A prediction approach for the estimation of hours worked using integrated register and survey data

Fabiana Rocci [1], Silvia Pacini [2], Laura Serbassi [3], Marina Sorrentino [4], Maria Carla Congia [5]

Abstract[6]

Istat has released, starting from the reference year 2014, a new estimate of the variable 'hours worked per employee' to complete the integrated system of estimates FRAME, designed to meet the European regulation SBS. The result is an estimate of hours worked for all businesses with employees in industry and services. The methodology is based on joint information from statistical registers and business surveys, both structural and short term. The used method is based on a predictive approach that estimates on observed data the relationship between hours worked and hours paid, as register auxiliary variable, and imputes the same relationship on the rest of the population. Through the integrated use of information from various sources, and in particular those from the census register, it has been possible to identify subpopulations of companies with significant characteristics from the input of work and the relationship being valued. This paper describes the results of the analysis and outline of the final methodology.

Keywords: hours worked, prediction approach, data integration.

1 Istat, email: rocci@istat.it
2 Istat, email: pacini@istat.it
3 Istat, email: laserbas@istat.it
4 Istat, e-mail: mrsorren@istat.it.
5 Istat, e-mail: congia@istat.it.
6 The authors thank Fabrizio Solari as support for the model specification. Although the paper is the result of the combined work of the authors, the Sections are to be awarded as follows: Sections 1 and 4 to Fabiana Rocci; the introduction to Section 2 and Sections 2.1 and 3.1 to Marina Sorrentino; Sections 2.2 and 2.3 to Silvia Pacini and M.Carla Congia; Section 3.2 to M.Carla Congia; Section 3.3 to Silvia Pacini; Sections 3.4 and 5.1 to Laura Serbassi. Section 5.2 to all the authors.

1. Introduction

Structural business statistics (SBS) aim at describing the structure, behaviour and performance of businesses across the European Union on a yearly basis,.

The Italian National Statistical Institute (Istat) has developed a new system for the production of estimates on economic accounts of businesses for the SBS regulation (Reg. EU 58/1997 e 295/2008). The system is based on the use of administrative and fiscal data as primary source of information (at firm level), integrated with direct survey data as complementary information on specific variables or businesses' sub-populations, for which administrative information is not available. Thus, a new system of a multidimensional set of estimates (FRAME in the following) at an extremely refined level of detail can be annually released (AA.VV. 2016). The FRAME system is made up of several components, the main one being a statistical register of a number of key SBS variables that are available at firm level for the overall target population (~4,4 million units), which is linked to the Business Register Asia (BR).

For the SBS variables that are not covered by the administrative sources, further quality gains are achieved by improving the direct sample surveys and/or by developing alternative estimation strategies exploiting as much as possible the increased amount of available auxiliary information.

This paper focuses on the new estimation of the variable of hours worked by employees that belongs to the set of variables required by the SBS Regulation. It represents the most suitable measure for quantifying the real use of labour in the income production process. The statistical measurement of this variable has always been challenging, because there is not a unique way of registering it into the business accounts system and neither there is a variable defined properly in the administrative data.

A new methodology has been introduced to deliver the estimates of the hours worked for the SBS regulation, starting from the reference year 2014. Many studies on the available information, both direct and correlated to the target variable, have been done to identify the eventual statistical structure underlining all the data and the proper model representing it.

Several Istat surveys, designed to satisfy different EU regulations, comprehend the hours worked among their target variables. Some of them are primarily designed to gather high quality information about this variable, but they can differ with respect to coverage and level of disaggregation. On the other side, some administrative sources contain information about the remunerated days, that properly elaborated within a statistical register can produce a proxy of the paid hours, that is a variable very correlated to the target one.

The variety of information at disposal has made it necessary to adopt a mixed-sources statistical process as a solution, merging them into a physically-consistent dataset and rendering them suitable for a model application across all the population enterprises.

The final integrated dataset would deliver all the available information, from both administrative and surveys sources: the target variable hours worked measure for the surveys observed units and the administrative hours paid data on all the population units that represents auxiliary information correlated to the target one.

In this way, all the available information about hours worked and about the correlated hours paid on the whole target population is organized.

The prediction theory addresses the problem to estimate the total of a finite population variable from a sample to be equivalent to predict the total of the non-sample values (Valliant et al., 2000). In these terms, the prediction approach means to model the relationship between the two variables on the observed units and to impute the same relationship on the unobserved units on the basis of the model based estimation results.

In this context, the basic assumption is that there is a strong relation between hours worked and hours paid, so building a consistent integrated dataset of all related variables would create the suitable dataset to perform the estimation model representing the relationship properly.

