominios) dado que utiliza únicamente para su cálculo la información obtenida en la muestra y los pesos de muestreo (inversos de las probabilidades de inclusión) derivados exclusivamente del diseño muestral.

Estimación asistida por modelos

Hace uso de información auxiliar de la muestra. Este tipo de estimadores utilizan modelos de regresión como un medio para conseguir estimadores consistentes desde el punto de vista del diseño.

Estimador de Regresión Generalizado GREG:

Presentamos brevemente la familia de estimadores GREG, pero no los aplicaremos directamente a la Encuesta Industrial. El motivo de incluirlos en este trabajo radica en el uso que se hará de ellos como alternativa a los estimadores de Horvitz-Thompson a la hora de construir estimadores compuestos.

La familia de estimadores GREG fue propuesto fundamentalmente por Sarndal, Swensson y Wretman (1989). La idea inicial está basada en el estimador de Horvitz-Thompson y se trata de mejorar las estimaciones mediante la utilización de datos auxiliares. Se trata de utilizar modelos de regresión como un medio para conseguir estimadores consistentes desde el punto de vista del diseño. Requieren que el muestreo sea aleatorio.

El estimador GREG se diferencia del estimador de regresión lineal habitual en que introduce pesos en la estimación de los coeficientes del modelo. Aún haciendo uso de información auxiliar no se considera específicamente diseñado para proporcionar estimaciones en áreas pequeñas.

El estimador de regresión generalizado de cada área i-ésima $T = \sum_{i=1}^{N} y_i$ cuando el modelo elegido es un modelo de regresión lineal (es decir, $E(y_i) = x_i \beta_{GREG}$) viene dado por:

$$\hat{T}_{GREG} = \sum_{i=1}^{n} w_i^* y_i = \sum_{i=1}^{n} w_i g_i y_i = \hat{T}_{HT} + (\sum_{i=1}^{N} x_i - \sum_{i=1}^{n} w_i x_i)' \hat{\beta}_{GREG} ,$$

con $\hat{\beta}_{GREG} = (\sum_{i=1}^{n} w_i x_i^{|} x_i)^{-1} \sum_{i=1}^{n} w_i x_i^{|} y_i / c_i$, donde c_i son constantes especificadas que frecuentemente toman el valor $c_i = 1$, $\forall i$.

Alternativamente,

$$\hat{T}_{GREG} = \sum_{i=1}^{N} \hat{y}_i + \sum_{i=1}^{n} w_i (y_i - \hat{y}_i) .$$

donde $\hat{y}_i$ son los valores predichos en el dominio de interés $i = 1,..., N$. En cualquier caso, los pesos w_i^* satisfacen la propiedad de calibración, es decir, aplicados a x_i reproducen exactamente la población total de x_i :

$$\sum_{i=1}^{n} w_i^* x_i = \sum_{i=1}^{N} x_i$$

La varianza viene dada por:

$$\text{var}(\hat{T}_{GREG}) = \sum_{i=1}^{N} \sum_{i=1}^{N} (\frac{w_i w_j}{w_{ij}} - 1) E_i E_j ,$$

donde $E_i = y_i - x_i^{|} \beta_{GREG}$. La estimación de la varianza es:

$$\hat{\text{var}}(\hat{T}_{GREG}) = \sum_{i=1}^{n} \sum_{i=1}^{n} (w_i w_j - w_{ij})(g_i e_j)(g_i e_j) ,$$

donde $e_i = y_i - x_i^{|} \hat{\beta}_{GREG}$.

Estimación indirecta

La principal diferencia con los estimadores directos es que para estimar el total en un dominio dado se utilizarán observaciones de fuera de dicho dominio. Se suele decir que el estimador toma información prestada. Los estimadores directos pueden hacer uso de información auxiliar pero muestran como punto débil que se restringen al tamaño muestral efectivo del dominio. El número de observaciones en algunos dominios puede ser muy pequeño, por lo que las varianzas de los estimadores puede ser muy grande, por lo que las estimaciones pueden ser erráticas para dominios pequeños.

