TÉCNICAS DE ESTIMACIÓN EN ÁREAS PEQUEÑAS

JESÚS MANCHO CORCUERA

JUNIO 2002

TÉCNICAS DE ESTIMACIÓN EN ÁREAS PEQUEÑAS

Jesús MANCHO CORCUERA

Becario de colaboración en EUSTAT

RESUMEN

Las Técnicas de Estimación en Áreas pequeñas persiguen la obtención de estimadores de las medias y de los totales de las variables poblacionales de ciertas áreas, haciendo uso para ello de unos datos muestrales recogidos atendiendo a un diseño muestral en el que dichas áreas no recibieron una consideración específica, sino que son entidades contenidas en los estratos del diseño muestral. También se realiza un fuerte uso de la información auxiliar disponible, tanto la referente a los dominios como a las unidades muestrales.

El presente documento pretende ser una introducción a estos métodos, con especial atención a los basados en modelos, y servir para la divulgación y la extensión en la aplicación de los mismos.

La estructura del Cuaderno va a ser la siguiente:

En la *Introducción* se describe el ámbito de aplicación de estas técnicas y se ponen algunos ejemplos de situaciones en las que resultan de interés.

En el *Capítulo Segundo* se realiza una clasificación y una descripción de los distintos tipos de estimadores.

A lo largo del *Capítulo Tercero* se desarrollan de forma teórica los estimadores basados en modelos de unidad.

En el *Cuarto Capítulo* se da un ejemplo de aplicación de los "estimadores sintéticos" y de los "estimadores basados en modelos" a la obtención de estimaciones municipales de las principales variables de la Encuesta Industrial.

Finalmente, se presentan una serie de *Consideraciones generales* que afectan a las distintas fases de toda operación estadística y que habría que tener en cuenta para explotar al máximo las posibilidades de estos métodos.

Lo aquí recogido es fruto del trabajo realizado durante el disfrute de la Beca de formación e investigación en metodologías estadístico-matemáticas que me fue concedida en el año 2000 por el Euskal Estatistika Erakundea / Instituto Vasco de Estadística.

Agradezco al Área de Económicas, y en particular a Patxi Garrido, la colaboración prestada, la ayuda y el apoyo de todos los componentes del Área de Metodología y, en general, la amabilidad de todo el personal de EUSTAT.

Mi agradecimiento también para Marina Ayestarán, quien me ha guiado en estas lides y ha participado activamente en la elaboración de este Cuaderno.

PALABRAS CLAVE: Área pequeña, Diseño muestral, Estimador directo, Estimador sintético, Estimadores basados en modelos.

Indice

Capítulo

1

Introducción

La estimación en áreas pequeñas es un tema que ha despertado gran interés en los últimos años debido, fundamentalmente, a la importancia que tanto para el sector público como para el privado tiene la obtención de información fiable acerca de los dominios en torno a los que estas técnicas centran su atención.

Inmediatamente asociada a la expresión "área pequeña" está el problema de la ausencia de una muestra significativa. Comúnmente se usa esta expresión para referirse a áreas geográficas pequeñas, como municipios o comarcas, o a pequeñas subpoblaciones, como desempleados, juventud, minusválidos, minorías étnicas...

Las principales fuentes de datos acerca de estas subpoblaciones han venido siendo –y aún hoy en día lo son– los censos y los registros administrativos, pero la falta de un aprovechamiento adecuado de este último recurso, así como la necesidad de conocer una amplia variedad de aspectos económicos y sociales con relativa frecuencia han conducido al desarrollo de una amplia metodología en este campo.

Cabría situar los antecedentes de la estimación en áreas pequeñas en los métodos demográficos, que desde hace décadas vienen utilizándose para la estimación de la población en períodos intercensales o atender la falta de cobertura u otras carencias que pueden sufrir los censos y los padrones.

Muestra del creciente interés que existe sobre este tema es el gran número de congresos y simposiums organizados por los distintos institutos de estadística y organismos estadísticos internacionales, así como la gran cantidad de artículos publicados sobre éste y temas afines en las principales publicaciones estadísticas.

Los países donde más extendida está la utilización de metodologías para la obtención de información estadística sobre subpoblaciones son Australia, Canadá, Estados Unidos e Italia. Todos estos países han desarrollado una estrategia global entorno a su Encuesta de Población Activa que les permite obtener una información más desagregada. Especialmente reseñables resultan los sistemas de información que para cubrir las demandas de datos de áreas pequeñas han puesto en práctica Statistics Canada y el que nutre a los programas federales en los Estados Unidos.

EUSTAT, en su empeño por proporcionar a la sociedad una información estadística lo más exhaustiva posible, que describa de forma pormenorizada la realidad social y económica de Euskadi, ha abordado la cuestión de realizar estimaciones comarcales y municipales en algunas de sus encuestas.

La celebración en el año 2000 del curso "Metodología estadística para estimaciones indirectas en áreas pequeñas", dentro del *Seminario Internacional de Estadística en Euskadi* que anualmente organiza EUSTAT, es un ejemplo más del interés de este organismo por difundir e implantar en sus producciones estadísticas dichas metodologías. El encargado de impartir dicho curso fue el profesor Rao, una de las figuras más destacables de cuantos en la actualidad se dedican al estudio de los métodos de estimación en áreas pequeñas.

Capítulo
2

Descripción de los estimadores

En este apartado se abordan los distintos tipos de estimadores que existen para la evaluación de los totales o las medias de las variables en áreas pequeñas, realizando una somera descripción de los distintos tipos.

Existen tres grandes clases de estimadores: los estimadores basados en el diseño, los estimadores indirectos tradicionales y los basados en modelos. En los primeros, la escasez de muestra hace que tengan grandes varianzas, mientras que las otras dos clases de estimadores necesitan de una *buena* información auxiliar que permita "relacionar" las áreas pequeñas.

Dentro de los estimadores indirectos tradicionales se encuentran los métodos demográficos, que, por la importancia histórica que han tenido en el desarrollo de métodos de estimación en áreas pequeñas, merecen una mención especial.

El error medio cuadrático (MSE) es la medida comúnmente usada para determinar la acuracidad del estimador. Viene dada por:

$$\text{MSE}(\hat{Y}) = V(\hat{Y}) + \left[B(\hat{Y})\right]^2$$

Estimadores basados en el diseño

Son estimadores insesgados, pero suelen tener grandes varianzas debido al pequeño tamaño de la muestra, situación típica de las áreas pequeñas.

- Estimadores directos

  Están basados únicamente en los datos de la muestra para el área pequeña, pudiendo hacer uso de información auxiliar proveniente de censos o registros administrativos.

  Dentro de este tipo de estimadores cabe incluir a los siguientes:

  - Estimador expansivo:

    $\hat{Y}_{e,a} = \sum_{i \in S_a} w_i y_i$, donde S_a es la muestra del área pequeña a y w_i es el peso muestral de la unidad i.

- Estimador post-estratificado:

  Requiere el conocimiento del tamaño de la subpoblación, N_a.

$$\hat{Y}_{pst,a} = N_a \left. \sum_{i \in S_a} w_i y_i \middle/ \sum_{i \in S_a} w_i \right. = N_a \hat{Y}_{e,a} / \hat{N}_{e,a} = N_a \hat{\bar{y}}_{e,a}$$

  Si el diseño muestral es estratificado y se conocen los tamaños de la población en los estratos, h, del área pequeña a, un estimador post-estratificado alternativo al anterior es:

$$\hat{Y}_{st,pst,a} = \sum_h \left(N_{h,a} \left. \sum_{i \in S_{h,a}} w_i y_i \middle/ \sum_{i \in S_{h,a}} w_i \right. \right) = \sum_h N_{h,a} \hat{Y}_{h,e,a} / \hat{N}_{h,e,a} = \sum_h N_{h,a} \hat{\bar{y}}_{h,a}$$

- Estimador ratio:

  Requiere conocer el valor de la variable auxiliar x en el área pequeña a.

  $\hat{Y}_{r,a} = X_a \hat{R}_a$, donde $\hat{R}_a = \left. \hat{Y}_{e,a} \middle/ \hat{X}_{e,a} \right.$ es un estimador del ratio $\left. Y_a \middle/ X_a \right.$.

- Estimador regresión:

$$\hat{Y}_{r,a} = \hat{Y}_a + \hat{b}_a \left(X_a - \hat{X}_a \right),$$

$$\text{con } \hat{b}_a = \sum_{i \in S_a} u_i^{-1} \, w_i y_i x_i' \left\{ \sum_{i \in S_a} u_i^{-1} \, w_i x_i x_i' \right\}^{-1}.$$

  Este estimador cuantifica la diferencia entre el valor de la variable auxiliar de la subpoblación y el de la muestra mediante una relación lineal entre la variable objeto de interés, y, y la variable auxiliar, x.

- Estimadores directos modificados

  Este tipo de estimadores puede emplear datos muestrales de fuera del dominio, aunque siguen siendo insesgados. Por un estimador directo modificado entenderemos un estimador directo con un ajuste sintético para el sesgo del modelo.

$$\hat{Y}_{sreg,a} = \hat{Y}_a + \hat{b} \left(X_a - \hat{X}_a \right), \text{ con } \hat{b} = \sum_{i \in S} u_i^{-1} w_i y_i x_i' \left\{ \sum_{i \in S} u_i^{-1} w_i x_i x_i' \right\}^{-1}.$$

  Obsérvese que b se estima sobre la totalidad de la muestra, S.

La preferencia por este estimador o el anterior dependerá de lo grande que sea la varianza de $\hat{b}_a$ relativa a la variación de las b_a a lo largo de las áreas.

Estimadores indirectos tradicionales

Métodos demográficos[1]

Los demógrafos han venido utilizando desde tiempo atrás una variedad de métodos para la estimación de poblaciones locales y otras características de interés en años intercensales. Estos métodos, denominados "Técnicas con consideraciones sintomáticas" (Symptomatic Accounting Techniques), utilizan datos *sintomáticos* actuales obtenidos de registros administrativos –como número de nacimientos y defunciones– junto con datos del último censo.

Dichos métodos suelen hacer uso de la hipótesis de que existe un comportamiento *similar* entre el área pequeña y una mayor que la contiene.

- Método de los ratios vitales

  Este método usa sólo nacimientos y defunciones como variables sintomáticas. En dicho modelo se asume que las variaciones de los porcentajes de fallecimientos y de alumbramientos entre el último censo y la actualidad son iguales en el área local y un área mayor que la contiene, y en la que estos cocientes pueden estimarse sin problemas.

- Método de las componentes

  Considera los nacimientos, las defunciones y la migración neta habidos desde el último censo hasta la actualidad en el área objeto de atención. Estas cifras se obtienen de registros civiles.

- Procedimientos de regresión sintomática

  Como todos los métodos de regresión, la variable a estimar se quiere explicar mediante una relación lineal con ciertas variables auxiliares. Los coeficientes de la ecuación se estiman sobre los valores en algunas áreas, asumiéndose como válidos para el resto.

Estimadores sintéticos

Se denominan estimadores sintéticos a aquellos que son estimadores directos fiables de un área grande, que contiene a varias pequeñas, y se usa como estimador del área pequeña al considerar que ésta tiene las mismas características que el área mayor. Esta clase de estimadores tendrá una menor varianza, aunque pudiera estar fuertemente sesgado de no cumplir la hipótesis.

Veamos alguno de estos estimadores:

[1] Los métodos demográficos pueden verse ampliados en págs. 16-18 de [9].

