Designing the integration of register and survey data in earning statistics

C. Baldi[1], C. Casciano[1], M. A. Ciarallo[1], M. C. Congia[1], S. De Santis[1], S. Pacini[1]

[1] *National Statistical Institute, Italy*

Abstract
This paper addresses the redesign of the Structure of Earning Survey 2014 made possible by the availability of the new employer-employee register on wages, hours and labour costs. The redesign has aimed on one side to reduce the burden on the enterprises and improve the quality of the data and on the other side to collect the data not available in the register and to correct the definition of those available. Three main areas will be covered: the survey sampling, the questionnaire redesign and the development of regression models to mass-impute and correct the variables. As for the sampling the availability of the register information has allowed to dramatically reduce the sample size without lowering the target quality measures, while at the same time producing a survey structure more adequate to distributional analysis and micro econometrics. The complexity of the questionnaire has been reduced significantly and its content redefined in a way to be a bridge between the administrative and statistical definitions. Moreover the prefilling of key data on wages and hours should reduce to a large extent the measurement errors. Finally, an in depth analysis of the content of the Social Security data on which the register is based is allowing to specify a new class of regression models to impute the missing data on the register.

Keywords: Response burden, Register-based statistics, Micro integration.

1. Toward a system of integrated data in earnings statistics

Up to recent years the earning statistics in Italy were based on the 4-yearly Structure of Earning Survey (SES), required by the Regulation CE n. 530/99. No yearly statistics were released except for the Gender Pay Gap (GPG) obtained updating the SES figures with yearly information from EU-SILC.

In the last years, thanks to the availability of new statistical registers based on administrative data, a new integrated system of statistics on earnings has been designed to produce coherent

official statistics on earnings and hours paid/worked. The aim is to release annual earnings differentials indicators broken down for the characteristics of jobs, enterprises and employees, and the annual statistics of the Gender Pay Gap (GPG) based on earning per hour paid.

The 4-yearly release of microdata on employees for the SES Regulation, as well as the 4-yearly release for Labour Cost Regulation (LCS) have been involved in this system.

Starting from the innovative experience in integrating survey and register data carried out at firm level for the LCS 2012 (Baldi et al., 2014), the SES 2104 has been planned going beyond a classical micro-integration (Bakker, 2010; Zhang, 2012).

Following the availability of the new statistical register on wages (RACLI) with data on the entire population of employees of the private sectors referred to 2014, a new type of integration has been experimented introducing several solutions ex-ante, as a new sampling method and the pre-filling of the questionnaire which has also been completely redesigned.

The approach is for a "circular" integration, with the register data that assists the survey in the first phases of the process, and the survey that has been planned in order to check the register data and support the harmonization of some variables.

1.1 The new register availability and the new role of the survey

The recent developments in the availability of more detailed administrative data on employment and labor costs from Social Security sources has allowed the production of new statistical registers. Firstly, the Employment Register has been created using the individual Social Security declarations remitted monthly by firms since 2010, for each of their employee. It contains information on characteristics of the employee, the job and the firm, with a LEED structure, and it is also the micro base of the Business Register. The new RACLI wage register is an extension of the Employment Register to a set of variables on wages and paid time at employee level, covering the entire population of employees and firms of the private sector.

After an in depth study of the administrative definitions and their relations to the statistical ones, the variables of the statistical register RACLI has been derived from the administrative data. They have different status referring to their statistical usability. Focusing on the

requirements of the SES Regulation, some RACLI variables adequately fulfill the SES statistical definitions (statistical variables) while others represent 'proxy' measures that need to be harmonized as the number of hours paid which do not include overtime hours (proxy variables). Some other variables required by SES are not at all available in the register.

So the survey was unavoidable first of all for the unavailable variables and secondly to ask for information necessary to harmonize register proxy variable to statistical target concepts. Although necessary, the survey has been completely redesigned in order to be as much complementary as possible to the register taking into account available data and to build a bridge between register data and EU Regulation on definitional issues.

The "one variable collected only once" principle inspired the new survey design, assisted ex-ante by register data, reducing to a minimum if not eliminating all the need to reconcile similar variables collected in more than one source (Bakker 2010).

Also the sampling strategy has been completely changed due to the Register availability.