Estimador sintético

Se dispone de un vector auxiliar multidimensional $x_k = (x_{1k},...,x_{jk},...,x_{Jk})^{|}$, que proporciona información relevante sobre la variable de interés y_k . De forma general, se modela la relación entre estas variables mediante el modelo ξ de forma que:

$$E_\xi (y_k) = f(x_k \mid \beta)$$

donde $f(.|.)$ es una función lineal o no lineal cualquiera. Un modelo ξ particular utilizado frecuentemente (justificada o injustificadamente), especialmente para variables continuas, es el de regresión lineal:

$$y_k = x_k^{|}\beta + \varepsilon_k$$

con $E_\xi(\varepsilon_k) = 0$, para todo $k \in U$.

El primer paso del procedimiento es obtener un estimador adecuado de β, procedimiento que dependerá de las propiedades individuales y conjuntas de x_k y de ε_k. Una vez obtenido $\hat{\beta}$, parece natural estimar la variable de interés para todos los dominios como:

$$\hat{y}_k = f(x_k \mid \hat{\beta})$$

De esta forma, el estimador sintético del total de la variable y en un dominio d_j dado es:

$$\hat{T}_{yd_j,SYN_j} = \sum_{k \in d_j} \hat{y}_k$$

Para el caso particular de la Encuesta Industrial, la única información adicional para estimar el VABcf es el empleo de los establecimientos. Tras un análisis exploratorio de la relación de la variable VABcf y el Empleo se ha optado por un modelo ξ de regresión lineal sin intercepto (x es por tanto simplemente el empleo) común para los tres Territorios Históricos.

No es difícil comprobar que para este caso, el estimador sintético es en este caso:

$$\hat{T}_{yd_j,SYN} = T_{xd_j}\hat{\beta} = T_{xd_j}\frac{\hat{T}_{yd}}{\hat{T}_{xd}},$$

donde $T_{xd_j} = \sum_{k=1}^{n_{d_j}} x_k$, $\hat{T}_{xd} = \sum_{k=1}^{n_d} w_k x_k$, $\hat{T}_{yd} = \sum_{k=1}^{n_d} w_k y_k$, $w_k = \frac{N_d}{n_d}$ siendo $n_d = \sum_{j=1}^{J} n_{d_j}$ y $N_d = \sum_{j=1}^{J} N_{d_j}$ (con $J = 3$ siendo el número de territorios históricos). Notad que $\hat{T}_{xd}$ y $\hat{T}_{yd}$ no son más que estimadores de Horvitz-Thompson.

El cálculo de las varianzas de los estimadores sintéticos no es tarea fácil. Para el caso concreto que nos ocupa la varianza estimada se puede aproximar mediante:

$$\hat{\text{var}}(\hat{T}_{yd_j,SYN}) \approx T_{xd_j}^{|}\hat{T}^{-1}\hat{V}\hat{T}^{-1}T_{xd_j},$$

donde $\hat{T} = \sum_{k=1}^{n_d} w_k x_k x_k^{|} / c_k$, con $c_k = x_k$ escalar y por tanto $\hat{T} = \sum_{k=1}^{n_d} x_k$ también escalar y

$$\hat{V} = N_d^2 (1 - \frac{n_d}{N_d}) \frac{1}{n_d} \mathrm{var}(e)$$

,

donde $\mathrm{var}(e)$ es la cuasi-varianza muestral de los residuos del modelo $y_k = \beta x_k + \varepsilon_k$ para $k = 1,...,n_d$.

Este estimador sintético es sesgado si bien el sesgo será limitado cuando el modelo asumido ajusta adecuadamente (será insesgado en el caso en el que $\sum_{k=1}^{n_d} \varepsilon_k = 0$).