- Si suponemos que la media del área pequeña es igual que la del área mayor que la contiene.

$$\hat{Y}_{syn,m,a} = N_a \sum_{i \in S} w_i y_i \Big/ \sum_{i \in S} w_i = N_a \hat{\bar{y}}$$

- Si suponemos que determinados ratios son iguales en el área pequeña y el área grande.

$$\hat{Y}_{syn,r,a} = X_a \hat{Y}_e / \hat{X}_e \text{, con } \hat{Y}_e = \sum_{i \in S} w_i y_i .$$

Estimadores combinados[2]

Como antes mencionaba, los estimadores sintéticos hacen un fuerte uso de la información de otras áreas, permitiendo por tanto muy poca variación local, lo que puede contribuir a un importante crecimiento del sesgo de no existir cierta homogeneidad entre estas áreas. Para evitar el potencial sesgo del estimador sintético y la inestabilidad del estimador directo, consideramos como estimador una combinación lineal convexa de ambos, esto es, $\hat{Y}_{comb,a} = l_a \hat{Y}_{des,a} + (1 - l_a)\hat{Y}_{syn,a}$, con $l_a \in [0,1]$. l_a se estima minimizando el error cuadrático medio, MSE, con los datos muestrales.

Entre los distintos tipos de estimadores combinados cabe destacar los estimadores dependientes del tamaño de la muestra y los estimadores de James-Stein.

Estimaciones basadas en modelos

La estimación basada en modelos ha recibido mucha atención en los últimos tiempos, entre otras, por las siguientes razones:

- Los métodos basados en modelos permiten una variación local a través de complejas estructuras en los modelos que relacionan las áreas pequeñas. Se pueden obtener eficientes estimadores indirectos bajo estos modelos.
- Los modelos pueden ser validados con los datos muestrales.
- Estos métodos pueden utilizarse con casos complejos, tanto con datos longitudinales como transversales.
- A diferencia de los estimadores sintéticos y los combinados, los basados en modelos permiten obtener estables medidas de variabilidad de las estimaciones para cada área.

Como ya se ha hecho notar varias veces a lo largo de este Cuaderno, el problema de la estimación en áreas pequeñas radica en la escasez de muestra –que en ocasiones incluso puede llegar a ser inexistente–, que hace que los estimadores directos no sean

[2]Los estimadores combinados son tratados extensamente en págs. 24-30 de [9].

viables. Para incrementar esa muestra, o mejor dicho, para que la información muestral de las *áreas próximas* pueda incrementar la precisión del estimador, es necesario el uso de información auxiliar para establecer modelos que permitan relacionar estas áreas. Estos modelos pueden ser implícitos, como en el caso de los estimadores sintéticos o de los combinados, o explícitos, como es el caso de los basados en modelos. A todas luces parece más acertado optar por estos últimos porque dejan patente el modelo que las enlaza, siendo así susceptible de ser validado con los datos muestrales.

Los estimadores basados en modelos pueden clasificarse en dos grandes grupos: los que se basan en modelos de área y los que lo hacen en modelos de unidad.

Estimadores basados en modelo de área

Su nombre les viene de que la información auxiliar que emplean es del área o subpoblación. Cierta información, $\vec{x} = (x_1, ..., x_p)$, está disponible para todas las áreas, muestreadas o no.

El modelo básico de área asume que la media en la subpoblación de la variable objeto de interés, $\overline{Y}_i$, o una cierta función de ésta, $q_i = g(\overline{Y}_i)$, está en relación con las variables de $\vec{x}_i$ a través del modelo lineal con efectos aleatorios J_i:

$$q_i = \vec{x}_i' \vec{b} + J_i \quad i = 1, ..., m \qquad (1),$$

donde b es el p - vector de parámetros de regresión y las J_i son incorrelacionadas, con media 0 y varianza s_J. Habitualmente se asume la normalidad de las J_i. Suponemos que este modelo también es válido en las áreas sin muestra.

Por otro lado, denotemos por $\hat{\overline{Y}}_i$ el estimador de la media de la variable y en el área i-ésima, siempre que la muestra en el área, n_i, sea mayor o igual a 1. Supongamos que:

$$\hat{q}_i = q_i + e_i \qquad (2),$$

donde $\hat{q}_i = g(\hat{\overline{Y}}_i)$ y los errores muestrales, e_i, son independientes $N(0, \Psi_i)$, con varianzas, Ψ_i, conocidas.

Combinando este modelo muestral con el "modelo relacional" (1), obtenemos el modelo lineal de área con efectos mixtos:

$$\hat{q}_i = \vec{x}_i' \vec{b} + J_i + e_i \qquad (3)$$

Obsérvese que este modelo relaciona tanto variables aleatorias del diseño muestral, e_i, como variables aleatorias del modelo, J_i. En la práctica rara vez se conocen las varianzas muestrales , Ψ_i, pero a menudo se hace un ajuste de las varianzas estimadas, $\hat{\Psi}_i$, para obtener estimaciones estables, Ψ_i^*, que son empleadas como las verdaderas varianzas.

Una ventaja del modelo de área (3) es que los pesos muestrales se consideran a través de los estimadores directos $\hat{q}_i$.

La asunción de que $\mathrm{E}(e_i \mid q_i) = 0$ en el modelo muestral (2) puede no ser cierta si el tamaño de la muestra n_i es pequeña y q_i no es una función lineal del total Y_i, aun siendo su estimador directo, $\hat{Y}_i$, insesgado. Un modelo muestral más realista sería:

$$\hat{Y}_i = Y_i + e_i^*,$$

con $\mathrm{E}(e_i^* \mid Y_i) = 0$, es decir, $\hat{Y}_i$ es insesgado como estimador del total Y_i.

El modelo básico de área puede ser ampliado a casos de errores muestrales correlacionados, dependencia espacial de los efectos aleatorios de las áreas pequeñas, series temporales y en otros aspectos.

Estimadores basados en modelo de unidad

En este caso, los valores de las unidades poblacionales, y_{ig} –referido a la unidad g del área i,– están relacionadas con las variables auxiliares $\vec{x}_{ig}$ a través del modelo lineal con efectos mixtos:

$$y_{ig} = \vec{x}_{ig}'\vec{b} + J_i + e_{ig}, \quad g = 1,\ldots,N_i;\ i = 1,\ldots,m,$$

donde las J_i son $\mathrm{N}(0, s_u^2)$ independientes de e_{ig}, $\mathrm{N}(0, s_e^2)$ y N_i es el tamaño de la población del área i-ésima. Los parámetros de interés son los totales Y_i y las medias $\overline{Y}_i$.

Se asume que los datos muestrales $\{(y_{ig}, \vec{x}_{ig}) \mid g = 1,\ldots,n_i; i = 1,\ldots,m\}$ obedecen el modelo poblacional y que conocemos las medias poblacionales de las variables auxiliares para las áreas, $\vec{\overline{X}}_i$. Esto implica que el diseño muestral es ignorable o que no existe sesgo de selección, lo que se cumple, entre otros casos, cuando todas las unidades de la misma área tienen la misma probabilidad de selección.

Para diseños más generales, la variable indicadora de la muestra debería ser independiente de y_{ig} condicionada a $\vec{x}_{ig}$. Los estimadores basados en modelos de unidad no dependen de los pesos muestrales, luego la consistencia con el diseño se pierde cuando n_i crece, a excepción de cuando el diseño es autoponderado.

Para estimar los parámetros que aparecen en estos modelos existen dos enfoques posibles entre los que tradicionalmente ha existido una fuerte controversia. Estos son el enfoque frecuentista o clásico y el bayesiano[3].

[3] Ver [1], libro de referencia sobre métodos bayesianos.

La aproximación frecuentista se basa en el muestreo repetido, esto es, en la simulación del experimento muchas veces y la extracción de inferencias para valores de los parámetros. El frecuentista condiciona sobre los parámetros y replica –integra– sobre los datos. El concepto de estimación máximo-verosímil y los p-valores de los distintos tipos de contrastes de hipótesis estadísticas pertenecen a ese análisis estadístico clásico. La mayoría del software comercial existente atiende a esta filosofía.

El enfoque bayesiano se basa en un modelo de probabilidad para los datos y unas distribuciones a priori para los parámetros –éstas incorporan información a priori sobre los parámetros–. Los parámetros a estimar se consideran variables aleatorias y la inferencia para los mismos se basa en su distribución a posteriori, esto es, condicionada a los datos observados. El bayesiano condiciona sobre los datos y replica sobre los parámetros.

La filosofía bayesiana no ha gozado de mucha aceptación hasta hace pocos años, en parte por el componente de subjetividad en relación a la distribución a priori que hay que introducir y en parte por la dificultad de su implementación en coste computacional. El desarrollo de los métodos bayesianos ha sido importante a partir de la década de los 80, favorecido en gran medida por los avances en reducción del coste y mayor rapidez de los ordenadores.

Actualmente los métodos bayesianos se han introducido en una gran cantidad de campos de la teoría y la práctica estadísticas. La virtud principal de estos métodos radica en que los procedimientos que los soportan consiguen un equilibrio adecuado entre sesgo y varianza, las dos componentes de toda estimación estadística.

Capítulo

3

Modelos a nivel unidad

La mayoría de los tratamientos que sobre estimación en áreas pequeñas basada en modelos se han hecho en la bibliografía estadística no se han preocupado suficientemente de la influencia del diseño muestral, asumiendo que no aportaba nada y, que por lo tanto, era ignorable.

Al desechar la importancia del diseño muestral se pierde quizá la más importante contribución de la aleatorización a la inferencia. Como la mayoría de modelos estadísticos de inferencia en poblaciones finitas o son *incorrectos* o, en el mejor de los casos, incompletos, es deseable que el estimador cumpla la propiedad de consistencia con el diseño, que podría expresarse en los siguientes términos: "si la muestra es lo suficientemente grande, el estimador debería aproximarse a lo que está estimando sin que importe lo correcto que sea el modelo". Como cabría deducir de estas palabras se trata de una propiedad asintótica.

Para la estimación en áreas pequeñas son preferibles los estimadores que sean consistentes con el diseño, porque ofrecen cierta protección contra los fallos del modelo. Con esta premisa, pasemos a desarrollar un estimador para la media poblacional de la variable y en una cierta área pequeña.

El modelo básico a nivel unidad es:

$$y_{ig} = \mathbf{q}_i + e_{ig},\ g = 1,\ldots,N_i;\ i = 1,\ldots,m \qquad (1)$$

donde las e_{ig} son variables aleatorias incorrelacionadas de media 0 y varianza $\mathbf{d}_i^2$.

Centrémonos en un dominio particular, j. El problema es estimar la media del dominio:

$$\bar{y}_{jP} = \sum_{k=1}^{N_j} y_{jk} / N_j .$$

Sean p_{jk} la probabilidad de selección de la unidad k y n_j el número de unidades seleccionadas del dominio j.

Como es bien sabido, una estrategia de estimación insesgada con el diseño y lineal-eficiente establecería $p_{jk} = \dfrac{n_j}{N_j}$ como probabilidad de selección y $\sum_{k=1}^{n_j} y_{jk} / n_j$ como estimador, donde las unidades de la población han sido reordenadas de tal modo que las n_j primeras son las que han sido escogidas para la muestra.

$$d_j = \sum_{k=1}^{n_j} w_{jk} y_{jk}, \text{ donde } w_{jk} = \frac{p_{jk}^{-1}}{\sum_{l=1}^{n_j} p_{jl}^{-1}}.$$

w_{jk} representa por tanto el peso muestral de la unidad k.

