SMALL-AREA ESTIMATION IN THE INDUSTRIAL SURVEY OF THE A.C. OF THE BASQUE

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

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

Presentation:
EUSTAT
Euskal Estatistika Erakundea
Instituto Vasco de Estadística

Publication:
EUSTAT
Euskal Estatistika Erakundea
Instituto Vasco de Estadística
Donostia-San Sebastián, 1 - 01010 VITORIA-GASTEIZ

Print rune:
500 Copies
I-2006

Printing and Binding:
Estudios Gráficos ZURE, S.A.
Carretera Lutxana-Asua, 24-A
Erandio-Goikoa (Bizkaia)

I.S.B.N.: 84-7749-425-8

L..D.: BI-

PRESENTATION

Eustat, aware of the growing demand for increasingly disaggregated quality statistics, formed two years ago a research team made up of members of different departments of Eustat and University researcher. The main aim was to work on the improvement of estimation techniques in various statistical operations and to introduce model based small area estimation techniques in the official statistical production.

This work refers to the first survey worked upon in the search for a definition of a small-area estimation system. This survey has been the yearly Industrial Survey due mainly to its special relevance within Eustat economic surveys and the fact that it constitutes a key sector for the basque economy.

In this report, we present district-level estimations for the Industrial Survey of the A.C. of the Basque Country corresponding to 2002 and 2003. Due to the unfamiliarity and methodological complexity of this kind of techniques, this document offers a detailed description of the estimators and models applied, both for the totals and for the corresponding mean square errors.

The work presented in this document is just the starting point of bigger project. Within this project, small area estimation techniques will be introduced in other surveys of Eustat and useful resources will be provided for users interested in the knowledge and use of small area methods.

Vitoria-Gasteiz, January 2006

JOSU IRADI ARRIETA

General Director

Index

Chapter

1

Industrial Survey of the A.C. of the Basque Country

1.1.- Description of the Industrial Survey of the A.C. of the Basque Country

· Background

This statistical operation got underway in 1981 and from the outset its main objective has been a detailed understanding of the Basque industrial framework, given its importance in terms of value added as well as employment. The basic information is obtained via the main entries in the profit and loss account and the subsequent estimation of the principal macro-magnitudes.

This statistical operation is carried out in collaboration with the specific Statistical Organisation of the Statistics Service and Sectorial Analysis of the Department of Agriculture and Fishing.

· Technical Characteristics

Population: The population scope is limited to those establishments whose main activity, measured in terms of value added generated, is industrial.

This includes, according to the National Classification of Economic Activities 1993 (henceforth NACE-93), the following sections:

Section C: Mining industry;

Section D: Manufacturing industry;

Section E: Production and distribution of electric energy, gas and water.

Geographical. The statistical units located in the geographical sphere of the A.C. of the Basque Country, even when its central office or management is elsewhere.

Temporal. The reference period is the natural financial year. Exceptionally, in the case of establishments whose accountancy refers to time periods that do not correspond with the natural year, the information will refer to the financial period closing within the corresponding years.

Survey framework

The framework of the survey is the Eustat Directory of Economic Activities, which allows to obtain a probabilistic sample with controlled sampling errors.

Statistical Unit

The statistical unit is the establishment defined as a unit that chiefly or exclusively carries out one or various activities situated in the same geographical location.

Sample design and extrapolation

A probabilistic sample is made in two phases: a first one where all the units with more than 19 employees are selected with probability "one"; in the second phase a random stratified sample is obtained where the stratification variables are:

a) Province: Álava, Bizkaia and Gipuzkoa.

b) Activity: National Classification of Economic Activities (NACE-93) at sub-class level, which is to say, to 5 digits. For its subsequent publication the normalised classification of EUSTAT A84 is used. The A84 classification is a disaggregation of the A60 (NACE-93 to 2 digits) in relation to the specific economic structure of the A.C. of the Basque Country.

The sample size selected is approximately 3,000 statistical units.

Prior to extrapolation the sample establishments are post-stratified, according to the three Provinces (Álava, Bizkaia, Gipuzkoa), sub-class of the CNAE-93 to 5 digits and five sizes of establishment, which are:

Between 1 and 19 employees;

Between 20 and 49 employees;

Between 50 and 99 employees;

Between 100 and 499 employees;

500 or more employees.

The step from sampling data to estimations is made via a weight matrix for each stratum. The variable used to obtain the exponents is the number of people employed in the industrial establishments. This variable is used because it is the one most correlated with the main economic variables that this survey sets out to measure.

1.2.- Estimators employed in the Industrial Survey of the A.C. of the Basque Country

All the statistical units with more than 19 employees are self-weighted, so that the main interest lies in the estimations within the employment stratum of 1-19 employees. There follows an outline of the estimation within an activity sub-class (NACE-93 to 5 digits) the total of any variable y, as well as its corresponding variation coefficient.

At present, the indirect ratio estimator or synthetic estimator is employed for the estimation of the total of the variable y, using the number of employees in the establishment as auxiliary information. This decision was taken after carrying out an exhaustive descriptive study that proved that there was a strong positive correlation between the number of employees in establishments and the size of the main variables of the Industrial Survey.

The indirect ratio estimator of any variable of interest y, when an auxiliary variable x is available, is, in the case of the Industrial Survey, assisted by the simple heteroscedastic linear regression model of the type:

$$y_{hj} = x_{hj}\beta + \varepsilon_{hj} \quad \text{with} \quad \text{var}(\varepsilon_{hj}) = \sigma^2 x_{hj}\,, \tag{1}$$

where h refers to the stratum and j to the statistical unit. The strata are the Provinces (Álava, Bizkaia and Gipuzkoa), and in all that follows the interest lies in a single sub-class. The simple linear model (1) contains heteroscedastic disturbances where the variance is a growing linear function of the number of employees of the industrial establishments.

The estimator of the total of variable y in a given sub-class in the Province h is given by:

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

where $X_h = \sum_{j=1}^{N_h} x_{hj}$, w_{hj} is the sampling weight of unit j in Province h, x_{hj} covers employment in establishment j of Province h and n_h is the size of the sample in Province h.

The variance estimator of the indirect ratio estimator can be approximated via:

$$\hat{\text{var}}(\hat{t}_{yh.SYN}) \approx N^2\left(1-\frac{n}{N}\right)\frac{1}{n}\left(\frac{X_h}{\sum_{h=1}^{H}\sum_{j=1}^{n_h} w_{hj}x_{hj}}\right)^2 \hat{\text{var}}(\varepsilon),$$

where $\hat{\text{var}}(\varepsilon)$ is the sampling variance of the residuals of heteroscedastic model (1) with all the sampling data (which is to say that the residuals of the whole of the A.C. of the Basque Country are calculated, not just in Province h) and the rest of the notation is as usual.

Särndal and Hidiroglou (1989) offer an approximation of the bias of the synthetic estimator whereby $E\left(\hat{t}_{yh.SYN}\right) - t_{yh.SYN} \approx -\sum_{j=1}^{N}\hat{\varepsilon}_j$ where $\hat{\varepsilon}_j = y_j - x_j\hat{\beta}$.

Therefore, the estimator will be approximately unbiased if it is verified that $\sum_{j=1}^{N}\hat{\varepsilon}_k = 0$. This condition is not normally satisfied. If the model does not fit within the domain of interest, the sum of the residuals may be far from zero, indicating a considerable bias. Otherwise, we would expect a limited bias. Therefore, it is advisable to estimate the mean square error as a measurement of the accuracy of the estimator. This is given by:

$$MSE\left(\hat{t}_{yh.SYN}\right) = \text{var}\left(\hat{t}_{yh.SYN}\right) + \left(bias_{h.SYN}\right)^2,$$

and is estimated via the expression:

$$\hat{MSE}\left(\hat{t}_{yh.SYN}\right) = \hat{\text{var}}\left(\hat{t}_{yh.SYN}\right) + \left(\sum_{j=1}^{n_h}\hat{\varepsilon}_j\right)^2,$$

where $\hat{\varepsilon}_j = y_j - x_j^{|}\hat{\beta}$, for $j = 1,...,n$ are the residuals obtained from the estimated model (1) with all the sampling data, although in each Province solely the specific data of this Province is added. The variation coefficient is defined as

$$\hat{cv}\left(\hat{t}_{yh.SYN}\right) = \frac{\hat{rmse}\left(\hat{t}_{yh.SYN}\right)}{\hat{t}_{yh.SYN}},$$

where $\hat{rmse}\left(\hat{t}_{yh.SYN}\right) = \sqrt{\hat{MSE}\left(\hat{t}_{yh.SYN}\right)}$.