Rao (2003), propone la siguiente aproximación del error cuadrático medio para todo tipo de estimadores sintéticos:

$$MSE(\hat{T}_{yd,SYN}) \approx mse_M = \mathrm{var}(\hat{T}_{yd,SYN}) + N_d^2 b_a^2 (\hat{\bar{T}}_{yd,SYN})$$

,

donde $b_a^2 (\hat{\bar{T}}_{yd,SYN}) = mse_a (\hat{\bar{T}}_{yd,SYN}) - \frac{1}{m} \sum_{j=1}^{m} \mathrm{var}(\hat{\bar{T}}_{yd_j,SYN})$, N_d es el número de observaciones en el dominio, m es el número de dominios y además:

$$\mathrm{var}(\hat{T}_{yd_j,SYN}) = \frac{\mathrm{var}(\hat{T}_{yd_j,SYN})}{N_{d_j}^2},$$

$$mse_a (\hat{\bar{T}}_{yd,SYN}) \approx \frac{1}{m} \sum_{j=1}^{m} \frac{1}{N_{d_j}^2} (\hat{T}_{yd_j,SYN} - \hat{T}_{yd_j,HT})^2 - \frac{1}{m} \sum_{j=1}^{m} \frac{1}{N_{d_j}^2} \mathrm{var}(\hat{T}_{yd_j,HT})$$

Los resultados obtenidos para el sector 9 de la clasificación A84 se muestran en la **Tabla 3**.

Dominio	n_{d_j}	N_{d_j}	$\hat{\mathrm{var}}(\hat{T}_{yd_j,SYN})$	$\hat{MSE}(\hat{T}_{yd_j,SYN})$	$\hat{RMSE}(\hat{T}_{yd_j,SYN})$	$\hat{T}_{yd_j,SYN}$	$c.v$
Araba	2	9	846934.72	977169.82	988.52	**4137.86**	**0.237**
Bizkaia	3	24	7092981.08	7315444.24	2704.71	**11974.73**	**0.226**
Gipuzkoa	2	24	6654348.22	7347474.4	2710.62	**11598.56**	**0.233**
CAE	7	57	37984516.77	37984516.77	6163.16	**27711.15**	**0.222**

Estimadores compuestos

El estimador compuesto se construye para compensar el sesgo del estimador indirecto, frente a la inestabilidad de los estimadores directos:

$$\hat{T}_{yd_j,C} = \phi_{yd_j}\hat{T}_{yd_j,D} + (1-\phi_{d_j})\hat{T}_{yd_j,I}$$

donde $0 \le \phi_{d_j} \le 1$, $\hat{T}_{yd_j,D}$ es un estimador directo y $\hat{T}_{yd_j,I}$ es un estimador indirecto.

El error cuadrático medio de un estimador compuesto puede expresarse como:

$$MSE(\hat{T}_{yd_j,C}) = \phi_{d_j}^2 MSE(\hat{T}_{yd_j,D}) + (1-\phi_{yd_j,I})^2 MSE(\hat{T}_{yd_j,I}) + 2\phi_{d_j}(1-\phi_{d_j})E(\hat{T}_{yd_j,D} - T_{yd_j})(\hat{T}_{yd_j,I} - T_{yd_j})$$

Una forma de escoger el peso ϕ_{d_j} es minimizar $MSE(\hat{T}_{yd_j,C})$, aunque en este trabajo únicamente mostraremos los pesos utilizados en Eustat hasta el momento.

Dos de los estimadores compuestos que se han aplicado a la Encuesta Industrial de la Comunidad Autónoma de Euskadi han sido los siguientes:

$$\hat{T}_{yd_j,C_1} = \phi_{yd_j}\hat{T}_{yd_j,HT} + (1-\phi_{d_j})\hat{T}_{yd_j,SYN}$$
$$\hat{T}_{yd_j,C_2} = \phi_{yd_j}\hat{T}_{yd_j,GregDif} + (1-\phi_{d_j})\hat{T}_{yd_j,SYN}$$
$$\phi_{yd_j} = \frac{n_{d_j}}{N_{d_j}}$$

donde en ambos casos, se ha utilizado . Los errores cuadráticos medios son en este caso los siguientes:

$$MSE(\hat{T}_{yd_j,C_1}) = \phi_{d_j}^2 MSE(\hat{T}_{yd_j,HT}) + (1-\phi_{yd_j,I})^2 MSE(\hat{T}_{yd_j,SYN})$$
$$MSE(\hat{T}_{yd_j,C_2}) = \phi_{d_j}^2 MSE(\hat{T}_{yd_j,GregDif}) + (1-\phi_{yd_j,I})^2 MSE(\hat{T}_{yd_j,SYN})$$

Los resultados obtenidos se representan en las Tablas 4 y 5.