El estimador d_j es insesgado bajo el modelo (1), esto es, $\mathrm{E}_e\left(d_j - \bar{y}_{jP}\right) = 0.$

Bajo muchos diseños muestrales d_j también es consistente con el diseño, lo que matemáticamente se expresa: $\underset{n_j \to \infty}{plim_p}\left(d_j - \bar{y}_{jP}\right) = 0$, donde p denota el espacio probabilístico generado por el proceso de selección.

Pero d_j presenta problemas cuando n_j es muy pequeño. Una posible solución es "compartir información" entre los dominios. Un modo de hacer esto es tratar el parámetro fijo del modelo (1), q_j, como la realización de una variable aleatoria que satisface el modelo relacional:

$$q_j = m + J_j \qquad (2)$$

donde $\mathrm{E}(J_j) = 0$ y $\mathrm{E}(J_j J_k) = s^2 c_j^k$, siendo c_j^k igual a 1 si $k = j$ ó 0 si $k \neq j$.

Combinando las ecuaciones (1) y (2) resulta la forma reducida del modelo de componentes de varianza:

$$y_{jk} = m + J_j + e_{jk} \qquad (3)$$

Hemos separado el modelo básico y el relacional para subrayar el elevado nivel de confianza que ofrece este modelo.

Cualquier estimador de la forma:

$$f_j(a, \vec{c}) = (1-a) d_j + a\hat{m} \qquad (4)$$

donde $\vec{c} = (c_1, ..., c_{j-1}, 0, c_{j+1}, ..., c_m)$, tal que $\sum_{k=1}^{m} c_k = 1$, $\hat{m} = \sum_{g=1}^{m} c_g \bar{y}_{gS}$ y $\bar{y}_{gS} = \sum_{i=1}^{n_g} y_{gi} / n_g$, es un estimador insesgado bajo el modelo (3). Obsérvese que tanto $\vec{c}$ como $\hat{m}$ dependen del dominio.

Si asumimos que todas las varianzas d_i^2 son iguales a d^2, usando el método de los multiplicadores de Lagrange obtenemos los valores de a y $\vec{c}$ para los que se minimiza la varianza del modelo de $f_j(a,\vec{c})-\bar{y}_{jP}$:

$$a^* = \frac{\sum_{i=1}^{n_j} w_{ji}^2 - 1/N_j}{\sum_{i=1}^{n_j} w_{ji}^2 + \sum_{k=1}^{m} c_k^{*2}/n_k + (1+\sum_{k=1}^{m} c_k^{*2})(s^2/d^2)} \qquad (5)$$

$$c_k^* = \frac{\left[(s^2/d^2)+n_k^{-1}\right]^{-1}}{\sum_{l=1}^{m}\left[(s^2/d^2)+n_l^{-1}\right]^{-1}}, \text{ si } k \neq j \qquad (6)$$

En la práctica, s^2 y d^2 rara vez son conocidos. Un modo propuesto para estimar el ratio s^2/d^2 de forma consistente con el modelo es[4]:

$$L = máx\left\{0, \left[\frac{\sum_{k=1}^{m} n_k(\bar{y}_{kS}-\bar{y}_S)^2/(m-1)}{\sum_{k=1}^{m}\sum_{i=1}^{n_k}(\bar{y}_{ki}-\bar{y}_{kS})^2/(n-m)} - 1\right](m-1)/(n-\sum_{k=1}^{m} n_k^2/n)\right\}$$

Reemplazando en las ecuaciones (5) y (6) el ratio s^2/d^2 por L y sustituyendo éstas en la expresión (4), de la forma estimador, obtenemos el que denominaremos como 'estimador con efectos aleatorios': $b_j = f_j(a'(L),\vec{c}'(l))$, donde $\hat{m}$ en b_j es reemplazado por $m'(L)=\sum_{i=1}^{m} c_i'(L)\bar{y}_{iS}$.

Cuando el número de áreas pequeñas es muy grande, b_j es indistinguible de $f_j(a^*,\vec{c}^*)$.

Si el modelo (3) es correcto y $d_i^2 = d^2$, con $i = 1,\ldots,m$, para un número de áreas suficientemente grande, L será un número positivo. Incluso aunque el modelo no se ajuste perfectamente, al estar L acotado inferiormente por un número positivo, $|m'(L)|$ acotado y $n_j\sum_{i=1}^{n_j} w_{ji}^2$ acotado cuando n_j crece, b_j será un estimador consistente con

[4] GHOSH, M. & MEEDEN, G *Empirical Bayes estimation in finite population sampling*. Journal of the American Statistical Association, 81, págs.1058-1069. 1986.

el diseño cualquiera que sea d_j consistente con el diseño. Esto se debe a que $\underset{n_j \to \infty}{plim_p}\left[a'(L)\right]=0$, luego b_j converge hacia d_j.

Prasad y Rao, en págs. 68-69 de [7], proponen un estimador pseudo-EBLUP que en lugar de $m'(l)$, que es el mejor estimador de m bajo el modelo unidad (3) con medias $\bar{y}_{jS}$, emplea el mejor estimador de m basado en las medias ponderadas del estimador d_j.

Lo anterior puede extenderse al modelo de regresión de error anidado:

$$y_{ig} = \vec{x}'_{ig}\vec{b} + J_i + e_{ig}\,,\quad g=1,\ldots,n_i;\ \ i=1,\ldots,m \qquad (7)$$

donde x_{ig} es un p-vector de variables auxiliares relacionadas con y_{ig} y con media poblacional del área, $\vec{\bar{X}}_i$, conocida y $\vec{b}$ es el p-vector de coeficientes de regresión.

El modelo reducido viene dado por:

$$\bar{y}_{iw} = \vec{\bar{x}}'_{iw}\vec{b} + J_i + \bar{e}_{iw}\,,$$

con $\bar{y}_{iw}=\sum_{k=1}^{n_i} w_{ik}y_{ik}$, $\vec{\bar{x}}_{iw}=\sum_{k=1}^{n_i} w_{ik}\vec{x}_{ik}$ y $\bar{e}_{iw}=\sum_{k=1}^{n_i} w_{ik}e_{ik}$.

Por los resultados que se desarrollan en [6], obtenemos estimaciones de las varianzas s^2 y d^2 consistentes con el modelo (7) empleando el método de ajuste de las constantes.

El estimador pseudo-EBLUP de $q_i = \vec{\bar{X}}'_i\vec{b} + J_i$ viene dado por:

$$\hat{q}_i = \hat{g}_{iw}\bar{y}_{iw} + (1-\hat{g}_{iw})\vec{\bar{X}}'_i\hat{\vec{b}}_w\,,$$

donde $\hat{\vec{b}}_w = \left(\sum_{i=1}^{m}\hat{g}_{iw}\vec{\bar{x}}_{iw}\vec{\bar{x}}'_{iw}\right)^{-1}\left(\sum_{i=1}^{m}\hat{g}_{iw}\vec{\bar{x}}_{iw}\bar{y}_{iw}\right)$.

La expresión generalizada de los modelos a nivel unidad es:

$$y_{ig} = \vec{x}'_{ig}\vec{b} + \vec{z}'_{ig}\vec{h} + e_{ig}\,,$$

donde $\vec{x}_{ig}$ es un p-vector de variables auxiliares relacionadas con y_{ig} y $\vec{z}_{ig}$ es un q-vector de efectos aleatorios, que son los que relacionan las dominios. Los modelos de este tipo reciben el nombre de modelos lineales con efectos mixtos.

Capítulo
4

Aplicación a la Encuesta Industrial

Una vez expuestas las distintas técnicas existentes para la obtención de estimaciones en áreas pequeñas, veamos la aplicación de algunas de ellas a la obtención de estimaciones municipales de la Encuesta Industrial.

La Encuesta Industrial es una operación estadística que EUSTAT realiza anualmente y que tiene por objeto el conocimiento pormenorizado del entramado industrial vasco, atendiendo de este modo las demandas de información de los distintos agentes económicos y sociales. Estas necesidades se cubren con la estimación de las principales macromagnitudes del sector, base esencial para la elaboración de las Tablas Input-Output y de las Cuentas Económicas, así como con la elaboración de indicadores coyunturales, como el Índice de Producción Industrial (IPI) y el Índice de Precios Industriales (IPRI).

Veamos la ficha técnica de la Encuesta Industrial para situarnos en el contexto adecuado.

Descripción de la Encuesta Industrial

- Clase de operación:

  Encuesta por muestreo.

- Periodicidad:

  Anual

- Unidades de la población:

  El ámbito poblacional son aquellos establecimientos cuya actividad principal, medida en términos de valor añadido generado, sea industrial, incluyendo también todas aquellas actividades secundarias ejercidas por las empresas, bien sean industriales o de servicios.

  Estas unidades se clasifican según la Clasificación Nacional de Actividades Económicas (CNAE 93), y, según el número de empleados que tienen, se clasifican en uno de los 5 estratos de empleo existentes: estrato 1, de 1 a 19 empleados; estrato 2, de 20 a 49 empleados; estrato 3, de 50 a 99 empleados; estrato 4, de 100 a 499; y, estrato 5, de 500 ó más empleados.

- Directorio Industrial:

  El Directorio de Actividades Económicas de EUSTAT ha sido la base para establecer el marco de la encuesta. Su utilización permite la elaboración de un muestreo probabilístico que acote los errores.

- Diseño muestral:

  La distribución del número de establecimientos a encuestar se realiza atendiendo a los siguientes criterios:

  - Que sea representativo por Territorio Histórico.

  - Que sea representativo a nivel de actividad (según la CNAE 93 a 5 dígitos) y estrato de empleo.

  El número de establecimientos a encuestar está formado por todos los establecimientos censales –establecimientos con 20 ó más empleados– y un número representativo de los establecimientos muestrales para cada Territorio Histórico y cada actividad.

  Sobre el Directorio Industrial de que se dispone en el momento en que se efectúa la operación se realiza la siguiente discriminación:

  - Las empresas de estrato de empleo mayor que 1 son censales.

  - La muestra correspondiente al estrato de empleo 1 se reparte de forma proporcional a la raíz cuadrada del número de establecimientos por cada Territorio Histórico, sobrerrepresentándose de este modo Araba.

    La distribución sectorial, de acuerdo con la clasificación normalizada de EUSTAT A-84, se realiza de modo proporcional al número de establecimientos del sector por la raíz cuadrada de su empleo medio.

    El proceso finaliza distribuyendo la muestra de forma proporcional al peso de cada subclase (CNAE) en el total del sector.

    La selección de la muestra se realiza de forma aleatoria.

  El total de la muestra se compone de unas 3.000 unidades, de las que cerca de 1000 pertenecen al estrato de empleo 1. El Directorio Industrial consta de alrededor de 15.000 empresas, siendo las de estrato de empleo 1 aproximadamente unas 13.000 unidades.

- Fase de elevación:

  Se desean obtener estimaciones para los estratos formados por las variables “Territorio Histórico”, “Estrato de empleo” y “CNAE”.

  Dada la distinción realizada en el diseño muestral entre empresas del estrato de empleo 1 y empresas de estratos de empleo mayores, dicha diferenciación habrá de ser tenida en cuenta en el proceso de elevación:

  Las estimaciones para estratos en los que el estrato de empleo sea mayor que 1 será la mera agregación de los datos de las variables correspondientes a las empresas del estrato considerado.

Habrá de realizarse un tratamiento de la incidencia "no respuesta", para lo que, dependiendo de las características de la empresa de que se trate, o bien se recurre a fuentes externas, como el Registro Mercantil, o bien se realiza una imputación en función de los ratios medios del sector respecto del empleo.