Chapter

2

System of small-area estimation in the Industrial Survey

2.1.- Introduction

The A.C. of the Basque Country is divided into the following 20 districts:

- Álava: Valles Alaveses, Llanada Alavesa, Montaña Alavesa, Rioja Alavesa, Estribaciones del Gorbea y Cantábrica Alavesa.
- Bizkaia: Arratia-Nervión, Gran Bilbao, Duranguesado, Encartaciones, Gernika-Bermeo, Markina-Ondarroa y Plentzia-Mungia.
- Gipuzkoa: Bajo Bidasoa, Bajo Deba, Alto Deba, Donostia-San Sebastián, Goierri, Tolosa y Urola Costa.

Industrial activity of the A.C. of the Basque Country is not evenly distributed among the 20 statistical districts and both the importance of the industrial sector as well as its size vary hugely from district to district. In fact, there are districts where industrial activity is extremely limited, meaning that the task of district-level estimation has to involve model based small area estimation techniques. The increase of sampling size required in order to obtain quality district-level estimations would certainly be costly.

The small-area models assume the existence of an underlying model that all the population data follows, but which is estimated with the sampling data (Rao, 2003). In order to obtain district-level estimations for the Industrial Survey, Eustat employs two types of models: the fixed-effect linear regression model and the linear regression model with fixed and random effects, also called the mixed model.

In the mixed model the predictor includes a common fixed-effect term for all districts and another term differentiating the elements of each district d ($d = 1,...,t$). The differentiating term is made up of random effects (v_d), so that all the data from the same district shares the same random effect. In the case of the fixed-effect model, there are no differentiating terms for each district since the systematic part is common to all the districts. However, specificity is achieved by projecting the common estimated coefficient onto the specific auxiliary information of each district.

2.2.- Mixed linear model

2.2.1.- Projective version

Its basis is a population made up of N establishments of a given activity sub-class (NACE-93 to 5 digits). In each district d ($d = 1,...,t$) there are N_d establishments in the population, thus $N = \sum_d N_d$. In this sub-class n establishments are sampled of which n_d belong to district d. The following mixed linear heteroscedastic model is proposed:

$$y_{dj} = \beta_0 + \beta_1 x_{dj} + v_d + e_{dj}, \qquad d = 1,...,t \qquad j = 1,...,n_d, \qquad (2)$$

where, for establishment j of district d, y_{dj} is the value taken by the variable of interest and x_{dj} is the number of employees in the establishment. The total number of sampled establishments in district d is n_d. The fixed effects of the model are β_0 and β_1. The common random effect for all the establishments of district d is v_d and e_{dj} are the specific random errors of each establishment. Furthermore, it is assumed that $v_d \subset N\left(0, \sigma_v^2\right)$ and $e_{dj} \subset N\left(0, \sigma_e^2 c_{dj}^{-1}\right)$ are independent. To correct the

heteroscedasticity in the data, weights $c_{dj} = 1/x_{dj}$ are used. When $c_{dj} = 1 \ \forall d, j$, this model is similar to that proposed by Battese *et al* (1988).

The superpopulation model corresponding to model (2) written in matrix form is given by:

$$Y = X\beta + Zv + \varepsilon, \qquad v \subset N(0, \sigma_v^2 I_t), \qquad \varepsilon \subset N(0, \sigma_e^2 C^{-1}), \qquad (3)$$

where $C = diag(c_{dj})$ ($d = 1,...,t$), is the weight matrix of the model, and j is the establishment ($j = 1,..., N_d$). $Y = (Y_1^|,...,Y_t^|)^|$ is the ($N \times 1$) vector whose components $Y_d^|$ are the values of the variable of interest for each district, $\beta = (\beta_0, \beta_1)^|$ is the vector of coefficients of the model, X is the (**¡Error! Marcador no definido.**) design matrix formed by a column of ones associated with the intercept and another associated with the auxiliary variable which is, in this case, the number of employees in each establishment. Matrix $Z = diag(1_{N_d})$, $d = 1,...,t$, is the ($N \times t$) design matrix diagonal by blocks associated with the random effects. This is to say for each district d, matrix Z has an associated column of ones defined by vector $1_{N_d} = (1,...,1)^|$ of dimension N_d. Random effects $v = (v_1,...,v_t)^|$, are common to the N_d elements of the same district and $\varepsilon = (\varepsilon_1^|,...,\varepsilon_t^|)^|$ is the random error vector, where $\varepsilon_d = (\varepsilon_{d_1},...,\varepsilon_{d_{N_d}})^|$.

With the objective of unifying the theory and presenting the projective and predictive versions of model (3), we will go on to differentiate the sampled and non-sampled parts as follows:

$$\begin{pmatrix} Y_s \\ Y_r \end{pmatrix} = \begin{pmatrix} X_s \\ X_r \end{pmatrix}\beta + \begin{pmatrix} Z_s \\ Z_r \end{pmatrix}v + \begin{pmatrix} \varepsilon_s \\ \varepsilon_r \end{pmatrix},$$

where the sub-indexes s and r denote the sampled and non-sampled establishments respectively. Thus the sampling model can be written as:

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

where $C_s = diag(c_{dj} = 1/x_{dj})$, $d = 1,...,t_s$, $j = 1,...,n_d$ and t_s is the total number of sampled districts. The variance and covariance matrix of Y_s can be expressed as

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

where $V_d = \sigma_e^2 C_d^{-1} + \sigma_v^2 1_{n_d} 1_{n_d}^|$ and $C_d = diag(c_{d_1},...,c_{d_{n_d}})_{n_d \times n_d} = diag(c_{n_d})$. If we assume that the components of variance $\sigma^2 = (\sigma_e^2, \sigma_v^2)$ are known, the fixed-effect estimator as well as its variance and covariance matrix can be obtained by generalised least squares:

$$\tilde{\beta} = (X_s^{|} V_s^{-1} X_s)^{-1} X_s^{|} V_s^{-1} Y_s, \qquad \text{var}(\tilde{\beta}) = \Phi_s = (X_s^{|} V_s^{-1} X_s)^{-1},$$

where $V_s^{-1} = diag(V_1^{-1},...,V_d^{-1},...,V_{t_s}^{-1})$, $V_d^{-1} = \frac{1}{\sigma_e^2}\left(C_d - \frac{\gamma_{dc}}{c_{d.}} c_{n_d} c_{n_d}^{|}\right)$,

$\gamma_{dc} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2 / c_{d.}}$ and $c_{d.} = \sum_{j=1}^{n_d} c_{dj}$.

If $1_{n_d}^{|} = (1,...,1)$ is of dimension n_d, then the prediction of random effects is obtained as $\hat{v}_d = \sigma_v^2 1_{n_d}^{|} V_d^{-1}(Y_d - X_d^{|}\hat{\beta}) = \hat{\gamma}_{dc}(\bar{y}_{dc} - \bar{x}_{dc}^{|}\hat{\beta})$, where $\bar{y}_{dc} = \frac{1}{c_{d.}}\sum_{j=1}^{n_d} c_{dj} y_{dj}$, and $\bar{x}_{dc}^{|} = \frac{1}{c_{d.}}\sum_{j=1}^{n_d} c_{dj} x_{dj}^{|} = (1, \bar{x}_{dc})$, with $x_{dj}^{|} = (1, x_{dj})$. The projective predictor of the mean of the variable of interest in district d is given by:

$$\hat{\bar{y}}_d^* = \bar{X}_{d(p)}^{|}\hat{\beta} + \hat{\gamma}_{dc}(\bar{y}_{dc} - \bar{x}_{dc}^{|}\hat{\beta}), \qquad d = 1,...,t, \tag{4}$$

where $\hat{\beta} = \tilde{\beta}(\hat{\sigma}_e^2, \hat{\sigma}_v^2)$ is evaluated with the estimations of the variance components, $\bar{X}_{d(p)}^{|} = (1, \bar{x}_{d(p)})$, and $\bar{x}_{d(p)} = \frac{\sum_{j \in N_d} x_{dj}}{N_d}$ is the population mean of the number of employees in district d for all the establishments of the population: the sampled and the non-sampled in district d. The corresponding version for the total is given by:

$$\hat{t}_d^* = X_{d(p)}^{|}\hat{\beta} + N_d \hat{\gamma}_{dc}(\bar{y}_{dc} - \bar{x}_{dc}^{|}\hat{\beta}), \qquad d = 1,...,t, \tag{5}$$

where $X_{d(p)}^{|} = (N_d, X_{d(p)})$, N_d is the total number of establishments and $X_{d(p)}$ is the total number (population) of employees of district d. The predictor (4) can also be expressed as a weighted sum of a generalised regression estimator:

$$\bar{y}_{dc} + (\bar{X}_{d(p)} - \bar{x}_{dc})^{|}\hat{\beta},$$

and the synthetic regression estimator $\bar{X}_{d(p)}^{|}\hat{\beta}$, so that,

$$\hat{\bar{y}}_d = \gamma_{dc}(\bar{y}_{dc} + (\bar{X}_{d(p)}^{|} - \bar{x}_{dc})^{|}\hat{\beta} + (1 - \gamma_{dc})\bar{X}_{d(p)}^{|}\hat{\beta}, \qquad d = 1,...,t. \tag{6}$$

The weight $0 \le \gamma_{dc} \le 1$ measures the proportion of variance σ_v^2 relative to the total variance $\sigma_v^2 + \sigma_e^2 / c_d$. If the variance of the model is small, the γ_{dc} are small and more weight is given to the synthetic component. Analogously, more weight is given to the generalised regression estimator when $c_{d.}$ is larger. When $c_{dj} = 1$ the generalised regression estimator is approximately unbiased under the design if n_d is large enough. Generally, it is unbiased subject to the fulfilment of the effects v_d assuming that $\tilde{\beta}$ is conditionally unbiased for β. This is to say the BLUP (Best Linear Unbiased Predictor) estimator (4) is conditionally biased due to the presence of this synthetic component When the sample is a simple random one and $c_{dj} = 1$ the BLUP estimator is consistent under the design for the district mean $\overline{Y}_d$ when n_d grows, since $\gamma_d \to 1$.

When the model is homoscedastic and $c_{dj} = 1$ therefore $c_{d.} = n_d$, then the projective predictor of the mean is given by:

$$\hat{\overline{y}}_d^* = \overline{X}_{d(p)}^{|}\hat{\beta} + \hat{\gamma}_d(\overline{y}_{dc} - \overline{x}_{dc}^{|}\hat{\beta}) = \hat{\gamma}_d \overline{y}_{dc} + (\overline{X}_{d(p)}^{|} - \hat{\gamma}_d \overline{x}_{dc}^{|})\hat{\beta},$$

where $\hat{\gamma}_d$ measures the uncertainty associated with the modelisation of the predictor and takes the form:

$$\hat{\gamma}_d = \frac{\hat{\text{cov}}(v_d, \overline{u}_{d.})}{\hat{\text{var}}(\overline{u}_{d.})} = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_d}.$$

2.2.2.- Predictive version

When the sampling fraction by district $f_d = n_d / N_d$ is significant, the literature recommends the use of the predictive version to obtain the prediction of the total or the mean of district d instead of the projective version. This version consists of differentiating the sampled part from the non-sampled part. In this way, the prediction of the sampled part is the sample itself, while the non-sampled part is predicted with the projective predictor.

To obtain the predictive version, the total $\sum_{j \in N_d} y_{dj} = \sum_{j \in d_r} y_{dj} + \sum_{j \in d_s} y_{dj}$ is decomposed, where d_s indicates the sample in district d and d_r the rest of the establishments which are not part of the sample in district d. Below the population mean is decomposed:

$$\overline{Y}_d = \frac{\sum_{j \in d_r} y_{dj} + \sum_{j \in d_s} y_{dj}}{N_d} = \frac{(N_d - n_d)\overline{Y}_{dr} + n_d \overline{y}_{ds}}{N_d} = (1 - f_d)\overline{Y}_{dr} + f_d \overline{y}_{ds}. \quad (7)$$

The predictive estimator of the mean for area d, for all $d = 1,...,t$ is given by:

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

Substituting $\hat{\bar{y}}^*_{dr}$ for its expression $\bar{X}^{|}_{dr}\hat{\beta} + \hat{v}_d$ where $\bar{X}^{|}_{dr} = (1, \bar{x}_{dr})$ and $\bar{x}_{dr} = \dfrac{\sum_{j \in d_r} x_{dj}}{N_d - n_d}$, it follows that:

$$\hat{\bar{y}}_d = \hat{\bar{Y}}_d = (1 - f_d)\left[\bar{X}^{|}_{d(p_r)}\hat{\beta} + \hat{\gamma}_{dc}(\bar{y}_{dc} - \bar{x}^{|}_{dc}\hat{\beta})\right] + f_d \bar{y}_{ds}, \qquad (8)$$

which leads to the predictive version of the total:

$$\hat{t}_d = X^{|}_{d(p_r)}\hat{\beta} + (N_d - n_d)\hat{\gamma}_{dc}\left(\bar{y}_{dc} - \bar{x}_{dc}\hat{\beta}\right) + \sum_{j=1}^{n_d} y_{dj}, \qquad d = 1,...,t \quad (9)$$

where $\hat{\beta} = \hat{\beta}_c\left(\hat{\sigma}^2_e, \hat{\sigma}^2_v\right)$ has been evaluated with the estimations of the variance components and $X_{d(p_r)}$ is the total number of employees in district d for all the non-sampled establishments.

2.2.2.1.- Estimators of the mean and the total by Province and for the A.C. of the Basque Country

For both predictive and projective versions, the means and totals per Province and for the A.C. of the Basque Country are calculated as follows.

The estimator of the mean for each Province is given by:

$$\hat{\bar{y}}_h = \frac{1}{N_h}\sum_{d \in h} N_d \hat{\bar{y}}_d = \frac{1}{N_h}\sum_{d \in h} \hat{t}_d, \qquad (10)$$

where $d \in h$ indicates that the sum is made in all the districts of stratum h (in this case $h = 1,2,3$ are the Provinces) and $N_h = \sum_{d \in h} N_d$ is the population total of Province h.

The estimator of the total for each Province is given by:

$$\hat{t}_h = \sum_{d \in h} \hat{t}_d. \qquad (11)$$

The estimator of the mean for the A.C. of the Basque Country is given by:

$$\hat{\bar{y}} = \frac{1}{N}\sum_{h=1}^{3} N_h \hat{\bar{y}}_h \,, \qquad (12)$$

where $N = \sum_{h=1}^{3} N_h$.

The estimator for the total for the A.C. of the Basque Country is given by:

$$\hat{t} = \sum_{h=1}^{3} N_h \hat{\bar{y}}_h = \sum_{h=1}^{3} \hat{t}_h \,,. \qquad (13)$$

2.2.3.- Weighted estimation of variance components

Although there are several methods for estimating variance components, this paper puts forward estimation by the method of moments, which does not depend on the hypothesis of normality required by other methods such as maximum likelihood (ML) or restricted maximum likelihood (REML).

2.2.3.1.- Method of moments

The method of moments (Searle et al. 1992) can be applied to estimate σ_v^2 and σ_e^2 equalling the expectations of the sums of the squares with the estimators, so that:

$$\hat{\sigma}_e^2 = \frac{\varepsilon_s^{|} C_\varepsilon \varepsilon_s}{n - rank(X_s, Z_s)} = \frac{\varepsilon_s^{|} C_\varepsilon \varepsilon_s}{n - t_s - 2} \,, \qquad (14)$$

where $n = \sum_{d=1}^{t} n_d$ and $rank(X_s, Z_s)$ is the range of the extended matrix (X_s, Z_s). Therefore the resulting estimation could also be obtained by calculating the residual variance of the weighted regression between the sample data (dependent variable) and the auxiliary data (employment) and the random effects of the district as predictors (independent variables). The variance of the random effects is calculated via the expression:

$$\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0) = \max\left(\frac{1}{n_{*c}} \left\{ \sum_{d=1}^{t_s} \sum_{j=1}^{n_d} c_{dj} s_{dj}^2 - (n - k - 1)\hat{\sigma}_e^2 \right\}, 0 \right), \qquad (15)$$

where:

$$n_{*c} = tr(M_c Z_s Z_s^{|}),$$

$$M_c = C_s - C_s X_s (X_x^{|} C_s X_s)^{-1} X_s^{|} C_s,$$

and, $s_{dj} = y_{dj} - x_{dj}\hat{\beta}_0 = y_{dj} - x_{dj}(X_s^{|} C_s X_s)^{-1} X_s^{|} C_s Y_s$, are the residuals of the weighted regression of Y_s over X_s with weights $C_s = diag(c_{dj} = 1/x_{dj})$, $d = 1,...,t_s$ and $j = 1,...,n_d$. The truncated estimator $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$ is biased but consistent when t_s grows.