The whole estimation scheme has been designed in several steps, to guarantee at each stage the required coherence between the observed units and the underlying concepts among the sources at disposal. In particular, to define the proper classification in specific sub-populations of the so obtained set of observed units has been challenging. The usual variables of stratification have been taken into consideration; nevertheless many other aspects to profile the enterprises structure have been defined. Furthermore, also other peculiar issues suggested by subject matter experts have been deeply analysed. Many aspects proved to need to be constantly monitored, in order to define the proper model stratification and the best model specification.

This paper aims at describing the new methodology used for the SBS estimates of hours worked and focuses on the main aspects that have resulted to be important to profile the enterprises' structure, to better estimate the target parameter.

The results indicate important differences between the new estimates of hours worked with respect to the previous ones based on the direct surveys that are designed to satisfy the SBS population.

The description of the statistical process of every source and the several aspects of the final estimation scheme are useful to explain the reasons of the differences and to give a final assessment of the entire mixed-source estimation process.

2. Informative context

Statistics on working time are needed to construct economic indicators, such as the average hourly earnings, the average labour cost per unit of time and labour productivity, and to evaluate policies and programs, as well as to estimate time-related underemployment. For these purposes statistics on hours worked need to refer to the same reference period and cover the same group of workers as are covered in statistics of, e.g., earnings, labour cost, employment-related income and production.

Many different working time arrangements exist, due to different scheduling of the hours of work that can be combined on a daily, weekly or monthly basis. Working schedules other than regular full-time, such as night work, shift work, part-time work and flexible working time arrangements are very frequently connected to certain economic activities and are also intended to enable workers to achieve a balance of work and family life (ILO, 2005).

For these reasons, the accurate measurement of the hours worked involved in the production is complicated, because it can depend on many different aspects. What is usually measured from the enterprise side are the component of regular hours paid by the business, both for actual hours worked and for paid but not worked hours.

The usual components of paid hours are:

1. Normal hours worked
2. Overtime
3. Not worked but paid hours: vacations, permits, etc.

Several EU regulations on business statistics require the measurement of hours worked. Nevertheless, they can refer to different population coverages, timeliness and details of the variable components.

In order to identify information useful to perform the estimations of hours worked for the SBS regulation all the available sources have been studied, both the surveys and the administrative data.

Since the main goal is to model the relationship between the hours worked and hours paid, the definition of the two variables and their differences have been widely studied across the sources, in order to fix the rules to build a coherent set of units and variables, into an integrated dataset. The relation has resulted to be influenced not only by the usually considered variables, that is by size and economic activity of the enterprise, but also by many aspects related to the internal organisation, as the share of peculiar work contracts, the use of short-time working (STW)[7] etc. Indeed, the chance to describe in a very detailed way the events and the actual input of

7 Short time working occurs when employees are laid off for a number of contractual days each week, or for a number of hours during a working day.

work, through the information of the register on all the population units, has made it possible to test several hypotheses that resulted to be significant to the type of relationship.

In the following, all the sources are described and compared from the coverage and the process point of view, so as to have the full picture on both definitions and measurement aspects. Many features of the surveys' design and processes have been studied (the survey design, the editing and imputation procedures, the calibration methods and the specific operational arrangements), in order to allow an assessment of eventual source effects in the measurement of the final variables (see § 3.3). Furthermore, also specific issues that could influence the target relationship are considered.

2.1 Survey data

Four Istat business surveys collect data on hours worked. Two of them are annual and are aimed at producing the estimation of the number of hours worked by employees in the reference year for the SBS EU Regulation: PMI and SCI, respectively covering enterprises with less and more than 100 persons employed, classified in Nace Rev. 2 sections from B to S with the exclusion of K and O. The two other surveys are short-term and have the production of indicators on per capita hours worked among their main targets: GI, a monthly survey covering enterprises with at least 500 employees classified in Nace Rev. 2 sections from B to S with the exclusion of O, and VELA, a quarterly survey covering enterprises with 10 to 499 employees in the same economic activities. GI and VELA microdata are used jointly to produce quarterly indicators of hours worked for both national publication and the STS EU Regulation (Regulation EC n. 1165/1998 of the Council and its revisions and amendments).

The surveys statistical design and processes are different in several aspects. Starting from the sample design: PMI and VELA are sample surveys, while SCI and GI are censuses of their enterprises' target population.

About the measurement of the variables: in both PMI and SCI employees are defined as the annual average of the end of month stock for each month of the reference year. In both surveys, this measure includes managers.