- Cuando nos encontremos en el estrato de empleo 1, la estimación se realizará mediante ratios medios respecto del empleo.

  Para ello necesitamos conocer el empleo poblacional de los distintos estratos, pero, dadas las especiales características de las empresas con menos de 20 empleados, resulta difícil tener permanentemente actualizados sus datos en el Directorio, en especial el dato correspondiente a su número de empleados. Habrán de realizarse estimaciones de la variable "empleo".

  1. Primero se estima el empleo por sector de actividad (A84) y Territorio Histórico. Para ello se calcula el coeficiente de variación en el empleo entre los establecimientos que han sido encuestados dos años sucesivos, esto es:

$$CV(S,t) = \frac{\sum_{i=1}^{n} [(empleo < 20)S,t]i}{\sum_{i=1}^{n} [(empleo < 20)S,t-1]i},$$

     donde S denota el sector de actividad, t el año de la operación y $t-1$ el año anterior e i una empresa concreta.

     La estimación del empleo poblacional resulta de multiplicar el coeficiente obtenido por el empleo poblacional del año $t-1$.

  2. El empleo por subclase de actividad (CNAE) se obtiene distribuyendo el empleo estimado para su sector de actividad de modo proporcional al peso de la subclase en función del empleo en el Directorio del año t.

     La elevación de las demás variables corresponde al siguiente estimador de razón del total:

$$\hat{X} = E_j \frac{\sum_{i=1}^{n_j} X_{ij}}{e_j},$$

     donde X_{ij} el valor de la variable cuantitativa X del establecimiento i del estrato j, E_j el empleo poblacional antes calculado y e_j el empleo muestral del estrato j.

Software utilizado

SAS

SAS es un paquete estadístico que permite realizar la gestión y el análisis de los datos de un modo integrado.

Aunque permite realizar algunas tareas a través de un entorno de menús (versiones de SAS para PC), el modo usual de trabajar con SAS es utilizando su lenguaje de programación. Es un lenguaje de 5ª generación que permite realizar de manera sencilla el tratamiento específico de los datos.

Este software se compone de distintos módulos, cada uno de los cuales tiene un cometido específico, estando relacionados entre sí.

- El módulo SAS/BASE es respecto al cual se estructura el Sistema SAS. Permite la generación y el mantenimiento de ficheros de datos, la realización de informes, tabulaciones y análisis descriptivos.

- SAS/GRAPH permite la realización de análisis gráficos.

- SAS/STAT es el módulo encargado de los análisis estadísticos. Proporciona un amplio abanico de posibilidades, como el análisis de la varianza, de regresión, análisis multivariante, análisis de clusters, de modelos lineales, realización de muestras… Este software se actualiza constantemente recogiendo las metodologías más avanzadas.

Los procedimientos del módulo SAS/STAT empleados han sido PROC REG, PROC MIXED y PROC SURVEYREG. Realizo a continuación una breve descripción de los mismos.

- PROC REG

  - Permite realizar regresiones múltiples.

  - Existen 9 modos diferentes de selección de las variables regresoras para realizar el ajuste.

  - Se pueden imponer restricciones lineales a los parámetros.

  - Este procedimiento permite cierta interactividad en el momento de la ejecución.

  - Realiza tests sobre las hipótesis lineales y multivariantes.

  - Elabora diagnósticos de colinealidad, de correlación y algunos diagramas.

  - Permite obtener estimaciones, predicciones, residuos, límites de confianza y otros estadísticos de diagnóstico.

  - Genera gráficos de los datos y de algunos de los estadísticos que proporciona.

- PROC MIXED

  - Realiza análisis de varianza.

  - Permite realizar análisis de modelos lineales mixtos, generalizando de este modo al procedimiento PROC GLM, que no permite la inclusión de efectos aleatorios en el modelo.

  - Admite datos correlacionados o que puedan tener una variabilidad no constante.

  - Facilita tests estadísticos e intervalos, hace contrastes de los ajustes y las estimaciones y realiza estimaciones empírico-bayesianas.

- PROC SURVEYREG

  - Permite realizar el análisis de los datos muestrales por complejos que sean, incluyendo diseños con estratificación, clusters o con pesos muestrales diferentes.

  - El procedimiento ajusta los datos a modelos, computando los coeficientes de regresión y su matriz de covarianzas.

  - Realiza tests de significatividad para los efectos del modelo.

BUGS

Este es un programa basado en métodos de inferencia bayesianos, que como tal trata las cantidades como si fuesen variables aleatorias.

BUGS es especialmente apropiado para problemas en los que no existe una solución analítica exacta y para los que las técnicas habituales presentan dificultades.

BUGS está diseñado para el tratamiento de modelos complejos en los que pueden aparecer un gran número de cantidades desconocidas, pero para las que son aceptables ciertas suposiciones de independencia condicionada. Entre estos modelos cabe incluir los modelos lineales generalizados con efectos fijos y mixtos.

Mediante el modelo que establecemos damos lugar a una distribución conjunta sobre todas las cantidades, tanto las observadas –los datos– como las no observadas –parámetros y datos no recogidos–, condicionando la misma a los datos en orden a obtener una distribución a posteriori sobre los parámetros y los datos inobservados.

El cálculo de las marginales de esta distribución a posteriori para obtener inferencias sobre las cantidades objeto de interés se realiza mediante técnicas de Monte Carlo de integración numérica, conocidas como “muestreos de Gibbs”. Dichas integraciones se obtienen mediante un proceso de simulación.

Un pequeño número de comandos resultan suficientes para controlar la sesión en la que un modelo estadístico, implementado sencillamente con el lenguaje BUGS, es analizado.

Estimación en áreas pequeñas

Como ha quedado constatado en la exposición de los distintos métodos de estimación, la información auxiliar disponible resulta de suma importancia para los mismos. En el caso que nos ocupa esta información auxiliar proviene tanto del Directorio Industrial como de las proyecciones de empleo elaboradas a partir de la muestra.

En el Directorio Industrial se dispone de la totalidad de empresas localizadas y de algunos datos de las mismas, como su ubicación, la actividad que desarrollan, el número de empleados y, por tanto, también el estrato de empleo al que pertenecen.

En este apartado vamos a realizar el desarrollo de la obtención de estimaciones municipales para la variable "producción bruta a salida de fábrica" (PBSF) por dos métodos diferentes: mediante estimadores sintéticos y mediante estimadores basados en modelos a nivel de unidad. Únicamente nos ocuparemos del estrato de empleo 1, ya que los otros carecen de interés por tratarse de estratos censales.

La definición técnica de la variable PBSF es la siguiente: Comprende el total de los bienes y servicios producidos, así como la venta de productos fabricados, si existe actividad industrial, y cualquier otro ingreso de explotación; engloba, además, los trabajos realizados por el establecimiento para su inmovilizado y el valor de la variación de existencias de obra en curso. La valoración es a salida de fábrica, es decir, incluyendo los impuestos ligados a la actividad y excluyendo las subvenciones. Se excluye el IVA repercutido o devengado.

Estimaciones mediante estimadores sintéticos

Sobre la Encuesta Industrial realizada en 1996 el Área de Económicas de EUSTAT produjo estimaciones por municipio de las macromagnitudes que se analizan en esta encuesta. Describir la metodología empleada es el propósito de este apartado.

Como ya ha quedado patente, la dificultad con que nos encontramos cuando queremos realizar estimaciones para un dominio más desagregado que aquel para el que se ha elaborado el diseño muestral es la escasez, cuando no ausencia, de una muestra significativa que nos permita utilizar los estimadores directos sin cometer grandes errores.

En este caso, la forma de solventar el problema es estableciendo una recurrencia que nos permita acudir a niveles más agregados hasta encontrar una muestra significativa. Adoptar este modo de operar conlleva aceptar la hipótesis de que existe un grado de homogeneidad bastante elevado entre los distintos dominios. Resulta bastante intuitivo pensar que no existen grandes diferencias en los ratios de producción por empleado –las leyes del mercado y la competencia se encargan de que esto sea así– , y así lo corroboran los análisis realizados.

Para la obtención de estimaciones municipales, en primer lugar, habremos de estimar el empleo poblacional, como se hacía para obtener las estimaciones por Territorio Histórico y CNAE. Para hacer esto procederemos del mismo modo que cuando obteníamos las estimaciones de la variable "empleo" por Territorio Histórico y CNAE a partir de las estimaciones para el estrato "Territorio Histórico * Sector A84" que le contiene. Es decir, se realiza un reparto proporcional al peso de la variable 'empleo' en

el Directorio Industrial del año en cuestión. Se verifica que las estimaciones realizadas tengan cierta coherencia con los datos recogidos en cada municipio.

Una vez que tenemos estimaciones de "empleo" por Municipio y CNAE ya podemos abordar la estimación de las demás variables. Operamos del siguiente modo:

- Los datos de los establecimientos encuestados se agregarán directamente a las estimaciones del Municipio y CNAE en los que se encuentren ubicados. Por lo general no se habrán recogido en la muestra todas las empresas de un municipio que desarrollan una determinada actividad.

- La estimación para la diferencia entre el empleo estimado y el empleo muestral –esta diferencia tiene que ser obligatoriamente positiva– se obtendrá a partir de la estimación del ratio de producción media por empleado en el estrato "Euskadi * CNAE" que contenga al área en cuestión.

  En los casos en que no exista muestra en el estrato antes citado, aplicaremos la siguiente recurrencia resultante de agregar dígitos en el código CNAE de la subclase de actividad que nos encontremos estimando:

  1. Ratios de la misma subclase (5 dígitos) de Euskadi.

  2. Ratios de la misma clase (4 dígitos) de Euskadi.

  3. Ratios del mismo grupo (3 dígitos) de Euskadi.

  4. Ratios de la misma división (2 dígitos) de Euskadi.

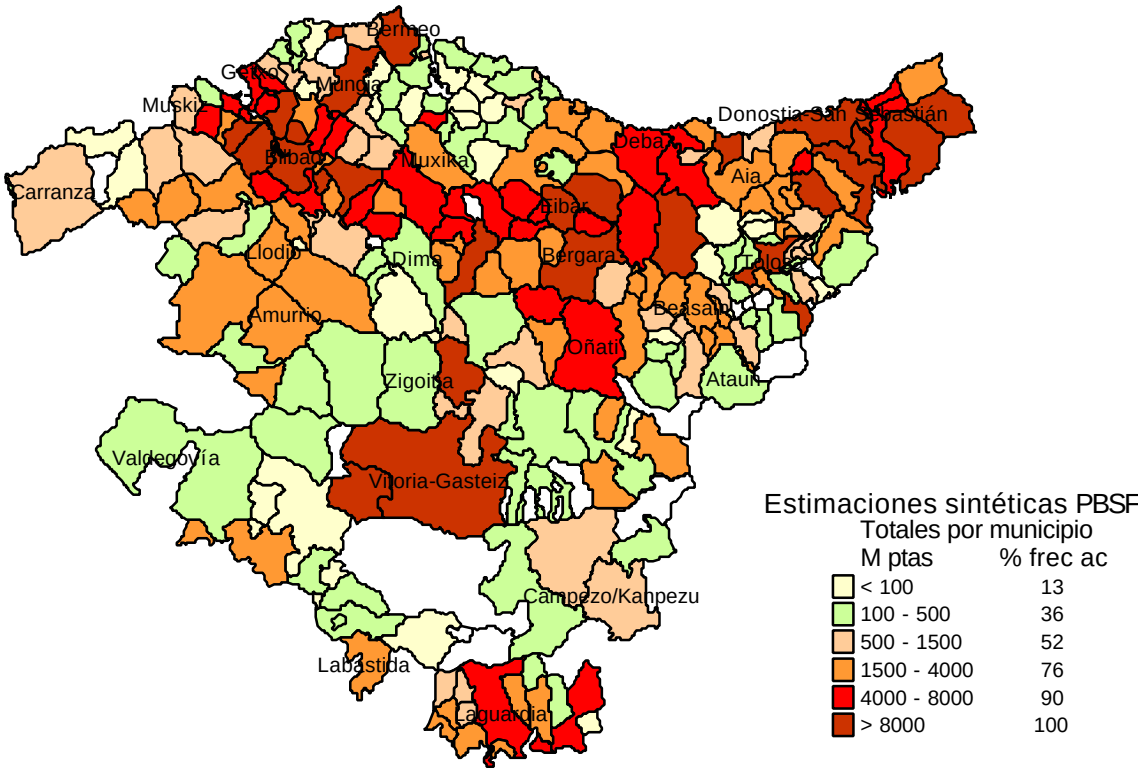
  5. Ratios de la misma actividad a un dígito de Euskadi.