2.2.4.- Mean square error by district

Kackar and Harville (1984) demonstrated that under the hypothesis of normality, the mean square error (MSE) of the generic BLUP $t(\hat{\theta}, Y)$ is given by:

$$MSE\left[t(\hat{\theta}, Y)\right] = MSE\left[t(\theta, Y)\right] + E\left[t(\hat{\theta}, Y) - t(\theta, Y)\right]^2, \qquad (16)$$

where $\theta = \sigma^2 = \left(\sigma_v^2, \sigma_e^2\right)$ and assuming that $\hat{\theta} = \hat{\sigma}^2 = \left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ is invariant to translations. For the predictor of the mean $t(\hat{\theta}, Y) = \hat{\bar{y}}_d$, $\tilde{\bar{y}}_d$ is the predictor of $\bar{y}$ assuming that the variance components are known, and therefore, $t(\theta, Y) = \tilde{\bar{y}}_d$. Thus the MSE of the predictor of the mean is given by:

$$MSE\left[\hat{\bar{y}}_d\right] = E\left[\bar{y}_d - \tilde{\bar{y}}_d\right]^2 + E\left[\hat{\bar{y}}_d - \tilde{\bar{y}}_d\right]^2. \qquad (17)$$

Henderson (1975) gave an expression for $MSE\left[\tilde{\bar{y}}_d\right] = g_{1d}(\sigma^2) + g_{2d}(\sigma^2)$, but the second term of (17), called $g_{3d}(\sigma^2)$ is not easily calculated except in special cases. Kackar and Harville (1984) obtained an expression based on the Taylor series expansion:

$$E\left[\hat{\bar{y}}_d - \tilde{\bar{y}}_d\right]^2 \approx E\left[h_d(\theta)(\hat{\theta} - \theta)\right]^2,$$

with $h_d(\theta) = \dfrac{\partial t_d(\theta)}{\partial \theta}$. Prasad and Rao (1990) proposed a subsequent approximation given by:

$$tr\left\{A_d(\hat{\theta}) E\left[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^{|}\right]\right\} \approx tr\left\{(\vee b_d^{|}) V_d (\vee b_d^{|})^{|} E\left[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^{|}\right]\right\},$$

where $\vee b_d^{|} = col_{1\le d\le p}\dfrac{\partial b_d^{|}}{\partial \theta_j}$ and p is the number of variance components. The estimators of $g_{2d}(\sigma^2)$ and $g_{3d}(\sigma^2)$ are given by $g_{2d}(\hat{\sigma}^2)$ and $g_{3d}(\hat{\sigma}^2)$. These estimators are correct up to order $O_p(t^{-1})$ (here t is the number of districts and not the predictor), since $\hat{\sigma}^2$ is a consistent estimator of σ^2. However, $g_{1d}(\hat{\sigma}^2)$ is not a correct estimator of $g_{1d}(\sigma^2)$, since its bias is of the order $O(t^{-1})$, and is obtained using a Taylor series expansion of $g_{1d}(\sigma^2)$ around σ^2 and taking its expectation. After performing several operations, we obtain:

$$E\left[g_{1d}(\hat{\sigma}^2)\right] - g_{1d}(\sigma^2) = -g_{3d}(\hat{\sigma}^2) + O(t^{-1}).$$

2.2.4.1. Projective version

Prasad and Rao (1990) offer an estimator of (17) in the projective version, which is valid when the estimators of the variance components have been obtained by REML or by the method of moments. It is given by:

$$\hat{MSE}\left[\hat{\bar{y}}_{d(p)}\right] = g_{1d}\left(\hat{\sigma}^2\right) + g_{2d}\left(\hat{\sigma}^2\right) + 2g_{3d}\left(\hat{\sigma}^2\right),$$

where:

$$g_{1d}\left(\hat{\sigma}^2\right) = \left(1-\hat{\gamma}_{dc}\right)\hat{\sigma}_v^2,$$

$$g_{2d}\left(\hat{\sigma}^2\right) = \left(\overline{X}_{d(p)} - \hat{\gamma}_{dc}\overline{x}_{dc}\right)'\hat{\Phi}_c\left(\overline{X}_{d(p)} - \hat{\gamma}_{dc}\overline{x}_{dc}\right),$$

$$g_{3d}\left(\hat{\sigma}^2\right) = \hat{\gamma}_{dc}\left(1-\hat{\gamma}_{dc}\right)^2\hat{\sigma}_e^{-4}\hat{\sigma}_v^{-2}h\left(\hat{\sigma}^2\right),$$

$$h\left(\hat{\sigma}^2\right) = \hat{\sigma}_e^4\,\hat{\text{var}}\left(\hat{\sigma}_v^2\right) + \hat{\sigma}_v^4\,\hat{\text{var}}\left(\hat{\sigma}_e^2\right) - 2\hat{\sigma}_v^2\hat{\sigma}_e^2\,\hat{\text{cov}}\left(\hat{\sigma}_e^2\hat{\sigma}_v^2\right),$$

$$\hat{\text{var}}\left(\hat{\sigma}_e^2\right) = 2\left(n-t_s-k\right)^{-1}\hat{\sigma}_e^4 = 2d_e^{-1}\hat{\sigma}_e^4,$$

$$\hat{\text{var}}\left(\hat{\sigma}_v^2\right) = 2n_{*c}^{-2}\left[\left(n-t_s-k\right)^{-1}\left(t_s-1\right)\left(n-k-1\right)\hat{\sigma}_e^4 + 2n_*\hat{\sigma}_e^2\hat{\sigma}_v^2 + n_{**c}\hat{\sigma}_v^4\right],$$

$$\hat{\text{cov}}\left(\hat{\sigma}_e^2,\hat{\sigma}_v^2\right) = -\left(t_s-1\right)n_{*c}^{-1}\,\hat{\text{var}}\left(\hat{\sigma}_e^2\right),$$

$$n_{*c} = tr\left(M_c Z_s Z_s^{|}\right),$$

$$n_{**c} = tr\left(M_c Z_s Z_s^{|}\right)^2,$$

$$M_c = C_s\left(I - X_s\left(X_s^{|}C_s X_s\right)^{-1} X_s^{|}C_s\right),$$

and k is the number of auxiliary variables, in this case $k = 1$.

