Using the Commercial Register to Reduce Response Burden in Economic Structural Statistics

María Victoria García Olea, Patxi Garrido, Haritz Olaeta

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

Keywords: Commercial Register, Reduction of Response Burden, Composite Estimation

1. Introduction

The growing burden that respondents are undertaking in the last years increases the difficulties that might arise from high non-response rates such as non-response and imputation bias and variance estimation.

EUSTAT, the Statistical Office of the Basque Country, is really concerned with this phenomenon and is actually working on a long-term project that deals with how to increase the quality of our Economic Structural Statistics without increasing the response burden of our respondents. Actually, the ultimate goal is to increase the quality (not only in terms of errors but also in terms of timeliness, decrease of duplications, etc.) by using data from different sources such as administrative registers.

In what follows, we will focus on the Commercial or Mercantile Register that due to its nature in Spain, it constitutes an important data repository for Economic Structural Statistics. The population that is covered in this register is a large subpopulation covered in Economic Structural Statistics. In this work our interest will focus primarily in the Industrial Statistics even though its application in other structural statistics is being developed.

2. Data Contained in the Commercial Register

All the mercantile companies with fiscal duties in Spain are forced by law to submit yearly to the Commercial Registry of the province they belong to, according to their fiscal identity code, the following information using normalized forms:

1. General identification data sheet
2. Balance sheet
3. Profit & Loss account

In 2003, EUSTAT signed a Cooperation Agreement with the Commercial Registrars of the Basque Country and the Association of Spanish Property and Commercial Registrars. The aim was to obtain information not only of those companies that are fiscally located in the Basque Country but of all the companies with an economic activity within the Basque Country. Due to these

Agreements, for the last few years the statistical use of the data included in the registers has become possible. EUSTAT carries out and releases the "Equity Accounts of Non-Financial Companies" statistics using the data sets of the registers.

The population of the Commercial Register includes all the activities covered by the Industrial Survey. However, there is no information at all about self-employees which are relevant in some sectors like, for instance, Textile and Clothing Industry where self-employees account for more than 50% of the establishments. However, overall they represent less than 30% of all the industrial establishments accounting for a moderate fraction of the value added and the employment. Therefore, the overlapped population in the Industrial Survey and the Mercantile Register is by all means significant.

3. Localisation of the activity

The Directory of Economic Activities of EUSTAT collects information from all the establishments that carry out economic activities in the Basque Country. It includes the reference population of the Industrial Statistics completely and, actually, it serves as a framework of reference to select and extract the samples.

Both for companies fiscally linked to one of the provinces of the Basque Country and for companies with fiscal identification in any other province but with at least one establishment in the Basque Country, an Indirect Ratio Estimation is performed to estimate for each of their establishments in the Basque Country the values of the variables. The ratio of the employment in each establishment, according to the Directory of Economic Activities, and the employment in the whole company, according to the Commercial Register, is applied to the values of the whole company to estimate the corresponding values of each of its establishments.

4. Estimation procedure

4.1. Composite Estimator

The estimator used to obtain estimates from the Mercantile Register is a composite estimator that can be expressed as a linear combination of a direct ratio estimator and an indirect synthetic estimator.

For a given sector of activity and employment group, the total of any given variable y in province h ($h = 1,2,3$), is given by:

$$\hat{t}_{yh.C} = \phi_h \hat{t}_{yh.D} + (1-\phi_h)\hat{t}_{yh.SYN} \quad \text{with} \quad \phi_h = \frac{n_h^{|}}{N_h}$$

where $n_h^{|} = n_h - n_{ho}^{|}$ is the number of establishments in the Commercial Register considered not to be atypical[1], N_h is the total number of establishments in province h and the direct estimator (D) and the synthetic indicator (SYN) are given by:

[1] Those establishments that show a particularly high or low performance so that should not be used in the estimation process in order to avoid biases. Their observed values are simply added to the corresponding total.