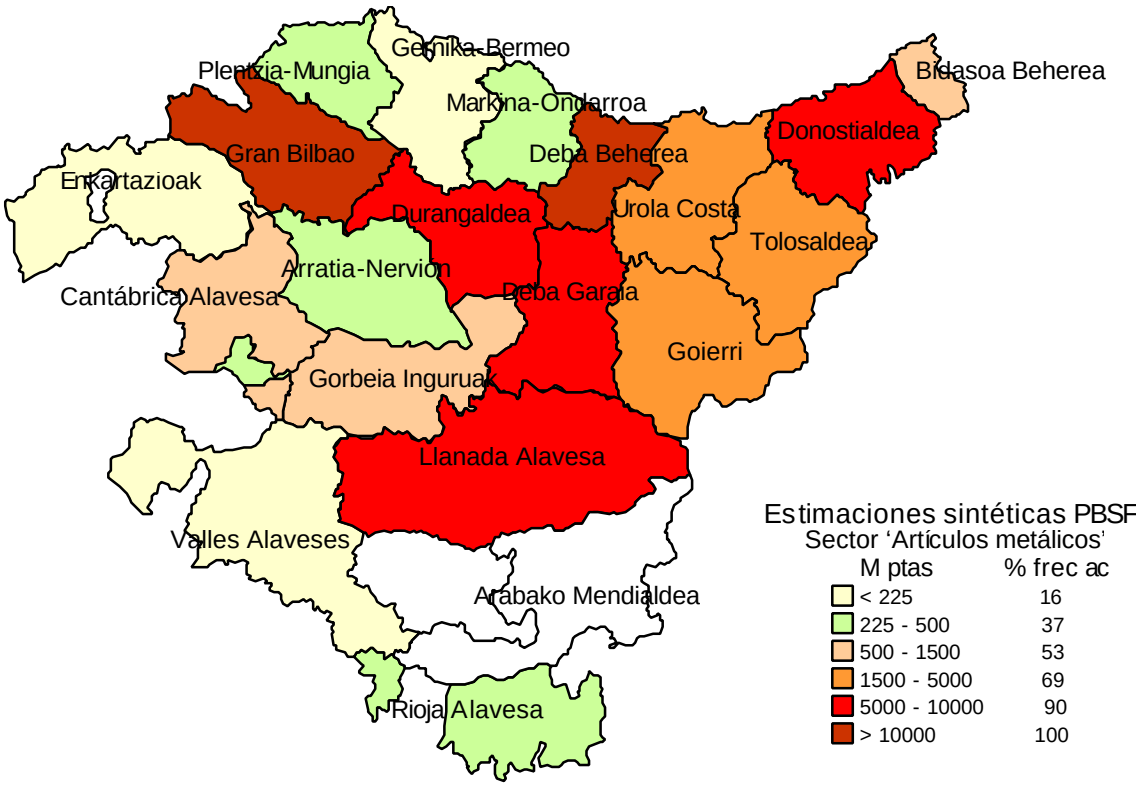
  Es al aplicar esta recurrencia cuando asumimos implícitamente la homogeneidad de la productividad entre los diferentes estratos a la que antes hacíamos referencia.

Según veíamos en la exposición de los distintos tipos de estimadores, el ahora descrito se correspondería con los estimadores sintéticos de razón. Un estimador de esta clase se caracteriza por ser un estimador directo de un área grande y utilizarse como estimador en un *área pequeña* contenida en la anterior al considerar que ésta tiene las mismas características que el área mayor.

No se ha realizado ningún estudio exhaustivo sobre la precisión de las estimaciones así obtenidas, aunque merecen la total confianza de los Técnicos que las han elaborado y supervisado.

Los siguientes mapas representan las estimaciones del total de la variable "producción bruta a salida de fábrica" por municipios y la PBSF correspondiente al sector "Artículos metálicos" por comarcas, la cual se obtiene agregando las estimaciones municipales obtenidas para dicho sector.


Bermeo
Gernika
Mungia
Muskiz
Donostia-San Sebastián
Deba
Bilbao
Muxika
Aia
Carranza
Eibar
Llodio
Dima
Bergara
Tolosa
Amurrio
Beasain
Oñati
Zigoitia
Ataun
Valdegovía
Vitoria-Gasteiz
Campezo/Kanpezu
Labastida
Laguardia
Estimaciones sintéticas PBSF
Totales por municipio
M ptas
% frec ac
< 100 13
100 - 500 36
500 - 1500 52
1500 - 4000 76
4000 - 8000 90
> 8000 100


Gernika-Bermeo
Plentzia-Mungia
Bidasoa Beherea
Markina-Ondarroa
Donostialdea
Gran Bilbao
Deba Beherea
Enkartazioak
Durangaldea
Urola Costa
Tolosaldea
Arratia-Nervion
Cantábrica Alavesa
Deba Garaia
Goierri
Gorbeia Inguruak
Llanada Alavesa
Valles Alaveses
Arabako Mendialdea
Rioja Alavesa
Estimaciones sintéticas PBSF
Sector 'Artículos metálicos'
M ptas
% frec ac
< 225 16
225 - 500 37
500 - 1500 53
1500 - 5000 69
5000 - 10000 90
> 10000 100

Estimaciones mediante modelos de unidad

La principal virtud de los estimadores basados en modelos es que se explicitan aquellas relaciones que en los estimadores indirectos aparecen de modo subyacente. Esto permite obtener estables medidas de variabilidad para cada área gracias a la utilización de la teoría de Modelos Lineales Generalizados.

La modelización de los datos se realizará con el procedimiento PROC MIXED de SAS, que permite establecer modelos lineales de efectos mixtos (también realizaremos una breve exposición acerca del uso de BUGS). Estos son una generalización de los modelos lineales en el sentido de que permiten modelar una variabilidad no constante y la correlación de los datos. Los modelos mixtos permiten modelar no sólo la media de los datos, como lo hacen los modelos lineales estándar, sino también las varianzas y covarianzas. Lo primero tiene que ver con la parte del modelo que contiene los efectos fijos y lo segundo con la parte de los efectos aleatorios. Esta clase de modelos es adecuada para datos estructurados jerárquicamente, datos longitudinales u otros casos en que los datos estén correlacionados entre sí o muestren una variabilidad diferente según ciertos grupos.

La información auxiliar utilizada procede tanto del Directorio Industrial actualizado al año que deseemos realizar las estimaciones como de las estimaciones del empleo poblacional por municipio y CNAE. Estas estimaciones de empleo poblacional se utilizan para el cálculo de los pesos de elevación que se asignan a cada unidad muestral de los estratos determinados por las variables “Territorio Histórico” y “CNAE”, y que vienen dados por la razón del empleo estimado del estrato entre su empleo muestral. El que los pesos se obtengan en función del empleo se debe a que es una variable muy correlacionada con todas las variables de la encuesta –en particular, veremos que lo está con la variable “PBSF”–.

Previamente a construir un modelo, realizamos un análisis descriptivo de las variables “producción bruta a salida de fábrica” (en adelante PBSF), cuyos agregados queremos estimar, y “empleo” (C60 en los análisis). De los datos muestrales hemos eliminado 55 registros atípicos o “outliers” –5% de la muestra del estrato de empleo 1–, que no se considerarán ni en los análisis ni en los modelos, y cuyos datos se agregarán a las estimaciones finales. Existen diversas opciones en los procedimientos de regresión de SAS que permiten determinar los “outliers”.

Los siguientes resultados muestran los estadísticos descriptivos por estrato y teniendo en cuenta los pesos muestrales:

```
                                The SAS System

---------------------------------------- ESTRATO=1 ----------------------------------------

                            The SURVEYMEANS Procedure

                                 Data Summary

                    Number of Strata                121
                    Number of Observations          933
                    Sum of Weights           121614.883




                                 Statistics

                                       Std Error        Coeff of
```

Variable	N	Mean	of Mean	Variation	Std Dev
PBSF	933	18238	1426.960579	0.078241	149495390
C60	933	1.928382	0.086304	0.044755	9849.127958

The SAS System

---------- ESTRATO=2 ----------

The SURVEYMEANS Procedure

Data Summary

Number of Strata	110
Number of Observations	735
Sum of Weights	735

Statistics

Variable	N	Mean	Std Error of Mean	Coeff of Variation	Std Dev
PBSF	735	552022	19819	0.035903	14567156
C60	735	31.775510	0.302792	0.009529	222.552393

The SAS System

---------- ESTRATO=3 ----------

The SURVEYMEANS Procedure

Data Summary

Number of Strata	92
Number of Observations	344
Sum of Weights	344

Statistics

Variable	N	Mean	Std Error of Mean	Coeff of Variation	Std Dev
PBSF	344	1669384	91177	0.054617	31364806
C60	344	70.290698	0.706066	0.010045	242.886650

The SAS System

---------- ESTRATO=4 ----------

The SURVEYMEANS Procedure

Data Summary

Number of Strata	83
Number of Observations	284
Sum of Weights	284

Statistics

Variable	N	Mean	Std Error of Mean	Coeff of Variation	Std Dev
PBSF	284	4818762	194880	0.040442	55345944
C60	284	199.239437	5.303733	0.026620	1506.260033

The SAS System

---------- ESTRATO=5 ----------

The SURVEYMEANS Procedure

Data Summary

Number of Strata	24
Number of Observations	28
Sum of Weights	28

Statistics

Variable	N	Mean	Std Error of Mean	Coeff of Variation	Std Dev
PBSF	28	40990193	702556	0.017140	19671569
C60	28	1076.750000	17.785463	0.016518	497.992972

Pasamos a analizar la relación lineal entre las variables y observamos que además esa relación es significativamente diferente por estratos de empleo. En realidad, nuestro problema de obtención de estimaciones se centra en el estrato de menor empleo, muestral, frente al resto de estratos, que son censales. En dicho estrato se obtiene un coeficiente de correlación lineal de 0.73, sin considerar los "outliers".