Frequently the root mean square error (RMSE) is used as a measurement of the accuracy of estimator $\hat{\bar{y}}_d$, given by:

$$RMSE\left[\hat{\bar{y}}_d\right] = \sqrt{MSE\left[\hat{\bar{y}}_d\right]}.$$

The MSE of the predictor of the total for each district is estimated by multiplying the estimator of the MSE of the mean by the square of the population size of the district, N_d^2. In effect,

$$\hat{MSE}\left[\hat{t}_d\right] = N_d^2\left[g_{1d}\left(\hat{\sigma}^2\right) + g_{2d}\left(\hat{\sigma}^2\right) + 2g_{3d}\left(\hat{\sigma}^2\right)\right],$$

and the square root of the mean square error of the total is estimated via the expression:

$$\hat{RMSE}\left[\hat{t}_d\right] = \sqrt{MSE\left[\hat{t}_d\right]}.$$

The variation coefficient is defined as:

$$CV\left[\hat{t}_d\right] = \frac{\hat{RMSE}\left[\hat{t}_d\right]}{\hat{t}_d}.$$

2.2.4.2.- Predictive version

In the predictive version, the estimator of the mean square error of the predictor (8), valid when the estimators of the variance components are obtained by the REML or by the method of moments, is given by:

$$\hat{MSE}\left[\hat{\bar{y}}_d\right] = g_{1d}\left(\hat{\sigma}^2\right) + g_{2d}\left(\hat{\sigma}^2\right) + 2g_{3d}\left(\hat{\sigma}^2\right) + g_{4d}\left(\hat{\sigma}^2\right),$$

where:

$$g_{1d}\left(\hat{\sigma}^2\right) = \left(1 - f_d\right)^2\left(1 - \hat{\gamma}_{dc}\right)\hat{\sigma}_v^2,$$

$$g_{2d}\left(\hat{\sigma}^2\right) = \left(1 - f_d\right)^2\left[\left(\overline{X}_{d(p_r)} - \hat{\gamma}_{dc}\overline{x}_{dc}\right)^{'}\hat{\Phi}_s\left(\overline{X}_{d(p_r)} - \hat{\gamma}_{dc}\overline{x}_{dc}\right)\right],$$

$$g_{3d}\left(\hat{\sigma}^2\right)=\left(1-f_d\right)^2 c_d^{-1}\left(\hat{\sigma}_v^2+\hat{\sigma}_e^2/c_{d.}\right)^{-3}\left[\hat{\sigma}_e^4\,\hat{\mathrm{var}}\left(\hat{\sigma}_v^2\right)+\hat{\sigma}_v^4\,\hat{\mathrm{var}}\left(\hat{\sigma}_e^2\right)-2\hat{\sigma}_e^2\hat{\sigma}_v^2\,\hat{\mathrm{cov}}\left(\hat{\sigma}_e^2,\hat{\sigma}_v^2\right)\right]$$
$$g_{4d}\left(\hat{\sigma}^2\right)=\sigma_e^2 N_d^{-2}\sum_{j\in p_r} c_{dj}^{-1},$$

are the contributions to the MSE of the estimation of the random effects, the fixed effects, the variance components and the weights of the model. Furthermore, p_r represents domain d of census establishments not belonging to the sample and $c_{dj}=1/x_{dj}$, where x_{dj} is the census employment of the non-sampled establishments. If aggregations are made in order to achieve a minimum size before proceeding with the estimations, then we consider the population of the sub-class in which the projection is made as non-sampled population and not that of the aggregation which is usually higher.

The MSE of the predictor of the total for each district is estimated by multiplying the estimator of the MSE of the mean by the square of the population size of the district, N_d^2. In effect,

$$\hat{MSE}\left[\hat{t}_d\right]=N_d^2\left[g_{1d}\left(\hat{\sigma}^2\right)+g_{2d}\left(\hat{\sigma}^2\right)+2g_{3d}\left(\hat{\sigma}^2\right)\right], \qquad (18)$$

and the square root of the mean square error of the total is estimated via the expression:

$$\hat{RMSE}\left[\hat{t}_d\right]=\sqrt{MSE\left[\hat{t}_d\right]}. \qquad (19)$$

The variation coefficient is defined as:

$$CV\left[\hat{t}_d\right]=\frac{\hat{RMSE}\left[\hat{t}_d\right]}{\hat{t}_d}.$$

2.3.- Fixed-effect linear model

The superpopulation fixed-effect model is given by:

$$Y=X\beta+\varepsilon, \qquad \varepsilon\subset N\left(0,\sigma_e^2 C^{-1}\right), \qquad (20)$$

where $C=diag(c_{dj})$, is the matrix of the model weights, d represents the district ($d=1,...,t$) and j is the establishment ($j=1,...,N_d$). Vector $Y=\left(Y_1^{|},...,Y_t^{|}\right)$ is the

$(N \times 1)$ vector whose components $Y_d^{|}$ are the observed values of the variable of interest for each district d , β is the single fixed coefficient of the model, X is the $(N \times 1)$ column vector of the auxiliary variable, which is to say employment, and $\varepsilon^{|} = \left(\varepsilon_1^{|},...,\varepsilon_t^{|}\right)$ where $\varepsilon_d^{|} = \left(\varepsilon_{d_1},...,\varepsilon_{d_{N_d}}\right)$ is the random effects vector.

As was the case with the decomposition carried out in the mixed model, the sampled and non-sampled parts can be separated as follows:

$$\begin{pmatrix} Y_s \\ Y_r \end{pmatrix} = \begin{pmatrix} X_s \\ X_r \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_s \\ \varepsilon_r \end{pmatrix}$$

where the sub-indices s and r denote the sampled and non-sampled establishments respectively. Thus the fixed-effect sampling model can be written as:

$$Y_s = X_s \beta + \varepsilon_s \qquad \varepsilon_s \subset N(0_s, \sigma_e^2 C_s^{-1}) , \tag{21}$$

where $C_s = diag\left(c_{dj}\right)$, $d = 1,...,t_s$, $j = 1,...,n_d$ and t_s is the total number of sampled districts. In an expanded form, model (21) is expressed as:

$$y_{dj} = \beta x_{dj} + e_{dj} \qquad d = 1,...,t \ , \ j = 1,...,n_d \tag{22}$$

where for establishment j of district d , y_{dj} is the value taken by the variable of interest and x_{dj} is the number of employees of the establishment. The total number of sampled establishments in district d is n_d , β is the single fixed effect of the model and $e_{dj} \subset N\left(0, \sigma_e^2 c_{dj}^{-1}\right)$ are the random errors. Furthermore, $e_{dj} \subset N\left(0, \sigma_e^2 c_{dj}^{-1}\right)$. To correct the heteroscedasity present in the data, we use the weights $c_{dj} = 1/x_{dj}$.

If N_d is the total number of units of area d , the population mean of area d is given by

$$\overline{Y}_d = \frac{1}{N_d} \sum_{d=1}^{N_d} y_{dj} = f_d \overline{y}_{ds} + (1 - f_d) \overline{y}_{dr} , \tag{23}$$

where $f_d = n_d / N_d$, $\overline{y}_{ds}$ is the sample mean of the sampled units and $\overline{y}_{dr}$ is the sample mean of the non-sampled ones. Given that the second term of (23) has not been observed, it is substituted by its estimated value. An estimator of (23) obtained in a similar way to that given in (7) is given by:

$$\hat{\overline{y}}_d^F = f_d \overline{y}_{ds} + (1 - f_d) \overline{X}_{d(p_r)} \hat{\beta} , \tag{24}$$

where $\overline{X}_{d(p_r)} = \sum_{j \in d_r} x_{dj} / \left(N_d - n_d\right)$ is the population mean of the number of non-sampled employees in area d .

The estimator of β is given by:

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

which is the generalised least squares estimator of β and

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

is its variance and covariance matrix.

The estimator of the total for district d is obtained as:

$$\hat{t}_d^F = \sum_{j=1}^{n_d} y_{dj} + X_{d(p_r)}\hat{\beta} \ , \qquad (25)$$

and the estimator of the mean by Province is given by:

$$\hat{\bar{y}}_h^F = \frac{1}{N_h}\sum_{d\in h} N_d \hat{\bar{y}}_d^F \ , \qquad (26)$$

where $d \in h$ indicates that the sum is made in all the areas of stratum h (in this case $h = 1,2,3$ are the Provinces) and $N_h = \sum_{d\in h} N_d$ is the population total of Province h.

The estimator of the total by Province is given by:

$$\hat{t}_h^F = \sum_{d\in h} N_d \hat{\bar{y}}_d^F = \sum_{d\in h} \hat{t}_d^F \ , \qquad (27)$$

where $d \in h$ indicates that the sum is made in all the areas of stratum h (in this case $h = 1,2,3$ are the Provinces) and $N_h = \sum_{d\in h} N_d$ is the population total of Province h.

The estimator of the mean and the total for the A.C. of the Basque Country are given, respectively, by:

$$\hat{y}^F = \frac{1}{N}\sum_{h=1}^{3} N_h \hat{y}_h^F \ , \qquad \hat{t}^F = \sum_{h=1}^{3} \hat{t}_h^F \ . \qquad \text{(28) (29)}$$

The mean square errors of the estimators of the mean by district (24), in its predictive version, are given by:

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

(30)

where p_r is non-sampled census employment. If aggregations are made in order to achieve a minimum size before proceeding with the estimations, then we consider the population of the sub-class in which the projection is made as non-sampled and not that of the aggregation, which is usually higher.