Tabla 4. Estimaciones obtenidas con el estimador compuesto 1 para los establecimientos con menos de 20 empleados del sector 9.- Minerales no metálicos.

Dominio	n_{d_j}	N_{d_j}	$\hat{MSE}(\hat{T}_{yd_j,C1})$	$\hat{RMSE}(\hat{T}_{yd_j,C1})$	$\hat{T}_{yd_j,C1}$	$c.v$
Araba	2	9	1212751.6	1101.25	**4582.34**	**0.240**
Bizkaia	3	24	5600937.6	2366.63	**11701.89**	**0.20**
Gipuzkoa	2	24	6194872.1	2488.95	**11743.01**	**0.212**
C.A. Euskadi	7	57	29742953.0	5453.71	**28007.03**	**0.195**

Tabla 5. Estimaciones obtenidas con el estimador compuesto 2.

Dominio	n_{d_j}	N_{d_j}	$\hat{MSE}(\hat{T}_{yd_j,C2})$	$\hat{RMSE}(\hat{T}_{yd_j,C2})$	$\hat{T}_{yd_j,C2}$	$c.v$
Araba	2	9	686743.69	828.7	**4468.67**	**0.185**
Bizkaia	3	24	5600937.6	2366.63	**12913.14**	**0.183**
Gipuzkoa	2	24	6176219.0	2485.2	**11184.53**	**0.222**
CAE	7	57	29800899.0	5459.02	**27711.15**	**0.197**

Conclusiones

Analizando los resultados obtenidos se ve clara la necesidad de utilizar información auxiliar (empleo) para la estimación del VABcf para sectores A84 por Territorio Histórico. Los mejores resultados se obtienen con los estimadores compuestos y actualmente se está trabajando en esta línea (búsqueda de nuevos estimadores, cálculo de los errores cuadráticos medios asociados, etc.).

Capítulo

4

Estimación de áreas pequeñas

Las técnicas de Las técnicas de estimación en áreas pequeñas son necesarias cuando el tamaño muestral es insuficiente y hace imposible el uso de otro tipo de estimadores debido al tamaño de los errores asociados a las estimaciones.

Utilizaremos como ilustración los resultados obtenidos para el sector 49[2] de la clasificación A84, el sector de Fabricación de muebles. Se ha escogido este sector por considerar a priori que se trata de un sector muy heterogéneo en la C.A. de Euskadi. Modelos capaces de capturar adecuadamente esta heterogeneidad serán, con bastante seguridad, apropiados para capturar la variabilidad en sectores más homogéneos.

Tabla 6. Información poblacional y muestral de establecimientos con menos de 20 empleados del sector 49.- Fabricación de muebles.

Comarca	Tamaño poblacional	Tamaño muestral
Arabako Lautada	110	10
Gorbeia Inguruak	3	0
Arabako Ibarrak	1	0
Arabako Mendialdea	2	0
Errioxa Arabarra	5	1
Kantauri Arabarra	17	0
ARABA	138	11
Bilbo Handia	424	17
Plentzia-Mungia	12	0
Gernika-Bermeo	13	0
Markina-Ondarroa	10	0
Durangaldea	47	0
Arrati-Nerbioi	5	0
Enkartazioak	60	0
BIZKAIA	571	17
Donostialdea	171	5
Urola Kostaldea	111	2
Deba Beherea	8	0
Bidasoa Beherea	41	10
Tolosaldea	17	2
Goierri	21	0
Deba Garaia	14	1
GIPUZKOA	383	20

[2] Únicamente para los establecimientos con menos de 20 empleados, dado que para estratos superiores la información es censal.