PMI and SCI measure hours worked as the total number of hours worked by all the enterprise's employees (managers included) in the reference year. Hours worked include both normal time and overtime and the sum of the two components is measured as a unique variable. Furthermore, SCI also collects data on the total number of hours paid during the reference year by all the enterprise employees (managers included).

GI collects monthly data on employees, hours worked by employees (distinguishing between normal time and overtime), hours paid but not worked, wages and employers' social contributions. Data on employees include managers while those on hours and labour costs do not cover them. Employees are observed as the stock at the end of the reference and previous months. The number of employees who are not managers can be calculated by subtracting the number of managers from the total number of employees.

VELA collects quarterly data on employees, hours worked by employees (distinguishing between normal time and overtime) and hours paid but not worked. All data cover only employees who are not managers. Employees are observed as the stock at the end of the reference and previous quarters.

Based on GI and VELA data, total hours paid can be calculated as the sum of total hours worked (including normal time and overtime) and hours paid but not worked. This sum supplies the correct number of hours paid provided that no compensatory time or "time off in lieu" scheme is used in the enterprise[8].

The PMI and SCI surveys provide variables that describe the whole scheme of accounts and balance sheet of the enterprise. To this aim, the process of validation is standardised to achieve a coherent full set of information. Hence, there is not a specific editing and imputation process related to each specific variable, besides those that guarantee the given relation among them. Therefore the procedures to edit, impute and validate the microdata on hours worked aim mainly at obtaining a plausible average estimate with regards to the whole set of account balance rules. In the PMI survey, item non responses are imputed through the mean over respondent units, unit non-responses are imputed through the use of administrative data, that covers almost the full balance sheet, for which the hours worked are treated as item non responses. Being PMI a sample survey, a calibration to the known totals of employment and number of enterprises in the population is then carried out. In the SCI survey, the unit non-responses are treated as in PMI.

On the other hand, data on hours worked are treated with specific attention in both GI and VELA surveys, due to their relevance in the disseminated aggregate indicators. In particular, enterprise experts first check GI data of the responding units. In case of non-responses, normal time and overtime hours worked are imputed separately, using data of the longitudinal profile of the firm itself. Influential observations are then identified and checked by the experts (Rocci and Serbassi, 2008).

Also within the VELA production process normal time and overtime hours worked are checked and validated separately. The first checks are performed during

8 These schemes allow employees to compensate a longer than normal working time in a given period with less hours of work in another one. However, paid hours refer to the normal working time. Hence, in these cases, the sum of hours worked and hours paid but not worked is higher than the actual number of hours paid when the employees work longer than normal hours and lower when they work shorter than normal hours

data collection: the largest share of responses are obtained via CATI and are in this process subjected to an extensive set of plausibility controls on hours worked. Moreover, the validation procedures on collected data include both interactive checks of outliers and influential observations, carried out by subject matter experts, and automated editing and imputation procedures based on the per capita number of hours worked in the same enterprise in the same quarter of the previous year (taking into account working days changes), wherever this information is available, and hot deck nearest neighbour donations, in the remaining cases.

To be used in the estimation procedure of SBS hours worked, GI and VELA higher frequency data need to be annualized.

To this aim, the monthly data collected by GI are transformed into quarterly ones at the enterprise level, by summing monthly hours worked over the quarter and by measuring quarterly employees (managers excluded) as the average of the stocks at the end of the previous quarter and of the reference quarter. In this way, the definitions of average quarterly employees and quarterly hours worked are identical in the quarterly GI and VELA microdata.

Starting from quarterly GI and VELA microdata, annual data are calculated by summing quarterly hours worked and averaging quarterly employees across the four quarters of a year. This step requires the availability of microdata for each considered enterprise for all four quarters of the year. GI microdata satisfy this condition: all unit non-responses are imputed to produce the target monthly indicators of the survey (on jobs, hours worked, wages and labour costs). These imputed unit non-responses are used in the production of the quarterly indicators on hours worked. VELA microdata, on the other hand, are affected by wave non-responses for which a correction via calibration is carried out in the process aimed at the production of quarterly hours worked indicators. Therefore, for the estimation of SBS hours worked, an imputation procedure for per capita hours worked and employees, based on hot deck nearest neighbour donations, is used to compensate for VELA wave non responses.

In the following, the results of several assessment analyses are described. The importance of hours worked in the Short Term surveys plays a fundamental role in the choice of the rules to be followed to build the integrated dataset.