The MSE for the total of the district is estimated via:

$$\hat{MSE}\left[\hat{t}_d^F\right] = N_d^2(1-f_d)^2\left[\bar{X}_{d(p_r)}\,\hat{\text{var}}(\hat{\beta})\bar{X}_{d(p_r)}\right] + \sigma^2\sum_{j\in p_r} x_{dj} \qquad \text{(31)}$$

2.4.- Totals by sector

The models presented so far allow us to obtain predictions of totals by districts, Provinces and the A.C. of the Basque Country for each sub-class, as well as their standard errors (RMSE) and variation coefficients. For the calculation of totals by sector, for example for classification A84, we simply carry out an aggregation of the totals obtained by district for each one of its sub-classes, an aggregation that can be carried out by districts, Provinces and the A.C. of the Basque Country. The standard errors of the sector are obtained as the square root of the sum of the MSE of each sub-class. Sub-classes within the same sector may have been estimated either with mixed models or by fixed ones. To obtain the results at a second level of aggregation, for example for classification A31 we aggregate the totals of the corresponding A84 sectors. The calculation of standard errors is also made assuming that the sectors are independent and thus to calculate the RMSE we calculate the square root of the sum of the MSE of the corresponding sectors.

2.5.- Calibration Process

Due to calibration, exactly the same totals as those provided by the estimator of the Industrial Survey at Province and A.C. of the Basque Country level, by A84 or another level of aggregation are obtained. Let t_d be the estimation of the total obtained under the models in district d by a specific sector, t_h the estimation of models obtained per Province and C_h the estimation of the total per Province obtained by the Industrial Survey, then the new calibrated estimation by district for each sector is given by:

$$\tilde{t}_d = t_d \frac{C_h}{t_h}.$$

Thus the calibrated total by model $\tilde{t}_h$ for each Province in each sector coincides with the estimated total C_h for each Province with the estimator of the Industrial Survey for the same sector, given that:

$$\tilde{t}_h = \sum_d \tilde{t}_d = \sum_d t_d \frac{C_h}{t_h} = \frac{C_h}{t_h} \sum_d t_d = C_h.$$

2.6.- Estimation Plan in the Industrial Survey

Eustat has programmed an *ad hoc* SAS computer programme to introduce the calculation of small-area estimations into their statistical production. This programme is specific to the Industrial Survey but which could easily be adapted to other types of economic surveys. A wide range of decisions have been taken and applied to programming:

- Both the mixed linear model and the fixed-effect linear model are calculated in the predictive version. This is due to the fact that the populations of some sub-classes are very small, so that the sample fraction is not negligible.

- The predictive version separates the prediction in two parts. That observed for the non-sampled establishments and that observed for the sampled establishments. This form is especially useful in the Industrial Survey since in some sub-classes there are atypical establishments, which is to say establishments whose difference is markedly different from the rest and which could distort the estimations considerably. These establishments are classified as non-valid sampled and when using the predictive version they do not count in

the estimation procedure, although they do count in the final prediction. Those establishments classified as non-valid are added in the same way as the rest of the sample, to the prediction of the non-observed to obtain the prediction of the total.

- It was considered necessary to establish a minimum number of establishments to proceed to the calculation of the mixed or fixed models. If this minimum number of establishments is not available, then NACE aggregations are made with one digit less. First the mixed model at this level of aggregation is estimated and if $\sigma_v^2 = 0$ or $\sigma_e^2 = 0$ then the fixed-effect model is estimated. This minimum number is currently fixed (it can be varied) in 5 establishments.

- It was decided to use the fixed-effect model when the mixed model is not valid when the decision of not grouping sub-classes together is considered a priority. This is why a fixed-effect model with 5 digits, for example, is preferred to a 3 digit mixed model.

- When an aggregation is made, this allows the estimation of the coefficients of the model, but the predictions are made specific to the sub-class under consideration.

- The use of the auxiliary variable "number of employees" introduces heteroscedasticity to the models, since the variable response y usually has greater variability as the number of employees rises. Therefore all the mixed and fixed-effect models consider that the variance of the error is proportional to the number of employees.

- In each case the totals per Province and A.C. of the Basque Country are obtained as aggregates from the estimations per district. The totals per A84 sector are obtained by aggregating the predictions obtained at sub-class level. The same occurs for the totals per Province and A.C. of the Basque Country. The procedure is the same as that used for other types of aggregation.

- In each sub-class, specific formulas are applied to calculate the square roots of the mean square errors of the Province-level predictions, not obtained as a square root of the sum of the mean square errors of the predictions per district. This is because in each sub-class the estimations by district are not independent in any of the models.

- However, once the estimations of the RMSE per sub-class for each Province are made, the calculation of the A.C. of the Basque Country estimations is direct, since now the hypothesis of independence is fulfilled. Following the calculations by sub-class, the RMSE of the A84 variable or any other grouping is obtained directly, which is to say calculating he square root of the sum of the MSE of the sub-classes that make up each sector.

Chapter
3

Application of small area estimation techniques in the Industrial Survey of the A.C. of the Basque Country. 2002 and 2003.

3.1.- District-level estimations in the Industrial Survey of the A.C. of the Basque Country. 2002 and 2003

Below we present the estimations obtained using the aforementioned estimation system in the Industrial Survey of the A.C. of the Basque Country corresponding to 2002 and 2003. The macro-magnitudes chosen for its publication are:

- Value added at factor cost
- Net sales
- Gross operating surplus
- Personnel costs
- Investment
- Pre-tax results

We also offer the Employment estimations for districts, given that this was the exogenous variable used in the small-area models employed in the Industrial Survey of the A.C. of the Basque Country.

Personnel employed by industry by province and district. 2002-2003

	2002	2003	Δ 03/02
A.C. of the Basque Country	**246.063**	**248.922**	**1,2**
Alava	**47.804**	**49.019**	**2,5**
Valles Alaveses	1.889	1.953	3,4
Llanada Alavesa	30.978	31.912	3,0
Montaña Alavesa	354	294	-16,9
Rioja Alavesa	3.778	4.042	7,0
Estribaciones del Gorbea	3.942	3.949	0,2
Cantábrica Alavesa	6.863	6.869	0,1
Bizkaia	**103.749**	**104.619**	**0,8**
Arratia-Nervión	5.173	5.021	-2,9
Gran Bilbao	60.622	61.312	1,1
Duranguesado	21.614	21.736	0,6
Encartaciones	2.128	2.139	0,5
Gernika-Bermeo	4.418	4.612	4,4
Markina-Ondarroa	3.892	4.056	4,2
Plentzia-Mungia	5.902	5.743	-2,7
Gipuzkoa	**94.510**	**95.284**	**0,8**
Bajo Bidasoa	5.490	5.602	2,0
Bajo Deba	10.452	10.643	1,8
Alto Deba	20.275	20.091	-0,9
Donostia-San Sebastián	25.464	25.508	0,2
Goierri	12.831	13.233	3,1
Tolosa	8.001	8.185	2,3
Urola Costa	11.997	12.022	0,2

Source: EUSTAT. Industry and Construction accounts

Industrial value added at factor cost and variation coefficients (cv) by province and district (thousand €). 2002-2003

	2002	cv	2003	cv	Δ 03/02
A.C. of the Basque Country	**13.008.214**	**0,01**	**13.371.649**	**0,01**	**2,8**
Alava	**2.676.614**	**0,00**	**2.747.477**	**0,00**	**2,6**
Valles Alaveses	89.743	0,03	106.061	0,02	18,2
Llanada Alavesa	1.708.733	0,01	1.763.392	0,01	3,2
Montaña Alavesa	17.455	0,19	14.211	0,16	-18,6
Rioja Alavesa	243.155	0,02	270.594	0,04	11,3
Estribaciones del Gorbea	216.196	0,02	210.776	0,01	-2,5
Cantábrica Alavesa	401.332	0,01	382.443	0,01	-4,7
Bizkaia	**5.371.448**	**0,01**	**5.561.854**	**0,01**	**3,5**
Arratia-Nervión	252.658	0,02	259.622	0,01	2,8
Gran Bilbao	3.259.555	0,01	3.395.279	0,01	4,2
Duranguesado	1.109.460	0,01	1.113.048	0,01	0,3
Encartaciones	154.460	0,05	155.544	0,04	0,7
Gernika-Bermeo	180.951	0,03	199.627	0,02	10,3
Markina-Ondarroa	149.595	0,03	167.664	0,02	12,1
Plentzia-Mungia	264.769	0,01	271.070	0,01	2,4
Gipuzkoa	**4.960.152**	**0,01**	**5.062.318**	**0,01**	**2,1**
Bajo Bidasoa	218.659	0,02	231.525	0,02	5,9
Bajo Deba	482.592	0,02	482.471	0,02	0,0
Alto Deba	996.600	0,01	1.013.516	0,00	1,7
Donostia-San Sebastián	1.585.031	0,01	1.591.077	0,01	0,4
Goierri	655.356	0,01	682.139	0,01	4,1
Tolosa	422.004	0,02	442.768	0,01	4,9
Urola Costa	599.910	0,01	618.822	0,01	3,2