The SAS System

---------- ESTRATO=1 ----------

The CORR Procedure

1 With Variables: C60
1 Variables: PBSF
Weight Variable: peso

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
C60	933	1.92838	26.22394	234520	1.00000	19.00000
PBSF	933	18238	524303	2218015740	1450	772435

Pearson Correlation Coefficients, N = 933
Prob > |r| under H0: Rho=0

	PBSF
C60	0.73302 <.0001

The SAS System

---------- ESTRATO=2 ----------

The CORR Procedure

1 With Variables: C60
1 Variables: PBSF
Weight Variable: peso

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
C60	735	31.77551	8.42210	23355	20.00000	49.00000
PBSF	735	552022	611665	405736067	5068	9831453

Pearson Correlation Coefficients, N = 735
Prob > |r| under H0: Rho=0

	PBSF
C60	0.30375
	<.0001

The SAS System

---------- ESTRATO=3 ----------

The CORR Procedure

1 With Variables: C60
1 Variables: PBSF
Weight Variable: peso

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
C60	344	70.29070	13.83313	24180	50.00000	99.00000
PBSF	344	1669384	1978459	574268100	29123	19218340

Pearson Correlation Coefficients, N = 344
Prob > |r| under H0: Rho=0

	PBSF
C60	0.34450
	<.0001

The SAS System

---------- ESTRATO=4 ----------

The CORR Procedure

1 With Variables: C60
1 Variables: PBSF
Weight Variable: peso

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
C60	284	199.23944	95.30134	56584	100.00000	497.00000
PBSF	284	4818762	4390378	1368528390	230775	24802586

Pearson Correlation Coefficients, N = 284
Prob > |r| under H0: Rho=0

	PBSF
C60	0.59199 <.0001

The SAS System

---------- ESTRATO=5 ----------

The CORR Procedure

1 With Variables: C60
1 Variables: PBSF
Weight Variable: peso

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
C60	28	1077	685.92903	30149	502.00000	3156
PBSF	28	40990193	61234988	1147725400	3919494	326889519

Pearson Correlation Coefficients, N = 28
Prob > |r| under H0: Rho=0

	PBSF
C60	0.16167 0.4111

La parte básica del modelo para explicar la variable “producción bruta a salida de fábrica” es el efecto de la variable “empleo”. De los diversos modelos planteados destacamos los siguientes resultados:

- El modelo con el *empleo* como único efecto fijo explica un 54% de la variabilidad total de la variable dependiente “PBSF” –teniendo en cuenta la variable 'peso', que determina el peso de cada unidad muestral; sin tenerla en cuenta el modelo explica el 46%–.

- Considerando la estructura estratificada de los establecimientos por Territorio Histórico y sector de actividad, analizamos el efecto de estas variables en nuestra variable a explicar. Debido al escaso número de establecimientos en algunos sectores de actividad parece adecuado agregar la variable CNAE a un menor número de dígitos (los códigos de esta clasificación constan de 5 dígitos) e incluir esa variable sectorial como efecto aleatorio.

El mejor modelo se obtiene desestimando la influencia geográfica y con la agregación a 3 dígitos del código de la CNAE, variable a la que hemos denominado “z3”, incluyéndola como efecto aleatorio en la pendiente o parámetro de regresión de la variable 'empleo'. Se obtiene que el modelo así construido explica un 74% de la variabilidad total.

La especificación del módelo es la siguiente:

$$PBSF_{ig} = a + b_i C60_{ig} + e_{ig}\text{, con } e_{ig} \sim \mathrm{N}\left(0, s_e^2\right) \text{ y } b_i \sim \mathrm{N}\left(b, s_i^2\right).$$

La siguiente salida muestra el listado con los análisis que el procedimiento PROC MIXED ha realizado del modelo indicado:

The SAS System

The Mixed Procedure

Model Information

Data Set	WORK.ENC
Dependent Variable	PBSF
Weight Variable	peso
Covariance Structure	Variance Components
Subject Effect	Z3
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information

Class	Levels	Values
Z3	82	141 142 145 151 152 153 154 155 157 158 159 174 175 181 182 183 191 192 193 201 202 203 204 205 211 212 221 222 241 243 245 246 247 251 252 261 262 265 266 267 268 272 273 274 275 281 282 284 285 286 287 291 292 293 294 295 296 297 300 311 312 314 315 316 321 323 331 332 333 342 343 351 354 361 362 363 364 366 372 401 403 410

Dimensions

Covariance Parameters	2
Columns in X	2
Columns in Z Per Subject	1
Subjects	82
Max Obs Per Subject	111
Observations Used	933
Observations Not Used	0
Total Observations	933

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	22904.73935405	
1	2	22509.35566570	0.00028329
2	1	22505.75932759	0.00005760
3	1	22505.07959514	0.00000338
4	1	22505.04302788	0.00000001
5	1	22505.04287927	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
C60	Z3	56701614	11835113	4.79	<.0001
Residual		7.165E10	0	.	.

Fit Statistics

-2 Res Log Likelihood	22505.0
AIC (smaller is better)	22509.0
AICC (smaller is better)	22509.1
BIC (smaller is better)	22513.9

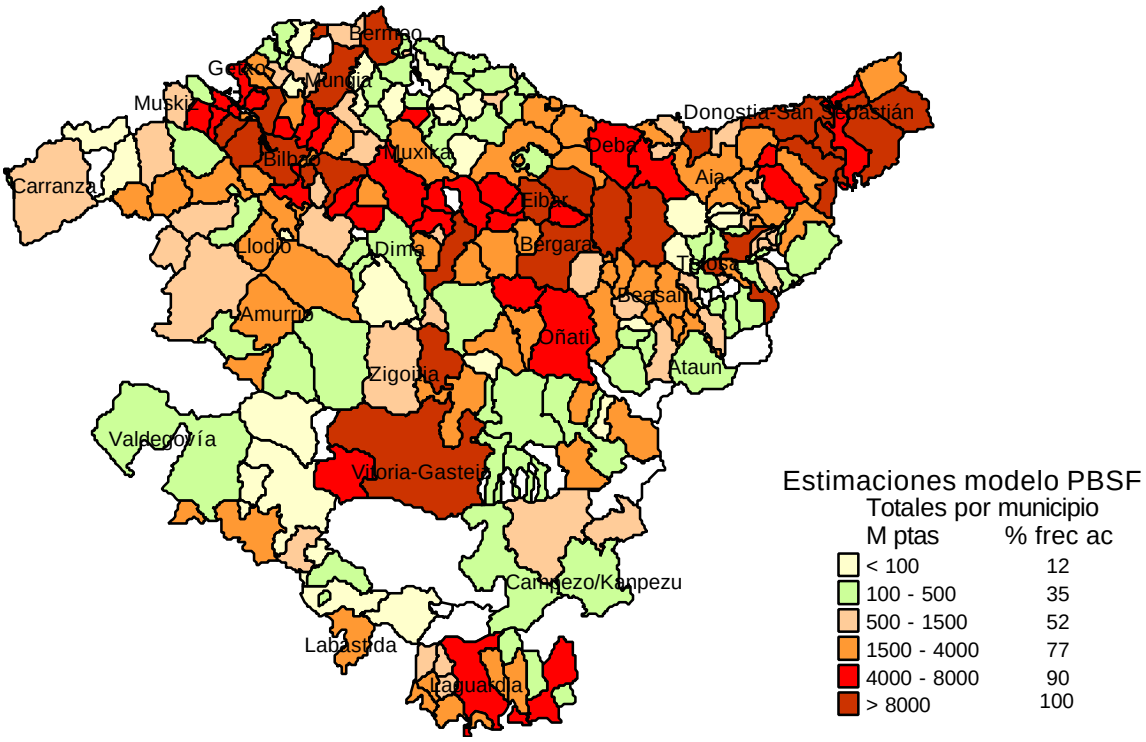
Solution for Fixed Effects

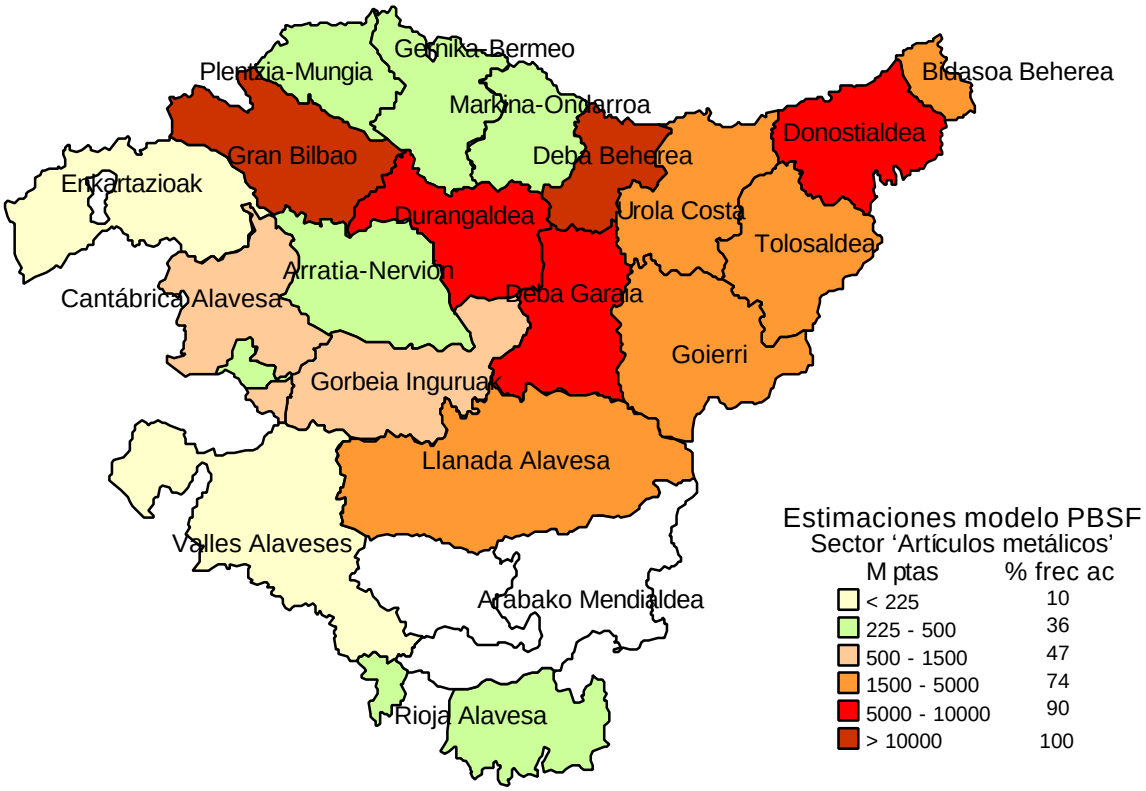
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-5082.62	1195.43	850	-4.25	<.0001
C60	15080	1069.41	81	14.10	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
C60	1	81	198.84	<.0001

Los siguientes mapas representan las estimaciones del total de la variable “producción bruta a salida de fábrica” por municipios y la PBSF correspondiente al sector “Artículos metálicos” por comarcas, la cual se obtiene agregando las estimaciones municipales obtenidas para dicho sector


Bermeo
Gernika
Muskiz
Bilbao
Muxika
Carranza
Donostia-San Sebastián
Deba
Aia
Eibar
Bergara
Ilodio
Dima
Amurrio
Oñati
Zigoitia
Ataun
Valdegovía
Vitoria-Gasteiz
Campezo/Kanpezu
Labastida
Laguardia
Estimaciones modelo PBSF
Totales por municipio
M ptas
% frec ac
< 100
12
100 - 500
35
500 - 1500
52
1500 - 4000
77
4000 - 8000
90
> 8000
100



BUGS

Un problema que surge con la utilización del procedimiento PROC MIXED es que requiere unas cantidades mínimas de muestra en cada estrato del nivel al que deseamos efectuar las estimaciones. Así, para niveles más desagregados, por los que hemos optado finalmente, resultan intratables con los métodos de estimación habitualmente implementados en los paquetes estadísticos comerciales. En este punto cabría reseñar dos cuestiones. Por una parte, habría que alcanzar ciertos compromisos a la hora de realizar el diseño muestral y la distribución de la muestra según el detalle y la desagregación de información demandadas (sobre este aspecto ahondaremos en el último capítulo de este cuaderno técnico). La otra cuestión es la utilización de métodos bayesianos para la obtención de las estimaciones.

