Small Area Estimation in the Industrial Survey

Iosune Azula, Patxi Garrido, Haritz Olaeta

EUSTAT, Economic Statistics
Donostia 1
Vitoria-Gasteiz, Spain
iosune-azula@eustat.es, patxi-garrido@eustat.es, Haritz.Olaeta@eustat.es

Keywords: Fixed Effects Models, Linear Mixed Models, Local Estimates, Coherence

1. Introduction

Aware of the increasing demand of small area estimations, Eustat started publishing estimates of the main variables of the Industrial Survey for *comarcas* (i.e., administrative clusters of municipalities) or counties in 2005. There are 20 such counties within the three provinces of the Basque Country, being some of them extremely small in terms of industrial activity.

The Industrial Survey is designed to provide estimates at province level, so that the use of small area estimation models is necessary in order to obtain reliable estimates. The industrial activity of the Basque Country is not evenly distributed among the 20 counties and both the importance and the size of the industrial sector vary hugely from county to county. In fact, there are counties where the industrial activity is extremely limited. The increase of sample sizes required to obtain reliable estimations for such counties would be certainly costly and senseless in terms of unnecessary burden increase of our respondents.

In this work, the small area estimation process implemented in Eustat for the Industrial Survey is explained in some detail together with the relevant issue of coherence with the estimates that provides the Industrial Survey at province level and for the whole of the Basque Country.

2. Estimation Procedure

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. In order to obtain county-level estimations for the Industrial Survey, Eustat employs two types of models in a pre-defined manner: 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 counties and another term differentiating the elements of each county d ($d = 1,2,3$). The differentiating term is made up of random effects (v_d), so that all the data from the same county shares the same random effect. In the case of the fixed-effect model, there are no differentiating terms for each county since the systematic part is

common to all the counties. However, specificity is achieved by projecting the common estimated coefficient onto the specific auxiliary information of each county.

The only auxiliary information used in the small area estimation process built for the Industrial Survey in 2005 is the employment of the industrial establishments according to Eustat´s Directory of Economic Activities, the framework used in the Industrial Survey to extract samples that constitutes the base in the estimation process as the Industrial Survey uses a model-assisted composite estimator.

2.1. The Mixed Linear Model

For each activity sub-class (5 digit NACE), in each *comarca* or county d there are N_d establishments in the population according to the Directory of Economic Activities, thus $N = \sum_d N_d$ is the total population . In this given sub-class, n establishments are sampled of which n_d belong to county d . The following mixed linear heteroscedastic model is proposed:

$$y_{dj} = \beta_0 + \beta_1 x_{dj} + v_d + e_{dj}, \qquad d = 1,2,3 \quad j = 1,...,n_d,$$

where, for establishment j of county d , y_{dj} is the value taken by the variable of interest and x_{dj} is the number of employees in the establishment according to the Directory of Economic Activities. The total number of sampled establishments in county d is n_d . The constants β_0 and β_1 are the fixed effects of the model (constants for all counties). The common random effect for all the establishments of county d is v_d , and e_{dj} are the specific random errors of each establishment. 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.

It can be shown that the predictive version of the estimator t of the total of the variable of interest y can be presented as follows:

$$\hat{t}_d = X_{d(p_r)}^{|} \hat{\beta} + (N_d - n_d)\hat{\gamma}_{dc}\left(\bar{y}_{dc} - \bar{x}_{dc}\hat{\beta}\right) + \sum_{j=1}^{n_d} y_{dj}, \qquad d = 1,2,3$$

where $X_{d(p_r)}$ is the total number of employees in county d for all the non-sampled establishments and any further details is omitted in order to keep it simple, including the estimator of the mean square error of the estimates.

2.2. The Fixed Effects Linear Model

For each establishment, the proposed fixed effects linear model is:

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

where β is the single fixed effect of the model (constant for all counties) and as in the previous case, to correct the heteroscedasity present in the data, we use the weights $c_{dj} = 1/x_{dj}$.

The predictive version of the estimator of the total for county d is then obtained as:

$$\hat{t}_d^F = \sum_{j=1}^{n_d} y_{dj} + X_{d(p_r)}\hat{\beta}, \qquad d = 1,2,3$$

where the notation has already been introduced. The mean square errors of the estimates are easily obtained .

2.3 Benchmarking and calibration for province level estimates

The estimation procedure starts for each county at activity sub-class (5 digit NACE) level. It is considered necessary to establish a minimum number of establishments in a given county to proceed to the calculation of the mixed or fixed models. This is currently fixed to 5 establishments considered, by the expert, to be valid to be used in the estimation process to estimate the economic values of others. 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.

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

In each activity sub-class, the totals per Province and A.C. of the Basque Country are obtained as aggregates from the estimations per county. These aggregated values are calibrated to totals provided by the Industrial Survey so that coherence between small area estimations and Industrial Survey estimates is guaranteed. 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 NACE or other levels of aggregation are obtained.

3. Results

There are yearly estimates at county level of variables Employment, Personnel Costs, Gross Value Added, Net Sales, Gross Operating Surplus, Investment and Pre-tax Results for the period 2002-2009. Estimates are published one month later than the results of the Industrial Survey.

In Table 1, the estimates of the variable Gross Value Added together with their corresponding coefficients of variation are shown for the last 5 years. There is a break in the series in year 2009 due to the exclusion of the sector of Energy from the estimates after this year.

