INTEGRATION OF SURVEY AND ADMINISTRATIVE DATA IN STRUCTURAL SERVICES STATISTICS

Laura Otero Franco, Patxi Garrido Espinosa; EUSTAT-Basque Statistical Office, Spain
E-mail: laura-otero@eustat.eus, patxi-garrido@eustat.eus

1. Introduction

The Basque Statistical Office (EUSTAT) has conducted Structural Services Statistics (SSS) since 1994 based on extensive samples with irregular frequencies and activity coverage due mainly to budget constraints. The growing demand for yearly estimates for National Accounts has been possible to satisfy thanks to a better access to administrative sources in the last years. From 2012, a yearly Structural Services Statistics (SSS) is carried out combining both survey and admin data that covers the whole services sector with a considerable decrease both in respondents burden and allocated budget.

The need of integration of data from different sources (both internal and external) brings new difficulties in all the stages of statistical production that are being handled in different ways. In what follows, we will focus on some of the strategies adopted in EUSTAT to overcome the challenges that the new multi-source yearly Structural Services Statistics (SSS) of the Basque Country has implied.

2. Administrative data sources

In 2003 EUSTAT signed a Cooperation Agreement with the Spanish Commercial or Mercantile Register (CR) aiming to obtain information of all the mercantile companies with economic activity within the Basque Country and has released yearly some statistics based purely on