Source: EUSTAT. Industrial and Construction accounts

Personnel costs for industry and variation coefficients (cv) by province and district (thousand €). 2002-2003

	2002	cv	2003	cv	Δ 03/02
A.C of the Basque Country	**7.398.012**	**0,00**	**7.808.500**	**0,00**	**5,5**
Alava	**1.477.956**	**0,00**	**1.576.926**	**0,00**	**6,7**
Valles Alaveses	57.176	0,01	61.286	0,01	7,2
Llanada Alavesa	969.544	0,00	1.024.924	0,00	5,7
Montaña Alavesa	7.968	0,06	7.258	0,04	-8,9
Rioja Alavesa	88.179	0,03	106.669	0,03	21,0
Estribaciones del Gorbea	119.351	0,01	126.500	0,01	6,0
Cantábrica Alavesa	235.738	0,00	250.291	0,00	6,2
Bizkaia	**3.108.386**	**0,00**	**3.273.633**	**0,00**	**5,3**
Arratia-Nervión	154.341	0,01	155.158	0,01	0,5
Gran Bilbao	1.879.256	0,01	1.976.019	0,01	5,1
Duranguesado	638.435	0,01	675.569	0,01	5,8
Encartaciones	57.201	0,02	62.725	0,02	9,7
Gernika-Bermeo	114.381	0,01	125.955	0,01	10,1
Markina-Ondarroa	93.588	0,01	102.151	0,01	9,1
Plentzia-Mungia	171.183	0,02	176.056	0,01	2,8
Gipuzkoa	**2.811.670**	**0,00**	**2.957.941**	**0,00**	**5,2**
Bajo Bidasoa	138.692	0,01	150.936	0,01	8,8
Bajo Deba	303.162	0,01	319.017	0,01	5,2
Alto Deba	625.672	0,00	652.605	0,00	4,3
Donostia-San Sebastián	749.887	0,01	775.827	0,01	3,5
Goierri	403.843	0,00	436.067	0,01	8,0
Tolosa	240.291	0,01	259.569	0,01	8,0
Urola Costa	350.123	0,01	363.920	0,01	3,9

Source: EUSTAT. Industry and Construction accounts

Net sales of industry and variation coefficients (cv) by province and district (thousand €). 2002-2003

	2002	cv	2003	cv	Δ 03/02
A.C. of the Basque Country	**42.393.031**	**0,01**	**43.768.410**	**0,01**	**3,2**
Alava	**8.336.340**	**0,01**	**8.614.962**	**0,01**	**3,3**
Valles Alaveses	321.016	0,02	390.287	0,02	21,6
Llanada Alavesa	5.521.407	0,01	5.564.479	0,01	0,8
Montaña Alavesa	46.122	0,17	40.494	0,17	-12,2
Rioja Alavesa	603.353	0,02	702.555	0,03	16,4
Estribaciones del Gorbea	705.855	0,02	763.485	0,01	8,2
Cantábrica Alavesa	1.138.586	0,01	1.153.661	0,01	1,3
Bizkaia	**19.341.307**	**0,01**	**20.078.450**	**0,01**	**3,8**
Arratia-Nervión	743.317	0,03	822.265	0,02	10,6
Gran Bilbao	12.733.700	0,01	13.268.830	0,01	4,2
Duranguesado	3.518.902	0,02	3.562.602	0,01	1,2
Encartaciones	422.018	0,05	417.302	0,03	-1,1
Gernika-Bermeo	673.681	0,03	727.858	0,02	8,0
Markina-Ondarroa	495.976	0,03	524.603	0,03	5,8
Plentzia-Mungia	753.713	0,03	754.989	0,02	0,2
Gipuzkoa	**14.715.384**	**0,01**	**15.074.998**	**0,01**	**2,4**
Bajo Bidasoa	581.630	0,03	630.369	0,02	8,4
Bajo Deba	1.273.758	0,03	1.238.176	0,04	-2,8
Alto Deba	3.134.142	0,01	3.202.536	0,01	2,2
Donostia-San Sebastián	4.600.028	0,01	4.557.629	0,01	-0,9
Goierri	2.039.115	0,01	2.224.916	0,01	9,1
Tolosa	1.239.249	0,04	1.315.686	0,02	6,2
Urola Costa	1.847.462	0,02	1.905.687	0,03	3,2

Source: EUSTAT. Industrial and Construction accounts

Gross operating surplus in industry and variation coefficients (cv) by province and district (thousand €). 2002-2003

	2002	cv	2003	cv	Δ 03/02
A.C. of the Basque Country	**5.610.202**	**0,01**	**5.563.149**	**0,01**	**-0,8**
Alava	**1.198.658**	**0,01**	**1.170.551**	**0,01**	**-2,3**
Valles Alaveses	32.931	0,07	44.762	0,05	35,9
Llanada Alavesa	739.335	0,01	737.566	0,02	-0,2
Montaña Alavesa	9.530	0,34	7.186	0,29	-24,6
Rioja Alavesa	156.177	0,03	164.114	0,05	5,1
Estribaciones del Gorbea	95.502	0,04	83.599	0,03	-12,5
Cantábrica Alavesa	165.183	0,01	133.325	0,03	-19,3
Bizkaia	**2.263.062**	**0,01**	**2.288.221**	**0,01**	**1,1**
Arratia-Nervión	98.254	0,04	104.809	0,03	6,7
Gran Bilbao	1.376.066	0,02	1.418.524	0,01	3,1
Duranguesado	474.025	0,02	438.507	0,02	-7,5
Encartaciones	97.238	0,07	92.094	0,05	-5,3
Gernika-Bermeo	66.908	0,07	73.657	0,05	10,1
Markina-Ondarroa	56.424	0,06	64.772	0,05	14,8
Plentzia-Mungia	94.146	0,03	95.858	0,04	1,8
Gipuzkoa	**2.148.482**	**0,01**	**2.104.377**	**0,01**	**-2,1**
Bajo Bidasoa	79.729	0,05	79.886	0,04	0,2
Bajo Deba	178.555	0,04	164.393	0,05	-7,9
Alto Deba	373.600	0,02	358.072	0,02	-4,2
Donostia-San Sebastián	831.687	0,02	814.317	0,02	-2,1
Goierri	254.308	0,02	247.468	0,02	-2,7
Tolosa	181.731	0,04	185.976	0,04	2,3
Urola Costa	248.871	0,03	254.264	0,03	2,2

Source: EUSTAT. Industry and Construction accounts

Investment carried out by industry and variation coefficients (cv) by province and district (thousand €). 2002-2003