Table 1. Estimates of Gross Value Added and Coefficients of Variation

	2005	cv	2006	cv	2007	cv	2008	cv	2009	cv
C.A. de Euskadi	15.701.676	0,01	16.624.164	0,01	17.615.267	0,01	17.997.138	0,01	11.928.556	0,01

Alava	**3.320.782**	**0,02**	**3.446.405**	**0,02**	**3.594.157**	**0,01**	**3.696.723**	**0,02**	**2.541.425**	**0,01**
Valles Alaveses	128.980	0,04	130.203	0,07	142.378	0,04	153.354	0,04	118.327	0,02
Llanada Alavesa	2.073.654	0,01	2.135.805	0,01	2.184.633	0,01	2.248.621	0,02	1.505.244	0,01
Montaña Alavesa	15.247	0,11	16.642	0,06	19.077	0,04	23.383	0,18	22.085	0,17
Rioja Alavesa	377.162	0,08	399.311	0,09	382.435	0,23	378.430	0,05	331.336	0,06
Estribaciones del Gorbea	255.874	0,02	259.191	0,04	259.903	0,02	260.487	0,03	153.292	0,02
Cantábrica Alavesa	469.864	0,01	505.252	0,01	605.730	0,01	632.449	0,01	411.140	0,01
Bizkaia	**6.609.114**	**0,01**	**6.911.272**	**0,01**	**7.281.370**	**0,01**	**7.546.849**	**0,01**	**4.740.470**	**0,01**
Arratia-Nervión	266.373	0,03	279.814	0,03	311.577	0,02	297.816	0,02	275.480	0,02
Gran Bilbao	4.232.537	0,02	4.379.899	0,01	4.538.472	0,01	4.721.396	0,01	2.744.788	0,01
Duranguesado	1.229.797	0,01	1.309.050	0,02	1.454.613	0,01	1.479.129	0,01	956.389	0,01
Encartaciones	187.981	0,01	182.071	0,07	140.240	0,07	150.699	0,07	97.944	0,05
Gernika-Bermeo	209.127	0,01	232.944	0,04	243.563	0,02	250.589	0,03	191.436	0,02
Markina-Ondarroa	189.944	0,03	200.550	0,02	248.099	0,01	261.217	0,02	165.338	0,05
Plentzia-Mungia	293.354	0,01	326.942	0,04	344.806	0,03	386.004	0,03	309.096	0,02
Gipuzkoa	**5.771.780**	**0,01**	**6.266.487**	**0,01**	**6.739.740**	**0,01**	**6.753.566**	**0,01**	**4.646.661**	**0,01**
Bajo Bidasoa	283.288	0,02	298.103	0,03	304.669	0,02	324.012	0,03	251.283	0,02
Bajo Deba	518.738	0,02	550.285	0,02	587.361	0,02	591.800	0,02	455.701	0,02
Alto Deba	1.103.086	0,01	1.202.213	0,01	1.273.822	0,01	1.156.899	0,01	864.729	0,01
Donostia-San Sebastián	1.826.527	0,04	1.940.767	0,02	1.951.588	0,01	2.059.400	0,02	1.139.975	0,02
Goierri	897.793	0,03	983.691	0,01	1.238.304	0,01	1.186.869	0,01	865.071	0,02
Tolosa	449.080	0,02	501.893	0,06	550.062	0,05	597.671	0,05	496.191	0,01
Urola Costa	693.268	0,02	789.535	0,02	833.934	0,02	836.915	0,02	573.711	0,01

There are large differences in the size of the industrial sector between counties. In fact, over half of the industrial value added is concentrated in the counties where the provincial capital cities are situated (Llanada Alavesa, Gran Bilbao and Donostia-San Sebastian). There are, on the other side, counties where the industrial activity is extremely small, specially in the province of Alava (for instance the county *Montaña Alavesa*, where the value added represents only 0.097% of the total).

It is clear analysing Table 1 that the greater the Value Added the more accurate the estimates in general. The estimated coefficients of variations are moderate for big and medium size counties, in terms of industrial activity, whereas for counties like the mentioned *Montaña Alavesa*, the coefficient of variation is as high as 18% in 2009.

In some counties, the accuracy of the estimates differs considerable over the years. For instance, in *Rioja Alavesa*, the coefficient in 2007 is 23% having had before and after that year one-digit coefficients. The fact that the sample of the Industrial Survey is designed to provide accurate estimates at provincial level can explain the variability in the accuracy over the years in some counties, specially in the smallest ones where the number of sampled establishments might vary considerably from year to year due to the probabilistic nature of the sample.

4. Conclusions and Future Work

The county-level estimates of the industrial activity in the Basque Country have provided reliable and coherent estimates for the last 6 years. This valuable yearly information is of great use for local authorities and there is an increasing demand for producing county-level estimates of other

economic sectors such as Construction and Services. Eustat is working actually on small area estimates on these sectors.

The recent availability of relevant auxiliary information like the Mercantile Register and Tax Information from the three tax authorities of the A.C. of de Basque Country (available in the near future), makes it necessary a complete revision of the estimation process. More sophisticated and valuable auxiliary information will be available so that alternative models and estimates will be analysed. Most probably, better estimates in terms of accuracy will be obtained together with a more complete set of estimated variables.

Another natural extension is to provide estimates for capital cities and medium-sized municipalities where the indusdtrial activity is relevant and have sufficient size to be estimated using more complete auxiliary information.