A continuación presentamos un ejemplo de utilización del software BUGS, que en la actualidad goza de un alto grado de popularidad entre los miembros de la escuela bayesiana. Este sofware ha sido desarrollado para su empleo principalmente, en casos de estudios clínicos, investigaciones en las que las unidades que conforman la muestra están autoponderadas y todas ellas tienen la misma importancia en el universo al que representan. También es habitual en este tipo de estudios conocer las distribuciones que siguen los parámetros del modelo, hecho que no se da en general, como ocurre en el ejemplo que estamos viendo y que supone una complicación añadida a la hora de elegir las distribuciones a priori.

Se deja abierta una puerta a las posibilidades que los métodos bayesianos y, en particular, el programa BUGS –o su versión de ventanas, WinBUGS– ofrecen, y este ejemplo únicamente pretende ilustrar caminos por escrutar, por ello rogamos al lector que tome lo aquí expuesto con ciertas cautelas y de modo meramente orientativo.

El modelo implementado en BUGS es el siguiente:

$$PBSF_{ig} = empleo_{ig}\, b + J_i + e_{ig},$$

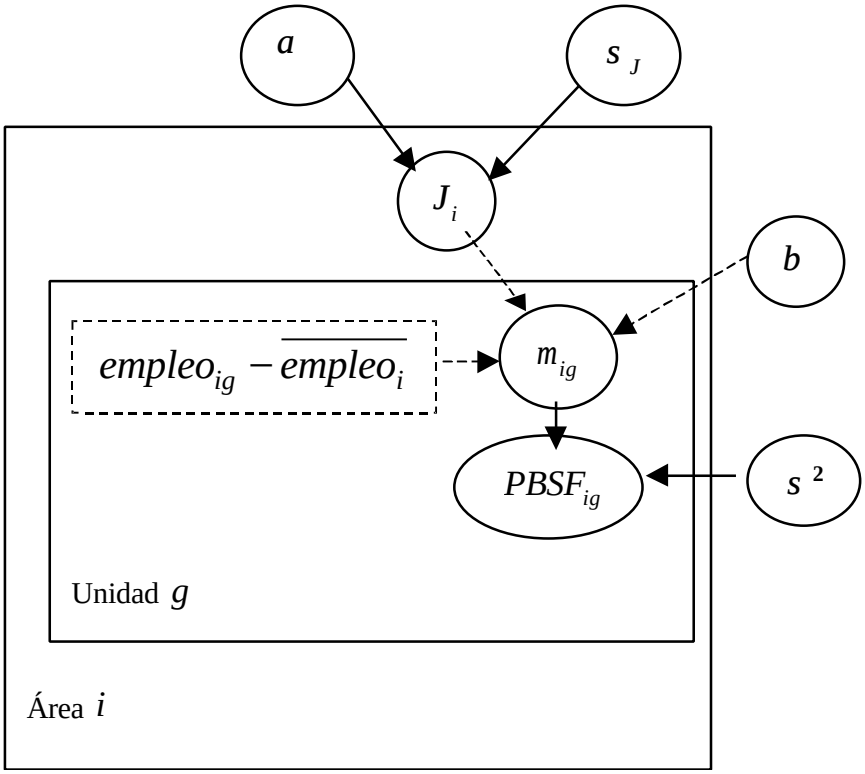
donde g indica las unidades dentro del área i, determinada por el Territorio histórico, el estrato de empleo y el sector de actividad A84.

Suponiendo que tanto el error muestral, e_{ig}, como la variable aleatoria J_i siguen distribuciones normales independientes con varianzas s^2 y s_J y medias 0 y a, respectivamente.

Podemos denotar lo anterior del siguiente modo:

$$PBSF_{ig} \sim \mathrm{N}\left(m_{ig}, s^2\right), \text{ con } m_{ig} = \left(empleo_{ig} - \overline{empleo_i}\right)b + J_i \text{ y } J_i \sim \mathrm{N}\left(a, s_J\right)$$

Este modelo puede expresarse gráficamente mediante el siguiente diagrama, muy útil para describir el modelo en el lenguaje de BUGS



En la siguiente tabla aparecen las estimaciones de los totales comarcales de la variable "producción bruta a salida de fábrica" obtenidas mediante estimadores sintéticos y mediante el modelo anterior implementado en BUGS:

COMARCAS	EST. SINTÉTICAS	EST. BUGS
Arabako Ibarrak/Valles Alaveses	30.473	26.727
Arabako Lautada/Llanada Alavesa	534.953	523.415
Arabako Mendialdea/Montaña Alavesa	4.327	4.575
Arrati-Nerbioi/Arratia-Nervión	70.559	84.519
Bidasoa Beherea/Bajo Bidasoa	58.756	63.203
Bilbo Handia/Gran Bilbao	1.491.184	1.480.520
Deba Beherea/Bajo Deba	126.939	133.638
Deba Garaia/Alto Deba	285.299	267.302
Donostialdea/Donostia-San Sebastián	475.979	505.824
Durangaldea/Duranguesado	345.246	352.547
Enkartazioak/Encartaciones	34.157	26.488
Errioxa Arabarra/Rioja Alavesa	52.133	68.898
Gernika-Bermeo	70.501	64.107
Goierri	202.989	190.222
Gorbeia Inguruak/Estribaciones del Gorbea	82.420	76.167
Kantauri Arabarra/Cantabrica Alavesa	133.176	137.700
Markina-Ondarroa	38.644	44.145
Plentzia-Mungia	90.481	88.446
Tolosaldea/Tolosa	136.602	149.660
Urola-Kostaldea/Urola Costa	179.130	155.847

En millones de pesetas

Hay que tener en cuenta también que para la estimación de los parámetros del modelo se han realizado 5.100 iteraciones, aunque sólo las 100 últimas han sido consideradas para darles un valor.

BUGS no cuenta con una opción clara que permita incluir los pesos muestrales, lo cual ha supuesto un serio escollo en la utilización, de este software en nuestro ejemplo. Finalmente, dado que el programa permite la inclusión de valores "missing" como datos, se ha optado por introducir la totalidad de los establecimientos, para los que la variable "empleo" viene dada por el Directorio Industrial, así como los datos referidos al Territorio Histórico en que están ubicadas y el sector de actividad A84 al que pertenecen. Para la variable a estimar, PBSF, se introducen los datos de la encuesta para los establecimientos muestrales y missing (.) en el resto.

Queda pendiente disipar las dudas que plantea operar de este modo y el estudio de los diagnósticos de convergencia, para los que BUGS dispone de vaias opciones y de la librería CODA de S-PLUS.

En general, la utilización de este software requiere estar bastante familiarizado con las familias de distribuciones y tener un gran conocimiento del tema de la investigación para poder fijar como distribuciones a priori aquellas que resulten más adecuadas.

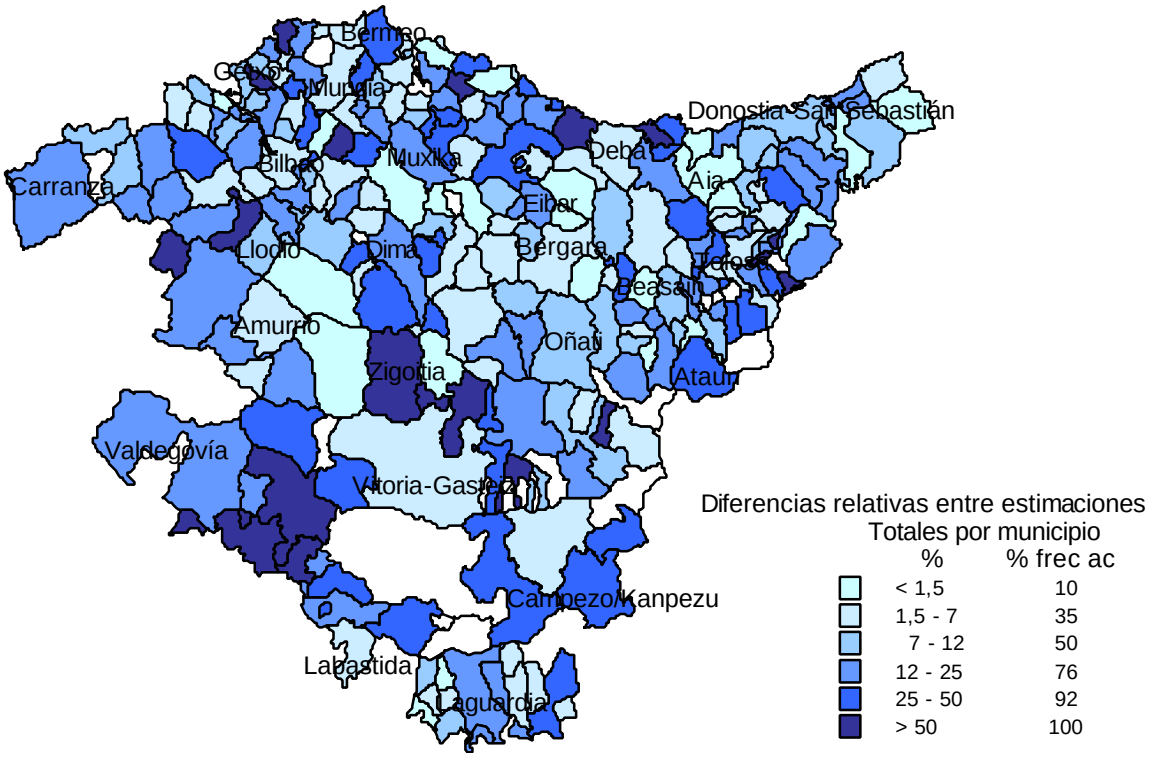
Errores y comparabilidad entre las estimaciones

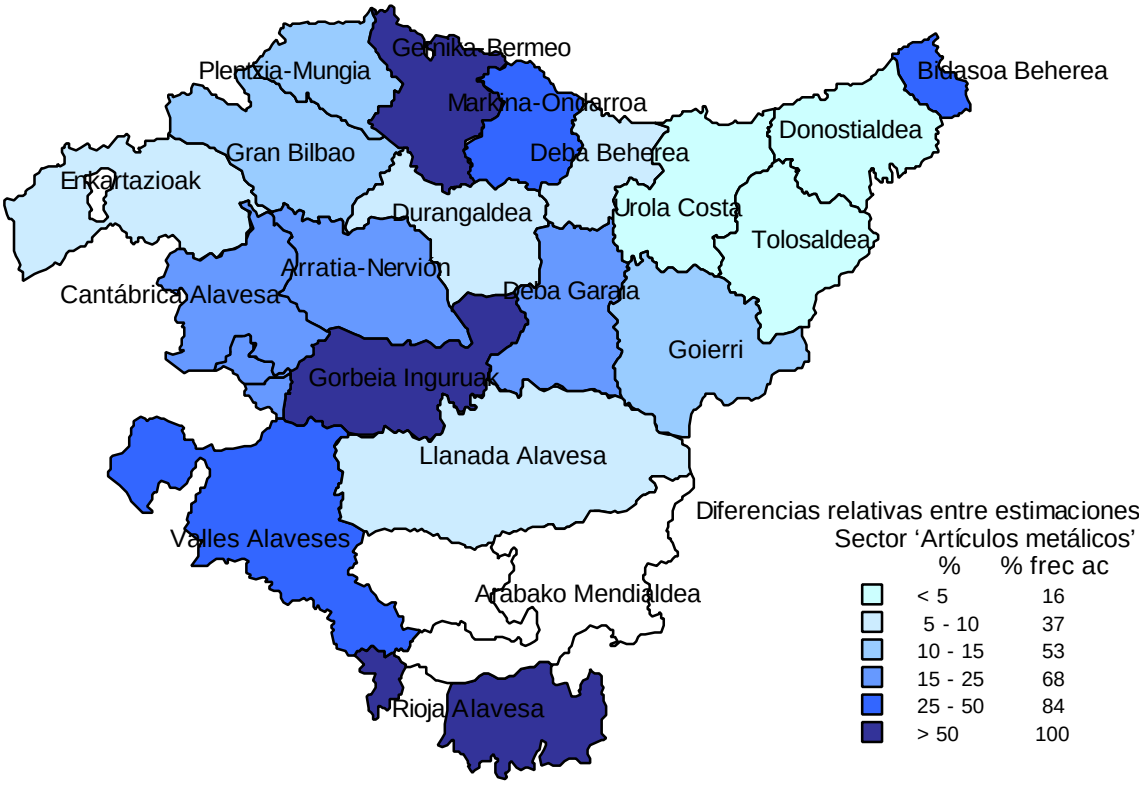
La virtud del empleo de estimadores basados en modelos reside en buena parte en la posibilidad de cuantificar los errores de las estimaciones obtenidas.

