Ensuring comparability over time and between domains by means of complex sample techniques

Francesca Inglese and Tiziana Tuoto

Abstract This paper describes the sample strategy defined for the Italian Structure of Earnings Survey of the 2010 year. The aim of the sample design is to allow to obtain estimates for the main variables according to both the NACE Rev.2 classification – that is one of the main domains for the 2010 survey- and the NACE Rev.1.1 one, that was considered in the previous 2006-2002 surveys. The adopted sampling coordination technique faces also with the necessity of contain the overall sample size. In this way, comparability over time and between domains is preserved and budgetary constrains are respected.

1 Introduction

The European Union Structure of Earnings Survey (SES) is a four-yearly large enterprise sample survey providing detailed and comparable information on relationships between the level of remuneration, individual characteristics of employees (sex, age, occupation, length of service, highest educational level attained, etc.) and their employer (economic activity, size and location of the enterprise). Several structural indicators such as the gender pay gap, the low-wage trap or the unemployment trap are based on earnings statistics.

The SES is conducted in the 27 Member States of the European Union and 2 countries of the European Free Trade Association (EFTA) in accordance with Council Regulation n° 530/1999 and the Commission Regulation 1916/2000 as amended by Commission regulation 1738/2005. The objective of this legislation is to provide accurate and harmonised data on earnings in EU Member States, EFTA countries and Candidate Countries for policy-making and research purposes.

National statistical offices collect the information on earnings used in the survey and it contains questions about the enterprise and on the individual employee, aiming to gather individual data on earnings and working hours, as well as personal characteristics and characteristics of the jobs. The national statistical institutes are

responsible for selecting the sample, preparing the questionnaires, conducting the survey and forwarding the results to Eurostat in accordance with the common coding scheme. The data are centrally processed by Eurostat. The data become available approximately 2 years after the end of the reference period.

The statistics of the SES refer to enterprises with at least 10 employees in the areas of economic activity defined by sections C-K of NACE Rev.1.1 for 2002 and 2006. For the reference year 2010, the economic activity is coded using NACE Rev.2

Due to the relevance of the gathered information in particular in this crisis period, and the long time elapsed between subsequent surveys, a further desire for the 2010 Italian survey is to preserve the comparability with the estimates of the previous survey, mainly overcoming the change of the economic activity code. This paper describes the sample procedure adopted in order to gain this comparability over time and between domains. The proposed sample strategy preserves the estimation domains for the year 2010 and constrains the overall number of sampled enterprises, as well.

2 The Sample Design

The Italian SES is traditionally based on a two-stages sample design: primary units are the enterprises while the secondary ones are the employees. All enterprises with at least 250 employees are surely included in the sample. The choice of the two-stages design allows to contain the total number of enterprises involved in the survey.

Both primary and secondary units are stratified, in order to introduce a gain in efficiency. As in most of the economic surveys, the stratification variables are the sectors of economic activity, the size of the enterprise in terms of the number of employees and the NUTS5 classification. The sample allocation is optimised with respect the secondary units (the employees) in the strata defined for the primary units (the enterprises). This kind of stratification produces thousands of strata, as reported in section 3.

In order to ensure comparability between the old and the new NACE codes, it is practically impossible to stratify considering at the same time the two classifications, given that too much strata with too few enterprises would be obtained. So, the main idea is to apply two separate stratifications to the same universe of reference, considering at the first time the NACE Rev.2 – the classification of the current survey - and at the second time the NACE Rev.1.1 – the classification of the previous surveys. Note that all the two economic activity classifications were available on the reference enterprises register. The two stratifications maintain stable the other stratification variables (size in terms of employees and NUTS5).

For each stratum defined on the primary units, the sample size of the secondary units is assigned by means of an optimal procedure for multivariate allocation, based on the Bethel algorithm (Bethel, 1989). It is important to underline that the optimal allocation procedure has been applied two times, one for the NACE Rev.2 - based stratification and one for the NACE Rev.1.1. The optimization algorithm ensures that the expected sampling errors, for the estimates of the main parameters of interest on the planned domains, do not overcome prefixed levels. Moreover, a minimum sample size (al least 200 employees) will be allocated in each stratum.