Como se observa en la Tabla 6 la información muestral es ciertamente insuficiente en muchas de las comarcas de la Comunidad Autónoma de Euskadi, por lo que tanto los estimadores directos como los asistidos por modelos no son válidos, dado que incluso cuando la estimación es posible ésta será muy errática.

En lo que sigue, utilizaremos un modelo explícito capaz de capturar y utilizar información que de otra forma queda diluida. En particular, presentaremos un modelo lineal mixto, es decir, un modelo lineal con efectos aleatorios que explica la variabilidad entre áreas que no ha podido explicarse con las variables auxiliares (empleo) del modelo. Se trata de un primer modelo que, tras un análisis exploratorio previo, se ha aplicado a la Encuesta Industrial de la C.A. de Euskadi.

La línea de investigación sigue abierta y se está actualmente trabajando con otro tipo de modelos. Es importante notar que cualquier resultado derivado de la aplicación de modelos en la estimación depende directamente de la bondad de dicho modelo. Es completamente imprescindible realizar estudios preliminares sobre la elección de variables auxiliares y de la forma funcional del modelo. De igual forma, hay que tener mucho cuidado con los supuestos realizados sobre las distribuciones de las variables aleatorias.

Modelo lineal general en áreas pequeñas

Debido a que el enfoque estadístico en la estimación en áreas pequeñas basadas en modelos es radicalmente distinto al utilizado en la estadística basada en el diseño muestral (asistido por modelos o no), introduciremos el modelo lineal general con bastante detalle.

La población de estudio (el dominio puede ser un área pequeña o no) está compuesta por N elementos donde cada elemento tiene asociado un valor de una variable de interés y^* (en nuestro caso tenemos N establecimientos y cada uno de ellos lleva asociado su Valor Añadido Bruto a coste de factores).

El vector poblacional[3] $y = (y_1,...,y_N)^|$ es tratado como una realización particular del vector aleatorio $Y = (Y_1,...,Y_N)^|$. El objetivo es estimar una combinación lineal de y's, $\gamma^|y$ donde $\gamma = (\gamma_1,...,\gamma_N)^|$ es un vector de N constantes. Por ejemplo, con esta notación, si definimos $\gamma = (1,...,1)^|$, $\gamma^|y$ es el total poblacional y si en cambio definimos $\gamma = (\frac{1}{N},...,\frac{1}{N})^|$ entonces $\gamma^|y$ es la media poblacional.

Seleccionamos de la población una muestra s de n unidades. Al resto de unidades (no muestreadas) las denotamos por r, cuyo tamaño es obviamente $N-n$. Sin pérdida de generalidad, reordenamos los elementos de la población y de modo que los n primeros elementos sean las muestrales, es decir,

[3] La notación adoptada no diferencia entre escalares y vectores.

$y = (y_s^|, y_r^|)$. De igual forma, dividimos el vector γ en partes correspondientes a la muestra y a la parte no muestreada, $\gamma = (\gamma_s^|, \gamma_r^|)$.

El objetivo es pues estimar $\gamma^| y = \gamma_s^| y_s^| + \gamma_r^| y_r^|$, que es una realización de la variable $\gamma^| Y = \gamma_s^| Y_s^| + \gamma_r^| Y_r^|$.

Estudiamos el problema de predicción bajo el modelo lineal general M:
$E_M(Y) = X\beta$,
$Var_M(Y) = V$,
donde:

$$X = \begin{bmatrix} X_s \\ X_r \end{bmatrix},\ V = \begin{bmatrix} V_{ss} & V_{sr} \\ V_{rs} & V_{rr} \end{bmatrix},$$

siendo X_s de dimensión $n * p$, X_s de dimensión $(N-n) * p$, V_{ss} de $n * n$, V_{sr} de $(N-n) * (N-n)$, V_{rs} de $n * (N-n)$ y $V_{rs} = V_{sr}$. Asumimos que V_{ss} es definida positiva. Se supone en lo que sigue que se dispone de los valores de las variables auxiliares para todos los elementos de la población.