data therein. The experience using this information has encouraged EUSTAT to use the data more widely and include it in the Structural Business Statistics System. Other admin sources like the Basque Registers of Cooperatives, Foundations and Associations are also available (after signing corresponding agreements). In what follows all these external sources will be included in the term CR.

The target population of the Structural Services Statistics (SSS) is made up of all establishments whose main activity is included in the following NACE (rev.2) divisions: 45-47, 52-63, 66-82 and 90-96. The variables of interest are the Profit and Loss Account entries and some other basic economic variables like Employment, Gross Value Added at factor cost or the Net Operating Surplus.

The CR covers all the divisions included in the SSS but no information regarding the self-employed is available and data refer to the company rather than to establishments. These are, together with the uneven coverage by subpopulation in the registers, the main difficulties that have needed to be faced.

3. Sampling Design

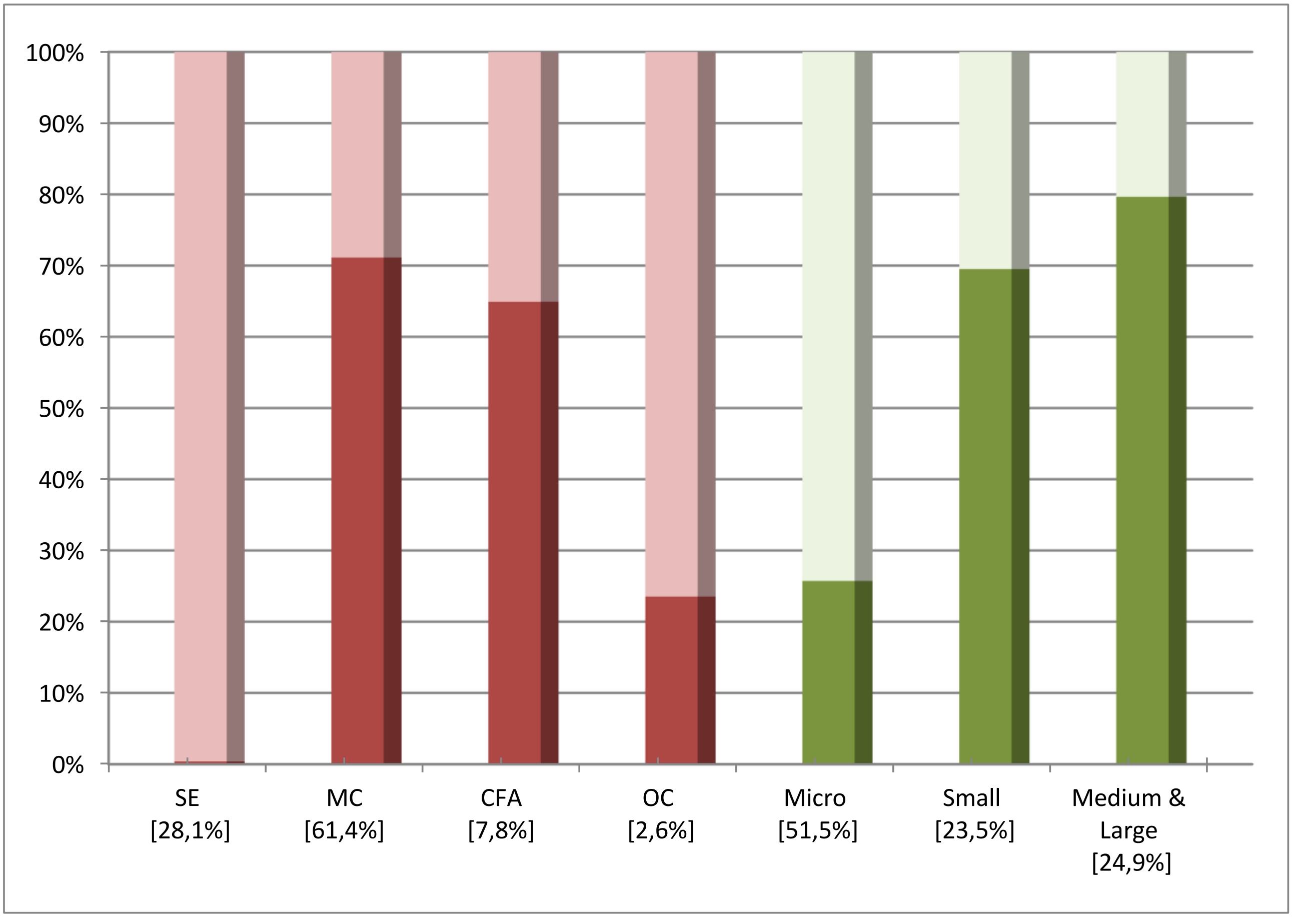
The heterogeneous coverage by subpopulation and, most importantly, the lack of information regarding the self-employed workers is solved by a tailor-made sampling design for under-represented subpopulations. The aim is twofold: to reduce the number of questionnaires (hence respondents burden) and to fulfill the information in those subpopulations not sufficiently represented by admin data.