The algorithm requires averages and standard errors of the main variables: the average worked hours and the hourly earnings have been considered as main variables of interest, while averages and standard errors have been calculated on the basis of

previous survey data (reference year 2006). For the standard errors estimates, approximated formulas have been used, the estimate of sample variability for *srs* (simple random sample) has been increased of both the cluster effect for the secondary units and a further effect taking into account the unequal selection probabilities (absence of the self-weighting condition, Kish 1992) . For the sample size of the enterprises with less than 249 employees, the non-response rates observed in the previous survey have been taken into account as well.

Once obtained the optimal sample size for the secondary units (the employees) both according to the NACE Rev.2 and NACE Rev.1.1 classifications, the resulting sample size of primary units (the enterprises) has been calculated according to the Eurostat recommendations and the previous surveys experiences, as showed in Table 1.

Table 1: *Fixed Number of employees to sample for size of enterprises in terms of employees*

Size in terms of Employees	***Fixed Number of sample employees***
From 10 to 19	all
from 20 to 49	20
from 50 to 99	25
from 100 to 249	35
from 250 to 499	40
from 500 to 999	50
from 1000 to 1999	60
from 2000 to 3999	65
from 4000 to 7499	75
from 7500 to 9999	100
10000 and more	200

2.1 The coordination phase

Within each stratum, equal probability has been assigned to the enterprises, both for classification in NACE Rev.2 and NACE Rev.1.1; then the first business sample has been drawn without replacement according to stratification in NACE Rev.2.

Subsequently, through a sampling coordination procedure, a second business sample has been drawn according to stratification in NACE Rev.1.1.

The sampling coordination has been performed by means of permanent random number (PRN) technique, known as collocated sampling (Ohlsonn, 1995, pp.161-162). The units (enterprises) in the stratum are ordered at random, giving unit i the rank L_i. Independent of this random ordering, a single random number ε is selected from the uniform [0, 1] distribution. For each unit i, define

$$R_i = \frac{L_i - \varepsilon}{N}.$$

The unit i is included in the sample if $R_i \leq \pi_i$ where π_i is the inclusion probability of the unit i .

The method, considering a unique value of ε and assigning to each unit a single random number R_i to be used in the two separate sampling selections, ensures a good overlap of the samples in each stratum and reduces (but not quite eliminate) the sample size variation.

At the end of the business sample selection, employees sample selections are drawn by the sampled enterprises them-self, according to the sample size reported in Table 1, on the basis of sampling rules defined by Istat.

3 Results

The size of the business register used as reference population is about 186000. The number of strata defined on the basis of NACE Rev.2, five classes of size in terms of employees and NUTS5 is 1708, while the number of strata defined according NACE Rev.1.1 and the other stratification variables is 1622.

The number of enterprises selected according to the NACE Rev.2 classification is 18388. Through the sampling coordination technique, the final size of business sample is 19535. This sample will allow to produce estimates on the considered variables, with sampling errors less than 8% for both NACE Rev.2 classification and NACE Rev.1.1 one.

References

1. Bethel J. (1989) "Sample Allocation in Multivariate Survey" Survey Methodology, 15, pp. 47-57
2. Cardinaleschi S. (2008) “Structure of Earnings Survey for the year 2006” Eurostat Quality Report for Italy
3. Kish L. (1992). Weighting for uniqual p_i . In *Journal of Official Statistics*, vol.8 n.2 1992, pp 183-200
4. Ohlsson E. (1995). Coordination of Samples Using Permanent Random Number. In *Business Survey Methods*. A Cura di Cox B. G., Binder D. A., Chinnappa B. N., Christianson A., Colledge M. J., Kott P. S., 153-170. New York, Wiley.