Definición. El estimador $\hat{\theta}$ es insesgado, o insesgado bajo la predicción, para θ bajo el modelo M si $E_M(\hat{\theta} - \theta) = 0$.

Definición. La varianza del error, varianza de la predicción o error cuadrático medio, de $\hat{\theta}$ bajo el modelo M es $E_M(\hat{\theta} - \theta)^2$.

El teorema general de la predicción proporciona el predictor BLUE (mejor predictor lineal insesgado) de $\hat{\theta}$ bajo el modelo lineal general M que minimiza la varianza del error:

$$\hat{\theta}_{opt} = \gamma_s^| Y_s + \gamma_r^| (X_r \hat{\beta} + V_{rs} V_{ss}^{-1} (Y_s - X_s \hat{\beta})),$$

donde $\hat{\beta} = (X_s^| V_{ss}^{-1} X_s)^{-1} X_x^| V_{ss}^{-1} Y_s$ es el estimador de mínimos cuadrados generalizados.

La varianza del error de $\hat{\theta}_{opt}$, es decir, el Error Cuadrático Medio viene dado por:

$$\begin{aligned} \mathrm{var}_M(\hat{\theta}_{opt} - \theta) &= \gamma_r^| (V_{rr} - V_{rs} V_{ss}^{-1} V_{sr}) \gamma_r \\ &+ \gamma_r^| (X_r - V_{rs} V_{ss}^{-1} X_s)(X_s^| V_{ss}^{-1} X_s)^{-1} (X_r - V_{rs} V_{ss}^{-1} X_s)^| \gamma_r \end{aligned}$$

Corolario. Bajo el modelo M, si $V_{rs} = 0$ entonces:
$\hat{\theta}_{opt} = \gamma_s^| Y_s + \gamma_r^| X_r \hat{\beta}$,
y la varianza del error es:

$$\mathrm{var}_M(\hat{\theta}_{opt} - \theta) = \gamma_r^| (V_{rr} + X_r (X_s^| V_{ss}^{-1} X_s)^{-1} X_r^|) \gamma_r$$

Aplicación al sector 49.- Fabricación de muebles

Suponemos el siguiente modelo a nivel de unidad para el Valor Añadido Bruto a coste de factores del establecimiento j de la comarca d:

$$y_{dj} = \beta_0 + \beta_1 x_{dj} + v_{dj} = \beta_0 + \beta_1 x_{dj} + u_d + e_{dj}$$

donde $d = 1,...,D$ donde D es el número de áreas (20 comarcas en nuestro caso), $j = 1,...,N_d$ donde N_d es el número de elementos poblacionales en el área d, y es el Valor Añadido Bruto a coste de factores, x es el empleo, β_0 y β_1 son parámetros desconocidos, u_d es el efecto aleatorio de la comarca d y e_{dj} es un error aleatorio específico del elemento j de la comarca d. Se supone, además, que $u_d \sim iidN(0,\sigma_u^2)$, $e_d \sim iidN(0,\sigma_e^2)$ y que u_d y e_{dj} son independientes, por lo que la estructura de covarianza de los términos aleatorios es:

$$\begin{aligned} E(v_{dj}v_{pq}) &= \sigma_u^2 + \sigma_e^2 && \text{si } d = p, j = q \\ &= \sigma_u^2 && \text{si } d = p, j \neq q \\ &= 0 && \text{si } d \neq p \end{aligned}$$

Esta matriz de varianzas y covarianzas es desconocida, por lo que el estimador BLU no se puede aplicar directamente. Habrá que utilizar un estimador empírico EBLUP ("Empirical Best Linear Unbiased Predictor") que únicamente lo derivaremos para el modelo que estamos utilizando.

Existen en la literatura diferentes procedimientos para estimar los componentes de varianza σ_u^2 y σ_e^2: el método de los momentos, el método de máxima verosimilitud y el método de máxima verosimilitud restringida. Describimos brevemente el primero de ellos (en el caso que nos ocupa los las estimaciones obtenidas son similares utilizando cualquiera de los tres métodos).