$$\hat{t}_{yh.D} = \sum_{j=1}^{n_h^{|}} y_{hj} + \sum_{j=1}^{n_{ho}^{|}} y_{hj} + E_h^{nm}\,\hat{\beta}_h = \sum_{j=1}^{n_h^{|}} y_{hj} + \sum_{j=1}^{n_{ho}^{|}} y_{hj} + E_h^{nm}\,\frac{\sum_{j=1}^{n_h^{|}} y_{hj}}{\sum_{j=1}^{n_h^{|}} e_{hj}},$$

$$\hat{t}_{yh.SYN} = \sum_{j=1}^{n_h^{|}} y_{hj} + \sum_{j=1}^{n_{ho}^{|}} y_{hj} + E_h^{nm}\,\hat{\beta} = \sum_{j=1}^{n_h^{|}} y_{hj} + \sum_{j=1}^{n_{ho}^{|}} y_{hj} + E_h^{nm}\,\frac{\sum_{h=1}^{3}\sum_{j=1}^{n_h^{|}} y_{hj}}{\sum_{h=1}^{3}\sum_{j=1}^{n_h} e_{hj}},$$

where $E_h^{|}$ is the total non-atypical employment in the register for the employment group in province h and E_h^{nm} the employment not included in the register in province h for the employment group. The variable of interest is given by y_{hj} and e_{hj} is the employment of establishment j of province h.

The direct estimator uses only auxiliary information from the province of interest (unbiased but often high variability when the sample size is small), whereas the indirect estimator uses aggregate information from all the provinces (bigger effective sample sizes, usually biased but more accurate). The higher the ratio of non-atypical establishments in the register in a given province, the more weight is given to the direct estimator. On the other hand, the higher the ratio of non-atypical establishments not included in the register in a given province, the more weight is given to the indirect estimator that uses auxiliary information from the three provinces. The idea behind this estimator using these particular weights is to try to capture the advantages of both the direct estimator and the indirect one trying to avoid the disadvantages of both.

The Mean Square Error can be estimated by the following approximation:

$$\hat{MSE}\left(\hat{t}_{yh.C}\right) \approx \phi_h^2\,\hat{MSE}\left(\hat{t}_{yh.D}\right) + \left(1-\phi_h\right)^2\,\hat{MSE}\left(\hat{t}_{yh.SYN}\right) + 2\phi_h\left(1-\phi_h\right)\left[\hat{MSE}\left(\hat{t}_{yh.D}\right) - \hat{t}_{yh.SYN}\left(bias_{h.SYN}\right)\right].$$

where the details will not be given in order to keep it simple.

4.2. Estimation using data from the Mercantile Register

As shown in Figure 1, from the Mercantile Registers, in 2009 there is information of 34% of the establishments of the population, compared to 18% of population coverage in the Industrial Survey. For the subpopulation of establishments with less than 20 employees, the difference is even bigger, with 25% of coverage with the Commercial Register versus a 7% in the Industrial Survey.

Overall, due mainly to a greater population coverage, the estimated coefficients of variation for the estimates of the variable Gross Value Added are smaller using the information included in the Mercantile Register. This gain in accuracy happens in almost all the sectors considered (15 sectors, using Eustat´s own A31 classification) in both cases, for the whole population and for small establishments with less than 20 employees.

The gain in accuracy, nevertheless, seems to be quite moderate given the greater “sample” sizes. This might be due to the lack of the subpopulation of self-employees and the lack of an optimum random sampling scheme in the Commercial Register.

Table 1. Population coverage and coefficients of variation for the Gross Value Added. Industrial Survey and Mercantile Register. 2009.

Total					Estab. <20				
	nresp IS	cv_ vabcf	nresp MR	cv_ vabcf		nresp IS	cv_ vabcf	nresp MR	cv_ vabcf
A31					A31				
Total	0,18	0,01	0,34	0,00	Total	0,07	0,04	0,25	0,01
4	0,29	0,03	0,37	0,05	4	0,14	0,04	0,24	0,07
5	0,16	0,03	0,21	0,02	5	0,09	0,07	0,14	0,05
6	0,07	0,08	0,14	0,06	6	0,05	0,11	0,12	0,09
7	0,10	0,00	0,08	0,00	7	0,08	0,00	0,05	0,01
8	0,12	0,03	0,27	0,02	8	0,08	0,05	0,23	0,03
9	0,10	0,02	0,29	0,01	9	0,04	0,06	0,24	0,03
10	1,00	0,00	1,00	0,00	10				
11	0,37	0,02	0,51	0,02	11	0,13	0,07	0,34	0,08
12	0,27	0,01	0,45	0,01	12	0,08	0,07	0,31	0,03
13	0,22	0,04	0,38	0,03	13	0,10	0,10	0,28	0,08
14	0,21	0,01	0,42	0,00	14	0,07	0,03	0,33	0,01
15	0,25	0,01	0,47	0,01	15	0,07	0,05	0,34	0,04
16	0,22	0,01	0,36	0,01	16	0,06	0,07	0,24	0,05
17	0,34	0,00	0,49	0,00	17	0,06	0,03	0,28	0,07
18	0,10	0,03	0,21	0,02	18	0,04	0,07	0,17	0,03
19	0,19	0,08	0,24	0,00	19	0,08	0,23	0,15	0,03

4.3. Estimation combining data from the Mercantile Register and the Industrial Survey

The next step is to combine the information obtained both from the Mercantile Register and the Industrial Survey.