	2002	cv	2003	cv	Δ 03/02
A.C. of the Basque Country	**3.211.667**	**0,02**	**2.717.879**	**0,01**	**-15,4**
Alava	**660.740**	**0,02**	**510.428**	**0,01**	**-22,7**
Valles Alaveses	36.549	0,55	26.231	0,03	-28,2
Llanada Alavesa	457.031	0,02	332.071	0,02	-27,3
Montaña Alavesa	2.396	0,17	2.198	0,20	-8,3
Rioja Alavesa	58.364	0,15	60.380	0,05	3,5
Estribaciones del Gorbea	55.831	0,45	34.254	0,04	-38,6
Cantábrica Alavesa	50.569	0,04	55.295	0,02	9,3
Bizkaia	**1.459.249**	**0,02**	**1.364.705**	**0,01**	**-6,5**
Arratia-Nervión	41.686	0,35	41.403	0,05	-0,7
Gran Bilbao	1.030.310	0,02	1.079.300	0,01	4,8
Duranguesado	248.724	0,27	165.607	0,03	-33,4
Encartaciones	42.907	1,80	18.728	0,08	-56,4
Gernika-Bermeo	27.135	0,54	8.636	0,18	-68,2
Markina-Ondarroa	25.626	0,07	24.139	0,06	-5,8
Plentzia-Mungia	42.860	0,35	26.891	0,06	-37,3
Gipuzkoa	**1.091.678**	**0,02**	**842.746**	**0,02**	**-22,8**
Bajo Bidasoa	53.852	0,54	37.669	0,04	-30,1
Bajo Deba	93.377	0,23	79.199	0,05	-15,2
Alto Deba	193.433	0,02	213.272	0,01	10,3
Donostia-San Sebastián	370.970	0,04	216.867	0,03	-41,5
Goierri	142.977	0,30	115.465	0,03	-19,2
Tolosa	137.092	1,17	79.849	0,15	-41,8
Urola Costa	99.976	0,05	100.425	0,04	0,4

Source: EUSTAT. Industry and Construction accounts

Pre-tax results for industry and variation coefficients (cv)
by province and district (thousand €). 2002-2003

	2002	cv	2003	cv	Δ 03/02
A.C. of the Basque Country	**3.213.405**	**0,02**	**2.904.928 0**	**0,02**	**-9,6**
Alava	**772.874**	**0,02**	**579.541**	**0,02**	**-25,0**
Valles Alaveses	13.060	0,15	18.734	0,10	43,5
Llanada Alavesa	468.551	0,02	316.075	0,04	-32,5
Montaña Alavesa	6.934	0,46	6.568	0,31	-5,3
Rioja Alavesa	145.513	0,02	103.511	0,07	-28,9
Estribaciones del Gorbea	40.156	0,07	61.471	0,03	53,1
Cantábrica Alavesa	98.660	0,02	73.182	0,05	-25,8
Bizkaia	**1.344.669**	**0,02**	**1.274.144**	**0,02**	**-5,2**
Arratia-Nervión	64.725	0,06	67.487	0,04	4,3
Gran Bilbao	838.007	0,02	774.138	0,02	-7,6
Duranguesado	263.632	0,03	243.464	0,03	-7,7
Encartaciones	63.818	0,10	62.697	0,08	-1,8
Gernika-Bermeo	37.564	0,11	31.403	0,11	-16,4
Markina-Ondarroa	23.769	0,13	34.147	0,08	43,7
Plentzia-Mungia	53.155	0,04	60.808	0,06	14,4
Gipuzkoa	**1.095.862**	**0,03**	**1.051.243**	**0,03**	**-4,1**
Bajo Bidasoa	37.231	0,10	35.167	0,09	-5,5
Bajo Deba	101.389	0,07	91.380	0,09	-9,9
Alto Deba	139.336	0,04	127.648	0,04	-8,4
Donostia-San Sebastián	429.737	0,03	418.766	0,03	-2,6
Goierri	127.438	0,04	108.258	0,05	-15,1
Tolosa	108.981	0,06	123.785	0,05	13,6
Urola Costa	151.750	0,05	146.238	0,05	-3,6

Source: EUSTAT. Industry and Construction accounts

3.2.- Value added at factor cost and Personnel employed by district in the Industrial Survey of the A.C. of the Basque Country. Conclusions.

Below we extract the main conclusions from the analysis of the results obtained at district level for the most relevant variables of the Industrial Survey: Value added at factor cost and Personnel employed.

From the estimation of the data for Basque industry by districts for the year 2003, we see that the district of Greater Bilbao represents a quarter of the industrial value added (25.4%) of the A.C. of the Basque Country. Second is the Llanada Alavesa with 13.2% and third the district of Donosita-San Sebastian, with 11.9%. These three districts contribute slightly over half the industrial value added (50.5%) of the A.C. of the Basque Country in 2003 and 50.4% in 2002. In terms of industrial personnel employed, these three provinces employ 47.7% of the industrial workers in the A.C. of the Basque Country.

The Duranguesado, Alto Deba and Cantábrica Alavesa contributed 18.8% of industrial value added in the A.C. of the Basque Country and represented 19.6% of the personnel employed in the sector. Therefore, 69.3% of industry is concentrated in six of the twenty districts in the A.C. of the Basque Country. However, this phenomenon of concentration-dispersion varies from one province to another.

In the province of Alava, almost two thirds of the industrial value added (precisely 64.2%) is concentrated in the Llanada Alavesa, while the district of the Montaña Alavesa houses only 0.5%. The four remaining districts in this province have an industrial weight that stands between 13.9% for Cantábrica Alavesa and 3.9% in the Valles Alaveses.

The effect of the concentration of industry by districts is also significant in Bizkaia, where 81.0% is accumulated in two districts, Greater Bilbao (61.0%) and the Duranguesado (20.0%), while the province of the Encartaciones hardly accounts for 2.8% of industrial activity in the province. The third district with the greatest industrial weight is Arratia-Nervión, with 4.7% of the industrial value added in this province.

In Gipuzkoa the concentration is lower and the inter-district balance is more weighted than in the other three provinces. Half of the industry is concentrated in two districts: Donostia-San Sebastian (31.5%) and Bajo Deba (20.0%), but at the same time, there are districts like Goierri and Urola Costa, with 13.5% and 12.2% relative provincial industrial weight, respectively. By contrast, the district with the lowest industrial weight is Bajo Bidasoa, in Gipuzkoa, but accounts for 4.6% of the industry in the province, compared to 0.5% of the Montaña Alavesa and 2.8% of Encartaciones.

Industry in the A.C. of the Basque Country obtained 2.8% nominal growth in 2003 and this average growth was surpassed by the districts of Valle Alaveses (18.2%), Rioja Alavesa (11.3%) and Llanada Alavesa (3.2%) in Alava; by Marquina-Ondarroa (12.1%),

Gernika-Bermeo (10.3%) and Greater Bilbao (4.2%) in Bizkaia; and by Bajo Bidasoa (5.9%), Tolosa (4.9%), Goierri (4.1%) and Urola Costa (3.2%) in Gipuzkoa.

With regard to the evolution of personnel employed in 2003, all the districts in Alava showed a positive profile, except for Montaña Alavesa. Only two districts in Bizkaia (Arratia-Nervión and Plentzia-Mungia) decreased employment. With reference to Gipuzkoa, only one district (Alto Deba) reduced industrial workers, the other six increasing by varying degrees.

Chapter

4

Conclusions

The small-area models described in this paper allow us to obtain estimations of totals of the various macro-magnitudes of the Industrial Survey at district level. In general, the results are highly satisfactory, given that the estimated variation coefficients are undeniably modest in those districts where the sample representation and the population size are relatively high. In districts where this was not the case, the variation coefficients estimated are higher.

Eustat will produce from this year on district level yearly estimations of the Industrial Survey with at least the same as the one in this document. This implies that in the near future district level estimation time series will be available and, therefore, more complete short time analyses, providing a better understanding of the evolution of the main economical macro-magnitudes within the A.C. of the Basque Country, will be possible. This information can be of great interest for the design of different regional policies.

It should be added that carrying out this small-area estimation project entailed a stringent review of the documentation processes of the Industrial Survey, which has translated into methodological improvements in the calculation of estimators and in the calculation of sampling errors.

Eustat continues with its research of the methodology of small-area estimation models with the intention of offering shortly quality estimations at higher levels of detail than those currently published. Work is being done in applying this methodology in other Eustat surveys, not only in surveys related to the industrial activity but also in surveys where all or some of the variables of interest are discrete.

Bibliography

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

[1] Henderson, C.R. (1975). Best Linear Unbiased Estimation and Prediction Under a Selection Model. *Biometrics*, 31, 423-447.

[2] Kackar, R.N. and Harville, D.A.. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. Journal of the American Statistical Association.

[2] Pfefferman, D. (2002). Small Area Estimation – New Developments and Directions. International Statistical Review, 70, 125-143.

[3] Prasad, N.G.N and Rao, J.N.K (1990). The Estimation of Mean Squared Error of Small Area Estimators. Journal of the American Statistical Association, 85, 163-171.

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

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

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