2.2 Register data

The administrative information on remunerated time, that has been used to build the integrated data set, is produced within the Statistical Register on Wages, Hours and Labour Cost at job level (hereafter RACLI).

The RACLI register is mainly based on social security data. It belongs to a system of registers, where it represents the extension of the Employment Register

[Istat, 2016] for variables related to wages, labour cost and labour input for all the employees of the enterprises in the private sector, agriculture excluded (with a sectorial coverage wider than that required by SBS Regulation). Both registers have a Linked Employer Employee Data (LEED) structure, based on the compulsory monthly information at job level that employers have to send to the National Social Security Institute. This implies the availability at enterprise level of details on the jobs and on the employees. The employment estimates obtained through the Employment Register are the source of BR Asia and they are the same used in FRAME. The RACLI variables at enterprise level are obtained by the summarization of the same variables' estimates at job level.

In this paper the focus is on the information related to labour input and to the employment characteristics available in RACLI, such as the type and the duration of the contract, the working time, etc. that makes the register a very rich and detailed source, very useful to delineate which enterprises' features mostly affect working time.

The evolution of the informative contents of the social security source on working time has led to the availability of new and more detailed data that have been used to improve the estimation method of hours paid.

At the moment, the estimates on labour input variables are produced in RACLI exploiting the administrative information available.

Information on labour input have to be declared to the social security system in different units of measure, depending on the type of employee contract. For full-time employees, in particular, enterprises have to declare the monthly paid days, while for the other categories the paid time is registered in terms of hours. This leads to derive a proxy variable of annual hours paid by each enterprise that is calculated using weekly and monthly information at job level.

For full-time employees information on the monthly paid days is declared, according to a standard social security calendar[9]. The number of hours paid have been derived multiplying the number of declared paid days by the contractual working time of the job established in the collective labour agreements. It is important to stress that for administrative aims, one remunerated hour in the day is sufficient to have an entire paid day declared, where paid means totally or partially at the expense of the employer. This has two main implications: i) a paid day may include hours not at all paid by the employer (i.e. hours of strike, etc.); ii) paid days can be both partially and totally remunerated by the employers (i.e. if the employee is sick and this day is partially paid by the employer and partially paid by the social security system it is declared as an entire paid day). The effect could be an overestimation of paid hours due to hours of absence not paid within paid days and days of absence partially remunerated by the employer.

9 The standard of the social security calendar is 26 days in each month, 312 days and 52 weeks in a year.

For part-time and job-on-call employees, employers have to declare the monthly numbers of hours paid in term of equivalent paid weeks of a full-time employee. Using the contractual time of the full-time employee, it is possible to derive the number of hours paid that should not have the measuring problems just described for the paid days of the full-time employee.

Nevertheless, it is evident that both the information on paid time and the proxy variable of hours paid derived on a contractual basis do not include overtime hours. This can cause a bias in the estimates of the levels of hours paid, that however does not prevent its use as auxiliary variable for hours worked.

The richness of all RACLI information can be very useful to distinguish different enterprises' structure and to study the effect on the relationship between hours worked and hours paid.

In this view, it is important to underline the possibility to identify clearly the jobs with contracts that have peculiar working time arrangements, like job-on-call, that characterize the labour input within enterprises that use them extensively. In particular, these types of jobs in general tend to reduce per capita hours worked and to affect the relationship with hours paid (Congia and Pacini, 2010).

Furthermore, very useful information is also available about the large use of short-time working schemes by some firms in many economic sectors during the economic crises (Congia and Pacini, 2012), that is expected to affect the relationship under study.

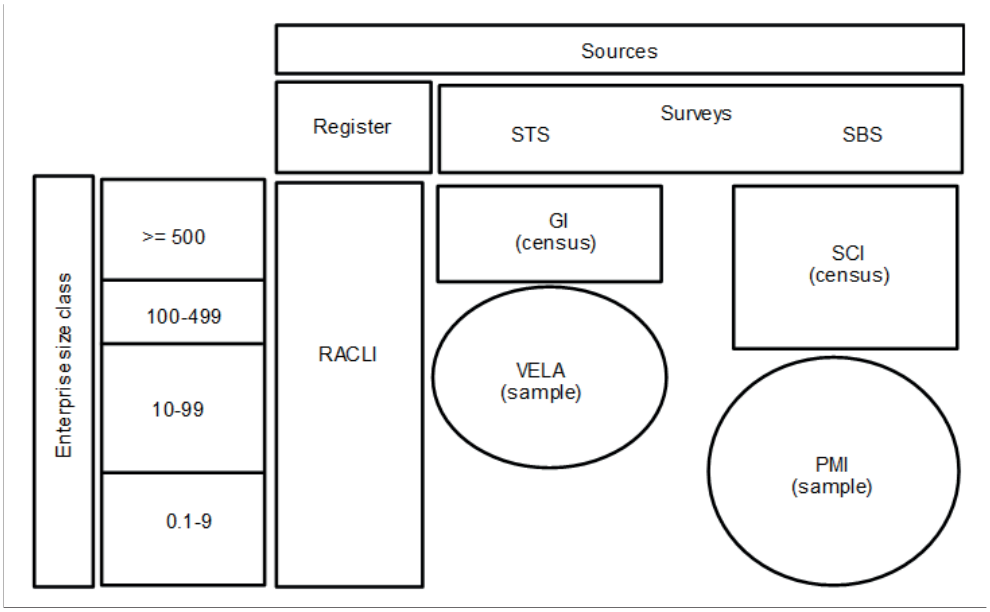
All the information on such employment features has been very useful to identify specific groups of enterprises according to the characteristics of their jobs and to specific events which may concern employees' working time.