La estimación de los componentes de varianza σ_u^2 y σ_e^2 se realiza mediante el método de ajuste de constantes:

- $\hat{\sigma}_e^2 = \dfrac{1}{n-t-p+1}\sum_{d=1}^{t}\sum_{j=1}^{n_d}\hat{\varepsilon}_{dj}^2$, donde $\hat{\varepsilon}_{dj}$ son los residuos de la regresión lineal ordinaria de $y_{dj} - \bar{y}_d$ sobre $x_{dj} - \bar{x}_d$.

$$\hat{\sigma}_u^2 = \max\left(\frac{1}{n^*}\left(\sum_{d=1}^{t}\sum_{j=1}^{n_d}\hat{v}_{dj}^2 - (n-p)\hat{\sigma}_e^2\right),0\right)$$ donde $\hat{v}_{dj}$ son los residuos de la regresión lineal ordinaria de y_{dj} sobre x_{dj} y la variable z_{dj} que toma el valor 1 para el área d y 0 para el resto de áreas y n^* es la traza de la matriz $MZZ^|$, donde M es la matriz de proyección idempotente $(I - X(X^|X)^{-1}X^|)$

El estimador BLU de los efectos fijos de la regresión y su matriz de varianzas y covarianzas vienen dados por:

$$\hat{\beta} = (X^{|}V^{-1}X)^{-1}X^{|}V^{-1}Y$$
$$\mathrm{cov}(\hat{\beta}) = (X^{|}V^{-1}X)^{-1}$$

La variable $\hat{\gamma}_d$ mide la incertidumbre en la modelización del predictor, que satisface:

$$\hat{\gamma}_d = \frac{\hat{\mathrm{cov}}(u_d, \bar{v}_{d.})}{\hat{\mathrm{var}}(\bar{v}_{d.})} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_d}}$$

Dado que la media poblacional $\bar{x}_{d(p)}$ es conocida, se puede demostrar que el predictor EBLUP de la media (el del total se obtiene trivialmente, multiplicando la media por el número de elementos en el dominio) viene dado por:

$$\hat{\bar{y}}_d = \bar{x}_{d(p)}\hat{\beta} + (\bar{y}_{d.} - \bar{x}_{d.}\hat{\beta})\hat{\gamma}_d = \hat{\gamma}_d\bar{y}_{d.} + (\bar{x}_{d(p)} - \hat{\gamma}_d\bar{x}_{d.})\hat{\beta}$$

El cálculo de errores cuadráticos medios en este caso complicado. Existen diferentes propuestas en la literatura, partiendo la mayoría de supuestos de normalidad. Se ha calculado, para este caso, la estimación del error cuadrático medio propuesto por Prasad y Rao:

$$\hat{MSE}(\hat{y}_d) = g_{1d}(\hat{\sigma}^2) + g_{2d}(\hat{\sigma}^2) + 2g_{3d}(\hat{\sigma}^2),$$

donde:

$$g_{1d}(\hat{\sigma}^2) = (1-\gamma_d)\sigma_u^2,$$
$$g_{2d}(\hat{\sigma}^2) = (\bar{x}_{d(p)} - \gamma_d\bar{x}_d)^{|}(X^{|}V^{-1}X)^{-1}(\bar{x}_{i(p)} - \gamma_i\bar{x}_i),$$
$$g_{3d}(\hat{\sigma}^2) = n_d^{-1}(\sigma_u^2 + \frac{\sigma_e^2}{n_d})^{-3}h(\sigma^2),$$
$$h(\sigma^2) = \sigma_e^4\,\mathrm{var}(\tilde{\sigma}_u^2) + \sigma_u^4\,\mathrm{var}(\tilde{\sigma}_e^2) - 2\sigma_e^2\sigma_u^2\,\mathrm{cov}(\tilde{\sigma}_e^2\tilde{\sigma}_u^2),$$