In Figure 1 the employment coverage in the registers in 2013 by nature (Self-Employed, Mercantile Companies, Cooperatives, Foundations and Associations, and Other Companies) and size of companies is pictured in shadowed. Below, in brackets, the percentage of the employment of the population that each of the groups represents. Overall, almost 50% of all the employment in Services sector is available by admin data, higher than in most standard fully-sample-based surveys. Nevertheless, some sampling is necessary in order to obtain more uniform and representative coverage of all subpopulations.

It is clear that the self-employed are strongly under-represented in available administrative sources (less than 0.5% of the employment covered), followed by Other Companies, with information available only for 23.5%. These two types of companies represent 28.1% and 2.6% of the total employment and there are some specific sectors where the picture would be strongly biased without taking them into account.

By size, micro companies, which amount for over a half of the global employment, are the ones most poorly represented in admin sources, with data covering just over 25%.

Figure 1. Coverage of employment provided by admin sources by nature and size of companies. %



Source: SSS 2013. EUSTAT

By sectors, and using data from previous surveys, a variability analysis of the variable Gross Value Added at factor cost (GVA) has been performed for each of the original sampling strata (size and activity). Significant differences were observed so that the variability in this key variable has been introduced as one of the new sampling criteria.

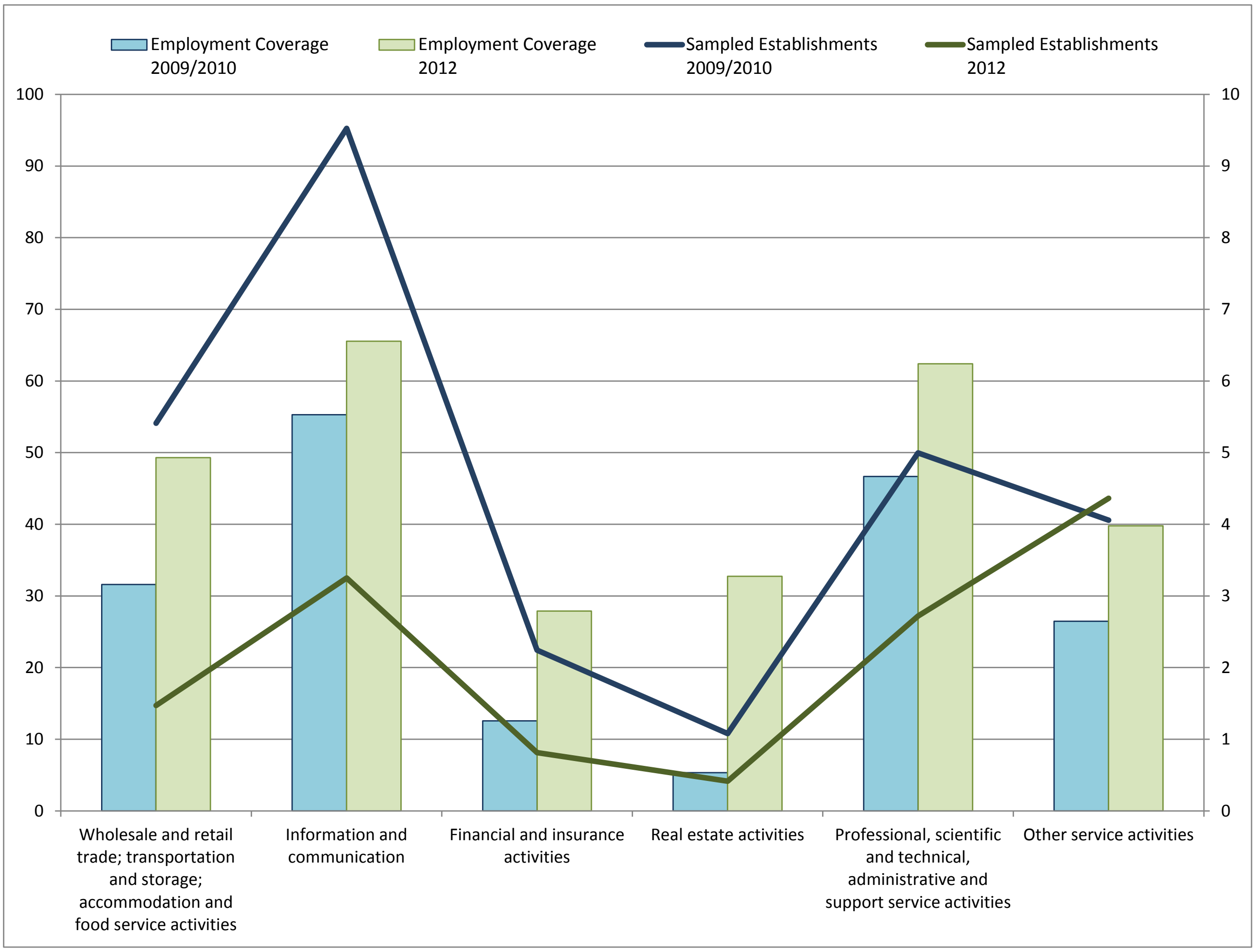
Considering both findings above and the general aims of the survey, a random *ad-hoc* sample has been designed stratified by **legal nature**, **size** and **activity** taking into account the **variability in GVA**. The Directory of Economic Activities of EUSTAT (DIRAE) serves as the framework of reference to select and extract samples.

The sampling unit (and statistical unit) is the establishment, in contrast to data in CR which refers rather to companies. Obtaining estimates at establishment level from information at company level is not straightforward neither trivial. An Indirect Ratio Estimation is applied combining information at establishment level from DIRAE with information at company level from registers. This method requires further research in the near future.

The final sampling design can be briefly described as:

- 50 employees or more: exhaustive
- 20-49 employees: exhaustive for all but mercantile companies
- 10-19 employees: exhaustive for associations, public administration and "other type" companies, and simple random sampling for self-employees with employees according to the variability observed for GVA by activity and the coverage in the admin sources.
- 1-9 employees: simple random sampling for self-employees with employees according to the variability observed for GVA by activity and the coverage in the admin sources for all type of companies but mercantile companies and cooperatives.
- without employees: simple random sampling proportional to the number of establishments by activity.

Figure 2. Employment total coverage and sampled establishments by activity group in SSS 2009/2010 and SSS 2012. %.