2. The survey sampling

2.1. The old sampling design

Up to now the SES Survey sample was based on a two stage design. In the first stage a sample of enterprises with at least 10 employees, were selected from the business register. All enterprises with more than 250 employees were included, while a random sample of those under this threshold in strata defined by the combination of 2 digit NACE, size class and NUTS1 has been drawn. The number of units to be selected in each stratum was defined as an optimum of a multivariate allocation procedure (Bethel, 1989). The sampling sizes, in terms of enterprises, was increased considering the response rate recorded in previous SES to prevent the negative effect of total non-response, and finally it included 22,000 enterprises.

The sampling size at second stage, that is the number of employees per enterprise, was not computed optimally for lack of information. Instead, the final number of sampled employees was established according to a predetermined number varying with the size class of the

enterprise (from a minimum of 10 to 200 employees). The selection was left to the enterprises according to a systematic sample scheme proposed by Istat in order to guarantee the randomization of the sample. The final number of employees was 480,000.

2.2. Multifold purposes of the new sampling design: pros and cons of one stage sampling

The large sample size at second stage was the main drawback of old SES sampling design. A second drawback derived from the impossibility of stratifying the sample in terms of characteristics of job/workers. This implied a lack of efficiency and possibly introduced bias in the sample through the selection operated by the enterprises. The availability of new data sources encouraged to optimize the integrated use of different sources. The sample should be enough large to produce efficient estimates for means and distributional parameters; at the same time, the ideal sampling design must ensure the representativeness of the planned sample with respect to many categories of workers, with a sample size able to provide a good set of microdata for modelling and for econometric analysis. Finally, the data collected via survey should be enough "robust" for the validation and updating of RACLI variables/records.

The employee level register constitutes a framework to test different sampling designs. 1) It makes possible to build strata based on job/worker level characteristics, where a sample of individuals can be directly selected. 2) The availability of target variables at individual level to drive the allocation procedure. 3) Known population totals of some core variables allow to derive measures of bias for sample estimates since the designing stage. 4) In the estimation stage, register variables can be used to impute the non-responses and/or calibrate the weights.

A one-stage stratified random sampling with strata defined by combinations of enterprise level variables with job and personal characteristics has been firstly experimented for the new sampling strategy. A one-stage sampling design would be much more efficient than the traditional two-stage, as it makes possible to achieve the same levels of accuracy of estimates with a much smaller final sample size.

To ensure the inclusion of all largest enterprises, each enterprise with over 250 employees formed an individual strata. The optimal allocation ensured that a sample of 220,000 jobs

produced estimates of the target variables with a sampling coefficient of variation much lower than the old sampling strategy, even in detailed domains (Guenther, 2014 for German SES).

However, the number of selected enterprises, being random, increased rapidly over 40-45 thousand units, too much than the reputed "politically" acceptable and technically feasible. In order to limit the first-stage sample size, a random selection with probability proportional to the enterprise dimension has been tested. This approach was efficient in reducing the number of final sampled enterprises, which decreased to 26,000 units. But enterprises of similar dimension still showed too different sampling rates; moreover, a significant increase of the expected sampling errors for the final estimates even on planned domains occurred. Finally, a comparison of the sample estimates with the totals known from the register highlighted that the estimates might be biased, particularly for quantiles, as this strategy tends to oversample larger enterprises, whose workers show socio-economic peculiarities.

2.3. The solution adopted: a two stage approach with sampling allocation in either stages

Given these drawbacks, the chosen design for new SES survey is a two-stage cluster sampling.

In the first stage, the clusters are the enterprises, stratified by crossing of 2 digit Nace, size class and NUTS1; the allocation has been optimally determined by constraining the precision of the estimates of target variables hourly monthly earnings and annual earnings to a maximum CV of 3.6%. The sampled enterprises are 24,465 out of the 168,290.

In the second stage, the employees were selected in strata built by splitting up each enterprise into groups defined by socio-economic characteristics of employees. The information available in the register would have allowed a very detailed stratification. However, a cross-classification according to the modalities of Nuts1 localization of worker, working time, Broad Occupation has been considered a relatively fragmented partition of reasonably homogeneous strata. The optimal sample size at strata level has been determined under the constraints for a maximum planned CV of 1.6% for the target variables on the estimation domains. The overall second stage sample size has been of 212,722 sampled employees out of 10,546,683. A further adjustment step has been performed by forcing the final sampling size to vary from a minimum of 3 to a maximum of 500 of workers per enterprise.