y bajo el supuesto de normalidad de u_d y e_{dj},

$$\mathrm{var}(\tilde{\sigma}_e^u) = 2(n-t-k)^{-1}\sigma_e^4,$$
$$\mathrm{var}(\tilde{\sigma}_u^2) = 2n_*^{-1}((n-t-k)^{-1}(t-1)(n-k-1)\sigma_e^4\sigma_u^2 + n_{**}\sigma_u^4,$$
$$\mathrm{cov}(\tilde{\sigma}_e^2\tilde{\sigma}_u^2) = -(t-1)n_*^{-1}vr(\tilde{\sigma}_u^2),$$
$$n_* = tr(MZZ^{|}) \text{ y } n_{**} = tr(MZZ^{|})^2,$$
$$M = I - X(X^{|}X)^{-1}X^{|},$$

y Z es la matriz de diseño de los efectos aleatorios.

Las estimaciones obtenidas con el correspondiente coeficiente de variación (medido en términos de la raíz del error cuadrático medio) se recogen en la Tabla 7.

Tabla 7. Estimación del VABcf de los establecimientos de menos de 20 empleados del sector 49 (Fabricación de muebles) por comarcas.

Comarca	Tamaño poblacional	Tamaño muestral	Estimación	RMSE	cv
Arabako Lautada	110	10	**9327**	1721.3	**0.18**
Gorbeia Inguruak	3	0	**120**	149.3	**0.83**
Arabako Ibarrak	1	0	**175**	49.6	**0.27**
Arabako Mendialdea	2	0	**226**	99.1	**0.39**
Errioxa Arabarra	5	1	**234**	192.4	**0.64**
Kantauri Arabarra	17	0	**1733**	842.7	**0.42**
Bilbo Handia	424	17	**26114**	5117.8	**0.19**
Plentzia-Mungia	12	0	**419**	597.4	**0.90**
Gernika-Bermeo	13	0	**1438**	644.3	**0.39**
Markina-Ondarroa	10	0	**943**	495.8	**0.45**
Durangaldea	47	0	**3656**	2332.1	**0.52**
Arrati-Nerbioi	5	0	**440**	248.0	**0.47**
Enkartazioak	60	0	**6248**	2974.1	**0.41**
Donostialdea	171	5	**9234**	3748.0	**0.36**
Urola Kostaldea	111	2	**29980**	3516.9	**0.11**
Deba Beherea	8	0	**841**	396.5	**0.41**
Bidasoa Beherea	41	10	**4881**	641.1	**0.13**
Tolosaldea	17	2	**1449**	530.7	**0.35**
Goierri	21	0	**1779**	1041.6	**0.49**
Deba Garaia	14	1	**1232**	535.7	**0.40**

Conclusiones

Los errores cuadráticos medios obtenidos en las estimaciones por comarca son en algunos casos elevados[4]. Esto implica que hay que dirigir esfuerzos a la búsqueda y especificación de modelos que capturen mejor la variabilidad de los establecimientos y las diferentes comarcas. Eustat está trabajando actualmente en este sentido.

[4] Recordar que no es publican estimaciones por comarcas y estratos de empleo, y que únicamente un 23% del VABcf de la industria es generado por los establecimientos con menos de 20 empleados.

Capítulo

5

Bibliografía

[1] DREW, D., SINGH, M.P. y CHOUDHRY, G.H.

Evaluation of Small Area Estimation Techniques for the Canadian Labor Force Survey". Survey Methodology, 8, pp. 14-47 (1982).

[2] GOVINDARAJULU, Z.

Elements of Sampling Theory and Applications. PrenticeHall (1999)

[3] SÄRNDAL, C.E. y HIDIROGLOU, M.A.

Small Domain Estimation: A conditional Analysis. Journal of the American Statistical Association, 84, pp. 166-175 (1989)

[4] SÄRNDAL, C.E.., SWENSSON, B y WRETMAN, J.

Model Assisted Survey Sampling. Springer (2003)

[4] SCHAIBLE, W.L.

Choosing Weights for Composite Estimators for Small Area Statistics. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp. 741-7476 (1978)