In Table 2, the results obtained combining both sources are compared to those obtained in the Industrial Survey for each of the 15 industrial sectors. The population covered using both data sources is 38% of all the establishments and 29% of small ones. This implies that there is a very important overlapping between both samples as most of the establishments that fill up the Industrial Survey bring also their economic information to the corresponding Commercial Register. When for a given establishment there is information in both sources, the administrative source will be used as it could imply a burden reduction with quality assurance.

Comparing the estimated coefficients of variation of the variable Gross Value Added, the estimates combining both sources are more accurate overall, specially compared to those obtained using only the Commercial Register. By sectors, the accuracy gain is generalised except for sector 6 (Textile and Clothing Industry) where with a higher coverage of the population (17% instead of 7%) the estimated coefficient of variation of the Gross Value Added is higher (12% versus 8%). This loss of accuracy is slightly higher in the subpopulation of small establishments. This sector is highly heterogeneous, with huge differences in productivity across establishments. In addition, over half of the population in this sector is constituted by self-employees whose information is not included in the Commercial Register and it is probably under-sampled in the Industrial Survey.

A similar circumstance occurs in sector 17 (Material for Transportation) in the subpopulation of small establishments. Using only the sample used in the Industrial Survey with a coverage of 6% the estimated coefficient of variation is of 3%. Adding the information provided by the Commercial Register, the coverage increases to 32% whereas the coefficient of variation raises to 5%.

Table 2. Population coverage and coefficients of variation for the Gross Value Added. Combining the Industrial Survey and the Mercantile Register. 2009.

Total					Estab. <20				
	nresp IS	cv_ vabcf	nresp MR-IS	cv_ vabcf		nresp IS	cv_ vabcf	nresp MR-IS	cv_ vabcf
A31					A31				
Total	0,18	0,01	0,38	0,00	Total	0,07	0,04	0,29	0,01
4	0,29	0,03	0,46	0,02	4	0,14	0,04	0,36	0,03
5	0,16	0,03	0,27	0,02	5	0,09	0,07	0,21	0,05
6	0,07	0,08	0,17	0,12	6	0,05	0,11	0,15	0,16
7	0,10	0,00	0,10	0,00	7	0,08	0,00	0,08	0,00
8	0,12	0,03	0,31	0,02	8	0,08	0,05	0,28	0,02
9	0,10	0,02	0,31	0,01	9	0,04	0,06	0,27	0,02
10	1,00	0,00	1,00	0,00	10				
11	0,37	0,02	0,56	0,02	11	0,13	0,07	0,40	0,07
12	0,27	0,01	0,48	0,01	12	0,08	0,07	0,35	0,03
13	0,22	0,04	0,44	0,02	13	0,10	0,10	0,36	0,06
14	0,21	0,01	0,45	0,00	14	0,07	0,03	0,37	0,01
15	0,25	0,01	0,50	0,01	15	0,07	0,05	0,38	0,03
16	0,22	0,01	0,39	0,01	16	0,06	0,07	0,27	0,05
17	0,34	0,00	0,52	0,00	17	0,06	0,03	0,32	0,05
18	0,10	0,03	0,24	0,02	18	0,04	0,07	0,20	0,03
19	0,19	0,08	0,31	0,01	19	0,08	0,23	0,22	0,05

5. Conclusions and Future Work

It has been shown that the use of the information submitted to Commercial Registers can be used to reduce respondents burden maintaining, or even decreasing, the size of the variance of the estimates.

Further data analysis needs to be done for several years in order to decide the extent of sample size reduction for each of the industrial sectors as data from the Commercial Registers is available shortly before data for the Industrial Survey is collected. Interesting lessons to optimise the sampling scheme will be extracted as well.

There is need of further research to build an indirect estimation procedures based on sample relationships between economic variables as in the Commercial Register there is no information about all the variables considered in the Industrial Survey.