El cálculo de estos errores es una cuestión que se escapa al contenido de este Cuaderno y sobre la que habrá que trabajar a partir de ahora

Para comparar las distintas estimaciones hemos tomado como referencia las estimaciones sintéticas obtenidas por los Técnicos del Área de Económicas, calculando el componente de sesgo de las estimaciones obtenidas a partir de los modelos. Las estimaciones sintéticas poseen una menor varianza y, además, dado que el modelo implícitamente adoptado está respaldado por el profundo conocimiento que los Estadísticos encargados de explotar la Encuesta Industrial tienen del tema, es de suponer que las estimaciones así obtenidas sean poco sesgadas.

Los siguientes mapas representan las diferencias entre ambas estimaciones:





Los mapas representados para los dos tipos de estimaciones tratados poseen un coloreado muy parecido. Calculando las diferencias relativas se obtiene que el 75 % de municipios de la Comunidad Autónoma de Euskadi tienen un error menor que el 25% y el 50% menor que el 12%.

Si nos restringimos a la producción de un sector particular, como el de "Artículos metálicos", y considerando las estimaciones por comarcas, en el 68% de las mismas se comete un error menor del 25% y en el 53% menor del 15%.

Capítulo
5

Consideraciones generales

A lo largo de estas páginas se ha atendido al problema de la estimación en áreas pequeñas únicamente desde la perspectiva de la estimación propiamente dicha, sin considerar otra serie de aspectos que también habrían de tenerse en cuenta.

Las consecuencias que pueden derivarse de las medidas adoptadas en base a resultados obtenidos mediante este tipo de técnicas, por parte de las administraciones públicas o del sector privado, hacen que dichas estimaciones se observen con muchas cautelas; cuando además, los errores resultan mucho más apreciables que los que pudieran cometerse en áreas o poblaciones más agregadas. Por estas y otras razones resulta necesario el desarrollo de una estrategia global, que abarque a todas las etapas de la operación estadística y que tenga en consideración el nivel de detalle que se va a requerir en ciertas estimaciones.

La cuestión de los diseños muestrales óptimos para los estimadores directos es algo que ha recibido gran atención en los últimos 50 años, y habría que considerar la influencia que éstos tienen en la estimación en áreas pequeñas. Las encuestas muestrales proporcionan estimaciones fiables para grandes dominios, pero estas encuestas no tratan a la población como un todo, sino que a partir de ellas se obtienen datos para dominios cruzados, como pudieran ser “sexo * rango de edad” o “comarca * CNAE * estrato de empleo”, resultando en algunos casos, cuando éstos se refieren a subpoblaciones *raras* o dominios geográficos pequeños, imposible obtener estimaciones o ser éstas de dudosa calidad.

Para en la medida de lo posible, evitar estos problemas y mejorar la precisión de las estimaciones, hay que tener en cuenta estos dominios desde la etapa del diseño muestral, alcanzando diversos niveles de compromiso según las necesidades de información requeridas.

La planificación de la operación, analizando las necesidades de información en determinados niveles, identificando dichos grupos y teniéndolos en consideración posteriormente en el diseño muestral, resulta de vital importancia para que las estimaciones que obtengamos sean dignas de crédito.

Singh, Gambino y Mantel, en [1], proponen varias ideas para la fase del diseño muestral con el fin de disminuir la influencia de los estimadores indirectos en áreas pequeñas. Entre ellas están la sustitución de conglomerados por “list frames” y usar muchos estratos para controlar mejor el tamaño de la muestra en dominios reducidos, alcanzando compromisos sobre la localización de la misma.

El que haya gran número de conglomerados afecta a aquellas poblaciones que no fueron consideradas en el diseño muestral, pudiendo ocurrir que algunas de ellas tengan gran número de unidades muestrales, mientras que otras apenas tengan alguna. El objetivo de la reducción del número de conglomerados es aminorar los efectos del diseño tanto como lo permitan las restricciones operativas.

En cuanto a la localización de la muestra, en lugar de hacer repartos proporcionales a la población, sería más óptimo reducirla en los dominios mayores, a los que apenas les afectaría esa disminución, en favor de esos otros dominios a los que les corresponde muy poca muestra, insuficiente para representar su diversidad.

La información auxiliar disponible procedente de registros administrativos o censos, resulta de enorme utilidad a la hora de elaborar el diseño muestral, tomar decisiones respecto a la reducción de los conglomerados y sobre la localización de la muestra.

En el artículo antes referido se analiza de forma detallada el modo en que se elabora el diseño muestral, considerando las cuestiones aquí apuntadas, de la Encuesta de Población Activa de Canadá, la mayor encuesta mensual realizada por Statistics Canada.

Hasta ahora se ha tratado de recalcar el hecho de que la estimación en áreas pequeñas no es una cuestión que se resuelva únicamente en la fase de explotación de la muestra recogida, sino que ha de considerarse a lo largo de todas las fases de la operación. Pero en cuanto a la estimación, también habría que hacer una mención especial acerca de la calidad de la misma.

La mayoría de los productores y usuarios estadísticos están acostumbrados a estimadores basados en el diseño y sus correspondientes inferencias. Interpretan los datos en el contexto de repetición de muestras seleccionadas dentro de un marco muestral y emplean el coeficiente de variación como medida de calidad.

En situaciones en que los dominios son muy pequeños y el diseño muestral no ha previsto la obtención de estimaciones para los mismos, los estimadores directos tienen grandes desviaciones y la estimación basada en modelos se convierte en la única opción posible. Valorar, comparar y explicar a los usuarios la precisión relativa de las estimaciones realizadas a diversos niveles de agregación, algunas usando estimadores directos, otras modelos, suponen un reto para los estadísticos. El coeficiente de variación de un estimador basado en modelos puede diferir mucho del de los estimadores directos.

Para los estimadores basados en modelos lo normal es obtener medidas de su error cuadrático medio, esto es, de la varianza y el cuadrado del sesgo del diseño. Pero también nos encontraremos con el problema de que “si la cantidad de datos resulta insuficiente para utilizar estimadores directos, difícilmente éstos van a proporcionar estimaciones adecuadas de la varianza y el sesgo”. Para evitar la dificultosa estimación del sesgo algunos autores proponen estimadores basados en modelos que sean consistentes con el diseño, aunque si el tamaño de la muestra en un dominio es lo suficientemente grande como para que el estimador sea consistente, entonces el propio estimador basado en el diseño proporcionaría estimaciones aceptables.

Se estima conveniente el empleo de la media de los errores mínimo-cuadráticos sobre los dominios como medida de la calidad de los estimadores. Como la necesidad de estimaciones sobre diferentes dominios surge de la suposición de que existen diferencias entre ellos, una cuestión central es explicar cómo estimaciones que se espera que difieran tienen asociada una misma medida acerca de su precisión. Otra posibilidad es realizar estimaciones basadas en modelos de la varianza y del sesgo para cada dominio. Encontrar métodos apropiados de estimación del error medio cuadrático es una cuestión que actualmente está siendo centro de atención de la investigación estadística.

Otra cuestión de gran importancia es la cuestión de la validación de los modelos y la protección contra los fallos de estos. Es necesario avanzar en la validación de los modelos en el caso de encuestas complejas. Hay que realizar una llamada de atención acerca de la utilización de modelos con datos de momentos temporales diferentes, pues la estimación del *cambio* entre distintos periodos de tiempo puede ser de dudosa calidad. Esto mismo puede suceder con la comparación entre áreas pequeñas incluidas en una mayor en la que las relaciona el modelo.

Bibliografía

[1] CARLIN, B.P. & LOUIS, T.A

Bayes and empirical bayes methods for data analysis

Chapman & Hall/CRC.1996.

[2] GHOSH, M. & RAO, J.N.K.

Small area estimation: An appraisal

Statistical science. Vol. 9, nº 1, págs. 55-93. 1994.

[3] IASS

Small area estimation

Publicación de las ponencias de la conferencia satélite celebrada en 1999 en Riga, Letonia.

[4] KOTT, P.S.

Robust small domain estimation using random effects modeling

Survey Methodology. Vol. 15, nº 1, págs. 3-12. Statistics Canada. 1989.

[5] PLATEK, R., RAO, J.N.K. y otros

Small area statistics. An international symposium

John Wiley & sons. Publicación de las ponencias presentadas en el congreso celebrado en 1985 en Ottawa, Canadá. 1987.

[6] PRASAD, N.G.N. y RAO, J.N.K.

The estimation of the mean squared error of small-area estimation

Journal of the American Statistical Association. Vol. 85, págs. 163-171. American Statistical Association. 1990.

[7] PRASAD, N.G.N. y RAO, J.N.K.

On robust small area estimation using a simple random effects model

Survey Methodology. Vol. 25, nº 1, págs. 67-72. Statistics Canada.1999.

[8] RAO, J.NK.

Some recent advances in model-based small area estimation

Survey Methodology. Vol. 25, nº 2, págs. 175-186. Statistics Canada.1999.

[9] RAO, J.N.K.

Metodología estadística para estimaciones indirectas en pequeñas áreas

Seminario Internacional de estadística en Euskadi. EUSTAT. 2000.

[10] SCHAIBLE, W.L.

Indirect estimators in U.S. Federal Programs

Springer. 1996.

[11] SINGH, M.P., GAMBINO, J. & MANTEL, H.J.

Issues and strategies for small area data

Survey Methodogy. Vol. 20, nº 1, págs. 3-22. Statistics Canada. 1994.

[12] SPIEGELHALTER, D. y otros

BUGS 0.5. Bayesian inference using Gibbs sampling

MCR Biostatistics Unit, Institute of Public Health, Cambridge. 1996.

[13] STUKEL, D.M. y RAO, J.N.K.

Small area estimation under two-fold nested error regression models

Journal of Statistical Planning and Inference. Vol. 78, págs. 131-147. Elsevier Science. 1999.