Small-area estimation in Official Statistics:
ICT survey in Enterprises of the Basque Country

Jorge Aramendi (j-aramendi@eustat.es), Elena Goni (egoni@eustat.es), Anjeles Iztueta (aiztueta@eustat.es), Jose Miguel Escalada (jm_escalada@eustat.es)

Euskal Estatistika Erakundea / Instituto Vasco de Estadística / Basque Statistics Office- Eustat

Key words: Small area estimate, Logistic regression model, Bootstrap

In 2003, Eustat, aware of the growing demand for increasingly disaggregated quality statistics, set up a research team made up of members of Eustat and the Public University of Navarre. The objective was to work on improving the estimation techniques in various statistical operations and to introduce small area estimation techniques based on statistical production models.

In 2004, work began on studying the first survey and so far, indirect small area estimates have been studied and disseminated[1] in 5 surveys: Industrial Survey (data dissemination in 2005), Labour Force Survey (data dissemination in 2008), Information Society Survey – Families (data dissemination in 2009), Technological Innovation Survey (data dissemination in 2010) and our last study which objective has been obtaining district-level estimates in the Information and Communications Technologies (ICT) Survey in Enterprises of the Basque Country.

The ICT Survey was implemented in 2001 to find out the level of penetration of the new technologies in the fabric of the Basque economy; and to monitor said penetration in view of the current and future revolutionary importance and scope of technological changes. The information is obtained for the main characteristics at the level of the provinces, in line with the sampling design and specific data collected from a questionnaire.

The ICT Survey is a panel of around 7500 establishments with an annual renewal between 15% and 20% of the elements. This updating is made according to a new sample distribution that, while respecting the original design, shows the new distribution of the population and the sample of title-holders in the strata needed to complete ones that are empty and over-represented.

The main characteristics of the sample design are:

[1] www.eustat.es

- Sampling units: The establishments listed in the Directory of Economic Activities of the Basque Country

- Sample size: Around 7,500 establishments, distributed according to three variables: province, activity and employment stratum.

- Sample type: Stratified

- Allocation: Optimal

- Drawing: Random within each province, employment stratum and activity. The sectorisation of activities is specific for this survey, based on the A38 classification of the National Classification of Economic Activities CNAE 2009 for this survey and disaggregating some services sectors (65 branches of activity).

- There are six employment strata: 0-5, 6-9, 10-19, 20-49, 50-99, >=100. The stratum of older than 99 employed is for census purposes. The rest of the strata arose from the combination of the results of an ad hoc study on optimal stratification, according to the data in the 2000 survey, to ensure comparability with the characteristics of the surveys in other European countries.

To estimate survey characteristics, an estimator based on the establishments in each grouped stratum is considered. The grouped strata are made up of the resulting combinations of the 3 provinces, 64 branches of activity and 3 employment strata. (Six employment strata are grouped into three so the number of strata will not be too high). Thus, the result is a theoretical matrix of 585 strata. Those that are not represented in the directory are eliminated, and the calculation of the elevators continues.

A direct estimator, the Horvitz-Thompson estimator, is used to calculate the estimates. The estimation for the total in a population or for the total in different population domains (province, activity sector, employment stratum, etc.) are calculated using the Horvitz-Thompson estimator.

This sample is designed to provide quality estimates, in terms of the estimated mean squared error, by province and by activity sector, but our users demand a more disaggregated information, in particular they demand district-level estimates (20 geographic areas) for the main variables.

Obtaining district estimations of their main variables is unfeasible using direct techniques given that the sample size is clearly insufficient and increasing the sample size would be obviously expensive. So we decided to improve the estimation methodology of this survey and to introduce small area estimation techniques. We performed a study to review the methodology, to analyze the relationship between the variables and to evaluate different estimates with the auxiliary information available.

The studied goal-variables for each district-level are total of: establishments with Internet and computer, establishments that make e-commerce, establishments with freeware operating-systems, establishments with electronic data exchange and establishments that make web proceedings with the public administration.

The auxiliary information available is not very extensive, but it is very reliable. We have a Directory for economic establishments that we are constantly updating with different surveys and administrative sources. The employment and the activity are key variables regularly updated.

Different types of estimates, design-based and model-based, have been evaluated using the information available in the latest business establishment register. The estimators have been assessed according to their Total Mean Square Errors, calculated with Bootstrap method.

The selected method is based on a logistic regression model. In a first level a logistic regression model using the employment strata, the activity sector and the district level. When there is not enough information (sample size or accuracy), we resort to a second level, choosing a logistic regression model, that uses information from the employment strata, the activity and an aggregation of district levels. The final results are produced with a province level calibration.

First level model:

$$logit(p^1) = \log \frac{p^1}{1-p^1} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{19} x_{19} + \beta_{20} x_{20} + \beta_{21} x_{21} + \beta_{22} x_{22} + \cdots + \beta_{47} x_{47}$$

where

- p1 is the proportion of establishments that responds affirmatively in a certain variable.
- β0 is the intercept.
- β1,…, β19 are the coefficients of explanatory variables for the 20 districts.
- β20, β21 are the coefficients of explanatory variables for the 3 employment strata.
- β22,…, β47 are the coefficients of explanatory variables for the 27 categories in the activity classification.

Second level model:

$$logit(p^2) = \log \frac{p^2}{1-p^2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \cdots + \beta_{30} x_{30}$$

where

- p2 is the proportion of establishments that responds affirmatively in a certain variable.
- β0 is the intercept.
- β1, β2 are the coefficients of explanatory variables for the 3 provinces.
- β3, β4 are the coefficients of explanatory variables for the 3 employment strata.
- β5,…, β30 are the coefficients of explanatory variables for the 27 categories in the activity classification

To obtain estimators of the accuracy measurement of the estimate the mean squared error is calculated using the bootstrap resampling method. The sub-samples are obtained using stratified random sampling, where the strata are defined by the province, activity and the 3 employment strata. The population used to calculate the mean squared error is based on the sample data, simulating the structure of establishments in the Basque Country for the study variables.

In general, the results offer acceptable levels of quality in terms of accuracy. The estimated coefficients of variation (CV) are not excessively high, given the relatively small samples and population in some districts.

The majority of the CV-s obtained in the estimations do not exceed 15%. The results for some variables are not so good but this is due to the low impact / frequency of the variable in the population, electronics commerce (Internet purchases + sales).

The result of the study is a computer programme based on SAS that is used to analyse this methodology and to apply the mentioned estimators. Specific macro programmes are written which execute the different phases: production of estimations by district and the calculation of the mean squared errors for the different methods.

District-level estimates for the ICT survey have been disseminated in our web http://en.eustat.es/estadisticas/tema_471/opt_0/ti_Companies/temas.html
of three years, 2010, 2011 and 2012, and in the coming months we will disseminate results of 2013.