3. The questionnaire redesign

3.1. The old way of designing a questionnaire: an output driven design

In the past editions of the SES Survey, the questionnaire was designed in the traditional way, that is starting from desired concepts and definitions set up by the SES Regulations. The aim of the questionnaire and its instructions was, as usual, to operationalize the definitions with an appropriate wording for the Italian enterprises. Considering that up-to-date suitable social security and fiscal sources for the private sector were not available it was necessary to acquire all target variables directly from respondents in order to produce the information requested.

In particular, since the regulation asked for employee (e.g. sex, age, education) and job (e.g. occupation, contractual working time full-time or part-time, share of a full-timer's normal hours; earnings, working time for both the reference month and year) characteristics for a sample of employees, the main section of the questionnaire was modular with a module for each sampled employee. Moreover the survey had a firm level section which was used to collect data at firm level on labour cost variables together with firm level information required by the regulation (e.g. such as economic activity, size of the enterprise, number of employees geographical location, referring local units). The inclusion of a firm level section to the modular one made the questionnaire additionally burdensome (probably the most among business surveys) and highly complex. This aspect together with the large number of sampled employees and the low motivation for respondents resulted in low response rates.

The availability of the new RACLI employer-employee register on wages and hours has pushed to question the traditional data collecting forms introducing new ones based on a different culture of questionnaire development, evaluation and testing aimed at reducing the burden on enterprises, improving data quality and making an optimal use of admin data.

3.2. The new way: bridging administrative data with requested outputs and the pre-filling

The previous principles have been implemented designing the new questionnaire starting from the availability of administrative data and their definitions. Statistical variables available in the Register, whose definition complies with the statistical ones, have been removed from the

questionnaire or prefilled for checking purposes. Geographical location, economic activity, enterprise size, employee sex and age, paid weeks are among the variables not included in the new questionnaire while contractual working time and share of full-time hours are prefilled.

Proxy statistical variables available in the Register, whose definition is close but not enough to the statistical one, have been prefilled and/or items necessary to adjust definitions are asked to the respondents. So the questionnaire becomes a bridge between administrative and statistical definitions: monthly register wage is used as pivot to measure the monthly earning requested by the regulation, and contractual working hours as pivot to measure monthly paid hours.

Finally the variables required by the Regulation not available in the Register, such as overtime hours and earnings and special payments for shift works, are added to the questionnaire. Some of these variables are also going to be used to integrate the Register.

The pre-filling not only allows to build questions that bridge available data with the information required, but it provides solid benchmarks to detect errors in the sub-items. Not all cases have been prefilled due to uncertainty in the measurement in the register.

So far, as regard monthly data, the proportion of administrative records with prefilled-variables is around 85% referring to earnings and 75% referring to working hours. The enterprises are given the possibility to modify the prefilled information: around 85% of respondents have confirmed the prefilled earning data, about 10% have reduced the data and more or less of 6% have increased them.

4. The development of regression models to mass-impute and correct the variables

We have seen that RACLI provides only a proxy for the number of hours paid, one of the main variables in the System of Earning Statistics. The variable estimated through RACLI can be defined as the number of hours paid without overtime hours. Although the share of overtime hours is usually a small part of the number of hours paid – around 2-3% on average - its absence make the variable provided by the register slightly downward biased. So the aim is to impute overtime hours in RACLI achieving the variables required by the Regulation.

One of the purpose of the SES survey is providing data on the number of overtime hours.

Another source of information that provides an estimate of overtime hours is the GI-VELA which is a combination of the monthly survey on large enterprises (GI) and the quarterly survey on job vacancies and hours worked (VELA). The first covers, with a take all sample, the enterprises with at least 500 employees and the second, with a SRS, all the enterprises between 10 and 500. GI-VELA produces quarterly estimates of the number of overtime hours and provides monthly information on the number of overtime hours for each GI enterprises.

The advantage of these estimates is that they come from short term surveys and thus provide information also in intra-SES periods. The SES instead is a four yearly survey but with the advantage of providing information at job level. How these two pieces of information can be combined to estimate (at least) yearly information of overtime hours, and thus the regulation defined number of hours paid at job level?