Source: SSS 2009/2010 and SSS 2012. EUSTAT

The resulting sample for SSS 2012 (also for SSS 2013), compared to previous fully-sample-based SSS 2009/2010 (Survey on Trade and Repairs 2010 and Survey of Other Services 2009), has implied a considerable saving in terms of number of establishments surveyed together with a significant increase in the coverage of the employment across all sectors. These two clear advantages are shown in Figure 2 where the sampled proportion and the obtained employment coverage are put together for SSS 2012 (in blue) and SSS 2009/2010 (in green) by activity group. The effective saving of total questionnaires has been of almost 60% with an increase of 17 points of the employment covered. This savings are expected to be higher in coming years as the sampling designs adopted in these first surveys have been quite conservative. For instance, those large establishments that year after year are considered to send consistent data to their corresponding register will not be sampled.

4. Coefficients of Variation

One of the immediate consequences of using more information is that estimates are more accurate in these multi-source surveys. In Figure 3 the estimated coefficients of variation of the estimates of variables Turnover and GVA are compared for both surveys (fully-sample-based 2009-2010 and multi-source 2012). In both cases the coefficients are moderate but the gain when using the new methodology is significant in most cases.

Figure 3. Estimated Coefficients of Variation for Turnover and Gross Value Added by activity. %

	Turnover			GVA		
	2009/2010	2012	12/10 (pp)	2009/2010	2012	12/10 (pp)
TOTAL	***3,6***	***1,3***	***-2,3***	***1,2***	***1,1***	***-0,1***
Wholesale and Reatil Trade	5,8	2,0	-3,8	1,6	0,6	-1,0
Transportation and Storage	3,7	5,7	2,0	4,3	12,8	8,5
Accommodation and Food Service Activities	1,7	1,2	-0,5	1,4	0,9	-0,5
Information and Communication	3,5	1,6	-1,9	1,6	1,0	-0,6
Financial and Insurance Activities	2,2	5,2	3,0	3,7	6,1	2,4
Real Estate Activities	11,7	5,4	-6,3	12,0	11,5	-0,5
Professional, Scientific and Technical Activities	3,1	1,9	-1,2	4,1	2,6	-1,5
Administrative and Support Service Activities	4,3	1,3	-3,0	2,2	1,1	-1,1
Arts, Entertainment and Recreation Activities	4,8	2,9	-1,9	4,3	2,9	-1,4
Other Service Activities	3,7	1,6	-2,1	3,2	1,2	-2,0

Source: SSS 2009/2010 and SSS 2012. EUSTAT

It is noticeable that for Transportation and Storage the coefficient of variation is actually higher when using the new multi-source Survey for both variables. This is actually the subsector with the highest increase in the employment coverage with the new methodology (from 11.3% up to 40.5%). In this particular subsector, in previous samples there were several estimation cells that relied just on a single establishment data so that the partial coefficient of variation was null. With the new multi-source approach, on the other hand, all cells have been more densely covered and, therefore, estimation variability added (most likely present in the population).

In general terms, as the variability observed is one of the criteria of the sampling design, the effective coverage gain implies that population heterogeneity is better reflected, which might result in higher resulting coefficients of variation for a number of sectors (those previously least sampled).

5. Estimation of breakdowns

The main variables of interest of the SSS are all included in the information provided by companies to the corresponding registers. However, there are missing breakdowns of the target variables which are of statistics interest. Sending yearly short questionnaires to companies asking for these breakdowns is not feasible when one of the main goals is to reduce respondents burden. Questionnaires have actually been simplified, as missing breakdowns in registers have been excluded from them.

Therefore, it is necessary to define indirect estimators for breakdowns that are considered to be essential for which there is no information from none of the sources. This is the case, for instance, for variable Other Operating Expenses which is divided into External Services, Taxes and Other Current Management Expenses.

After an exhaustive analysis of different possible methods, a Ratio Based Imputation system has been chosen for estimating unknown breakdowns of directly estimated variables using homogeneous units from a previous reference year. Units with same activity and employment group are assumed to be homogeneous. Every five years, coinciding with the extra requirements of the Input-Output Table estimation, larger questionnaires will be sent to a higher number of establishments in order to update the ratio structures.

6. Conclusions

The new SSS shows many advantages over previous services surveys of EUSTAT. It provides yearly estimates for the service sector as a whole for the first time. This is achieved by actually reducing the respondents burden and the cost of the survey. Besides, higher coverage of the population combining information from different sources either reduces the variability of estimates or allows to control possible biases in highly heterogeneous subpopulations.

Within the new estimation system there is still the need of massive samples with high costs both for EUSTAT and for respondents every 5 years that will probably need to be reconsidered in the near future.