A choice is to estimate directly the average number of overtime hours in SES years on classes of jobs defined by (a subset of) worker, job and enterprises characteristics (such as sex, age, occupational qualifications, type of contract, Nace sector, size class, etc..). These figures, obtained either by direct estimates of the totals of the SES data or through a regression of overtime hours on dummy variables representing the levels of the qualitative variables (and possibly their interactions), may then be used to calculate the share of overtime hours by groups, to be applied to the GI-VELA known totals.

This method has a couple of drawbacks. First, by calculating totals it underestimate the heterogeneity within the group and in particular no one individual will have zero overtime hours (in a group with a positive number of overtime hours), which, instead, is a very common characteristic of this variable distribution. Second, it neglects the use of overtime earnings, collected in the SES survey, and that can be possibly used in the estimates.

A refinement of the above method, more complex but hopefully more promising, is the estimate of overtime hours passing through the estimate of overtime earnings. The idea exploits the use of a register variable that is a proxy of the non-regular earnings highly correlated with the overtime earnings.

This second method can be described as follows. 1) Predict, for each job in the register, the overtime earnings by regressing the SES overtime earnings on the Register proxy of non regular earnings and other job, worker and enterprise variables. The regression might have enterprises fixed (or random) effect, exploiting the fact that, normally, multiple jobs for enterprise are requested by the sampling design. More generally the regression might result from a multilevel model. 2) Predict, for each job in the register, the number of overtime hours, by dividing the figures predicted in step 1 with the SES estimates of the hourly overtime earning. These should be calculated in groups as fine as possible. 3) Estimate the number of over-time hours at group level using the GI-VELA data. The groups might be built as Nace and size for the Small and Medium enterprises and the single enterprise for the Large ones. 4) Calculate for each job in the register the share of overtime hours over the group, as defined in step 3, to which it belongs. 5) Redistribute the number of overtime hours, estimated in step 3, across the jobs with the shares calculated in step 4.

This method can be applied in SES years. In intra SES years, the regression coefficients and the hourly overtime earnings, that is the parameters estimated in Step 1 and 2, will have to be used together with the current year register values and the VELA-GI known totals.

The passage through the overtime earnings has some advantages in theory. First, it should allow a better prediction of zero overtime, since when the proxy of non regular payments is zero by definition there is no space for overtime earnings. Second, being the proxy of non regular payments a continuous variable differentiated for each job in the register, it should allow to account for more heterogeneity in overtime. Third, being this proxy more variable over time than the characteristics of workers, jobs and enterprise, it should provide more up-to date signals of the changes of overtime hours.

The quality of the prediction depends substantially on the quality of regression of step 1. However this value is likely to be influenced by the huge tower of zeroes corresponding to zero overtime (for more than 80% of worker with positive proxy there has been reported no overtime). This issue must be coped with a different class of models. In the next future

mixture models, that accounts for zero inflations, will be analyzed to discriminate between zero and positive overtime.

5. Final remarks

The survey is still ongoing so only provisional remarks on the innovations introduced might be provided. The reduced number of variable required thanks to their availability in the register lowered the questionnaire burden and the pre-filling reduced measurement errors. In general the wide use of the register data in almost every phase of the production process had benefits in terms of monetary costs, reduction of time and sampling and non-sampling errors.

We are now involved in the challenge of improving the quality of some register variables for the entire target population through the integration of survey data and other sources available.

6. References

Baldi C., Ciarallo M. A., De Santis S. and Pacini S. (2014), The converging pattern between Business statistics and Administrative data. Towards an "industrialized" statistical production process, European Conference on Quality in Official Statistics (Q2014), Vienna, 3-5 June.

Bakker B. (2010), Micro-integration: State of the Art, Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses, Working paper 10, 12 May 2010 (The Hague, The Netherlands, 10-11 May 2010)

Bethel J. (1989), Sample allocation in multivariate surveys, Survey methodology, 15:47-57.

Joachimiak W. and Guenther R. (2014), Using administrative data for SES purposes– Germany, European Commision-Eurostat-Working Group on Labour Market Statistics. Doc.: Eurostat/F3/LAMAS/23/14.

Zhang L.-C. (2012), Topics of statistical theory for register-based statistics and data integration, Statistica Neerlandica (2012) Vol. 66, nr. 1, pp. 41–63