In the following sections, the use of such a wide set of information and the way in which the proxy of hours paid has been employed as auxiliary variable to estimate hours worked are illustrated.

2.3 An outline of the data sources coverage

There are many different kinds of information related to the time of work, so all the sources described above have been combined to evaluate their main characteristics. Figure 2.1 represents the available data in terms of coverage and target population.

Figure 2.1 – Sources available on hours paid/worked and their coverage of enterprises with at least one employee in the private sector



Source: Authors' depiction

Furthermore, a summary of the coverage in terms of variables and their components is shown in Table 2.1.

From this point of view, the proxy variable on hours paid from RACLI is available for each unit of the target population. On the other side, the information on hours worked variables from the surveys covers only parts of the population. The big enterprises are almost fully covered, because censuses are run over enterprises with at least 100 employees (SCI, for units with at least 100 employees included, and GI, for those with at least 500 employees included), the response rate are very high and all the unit no-responses are imputed.

Moreover, several samples on the enterprises with less than 100 employees are available from PMI and VELA. Since not all units in the theoretical sample are respondents, the actual response rates have to be taken into account.

From the variable point of view, the definition of hours worked is the same across the surveys but the measurement is not completely homogenous, mainly due to the different coverage of employees, because of the managers which are not always included and when included the data referring to them cannot always be separated from that for all the other employees.

Table 2.1 – Variables on hours paid and worked in the surveys and register

Data source	Reference period	Employment coverage	Hours worked (w)			Hours paid not worked (p)	Hours paid (w+p)
			Total (w_o+ w_s)	Normal (w_o)	Overtime (w_s)		
PMI	Year	managers included	x				
SCI	Year	managers included	x				x
VELA	Quarter	managers excluded	x	x	x	x	x
GI	Month	managers excluded	x	x	x	x	x
RACLI	Year	breakdown for managers					X[(a)]

Source: Authors' depiction
(a) It does not include overtime hours.

It is important to underline that the surveys under analysis have different data collection and validation processes. Consequently, here all aspects have been compared in order to monitor any significant difference in the measurement of the same variable.

On the other side, the register proxy variable has to be compared with survey variables in order to assess the effect of the lack of the overtime component in the register data.

A very deep and careful analysis of the data and the eventual underlying relationship structure has followed, to assess the possibility to build a data set of statistical units with consistent variables coming from different sources.

The steps that have been followed to establish a coherent mixed-sources process are described in the following. First of all the analyses on the measurement of variables by different sources and the rules according to which the coherence between units and variables can be ascertained are described. Then the methods applied in order to identify suitable groups of enterprises, with specific worth considering characteristics, are presented.

3. Assessment analysis and integration of the data sources

In the following, the main results of the analysis of the difference in measurement in the several sources are shown. As first step, the analysis of the released aggregate estimates are presented to assess the coherence at macro-level. Because of some unexpected results, different hypotheses have been analyzed. Since the final aim is to build a micro-integrated dataset of statistical units with homogenous variables, a main aspect has been studied: whether given the same definition, there is any difference

in the measurement of the variable due to the effects of the validation process. Some evidences, in this regard, led to deeper investigation, assessment and comparison among the sources. Hence, the following studies have been done over the group of respondent statistical units common to the several surveys. Once they are linked, the assessment of the similarity or difference in the measurement of the same variable on the same units could help in identifying a possible survey measurement bias.

The final aim has been to identify clear rules to build the integrated data set, covering the entire list of units for which information from various sources would be at disposal in a chessboard way.

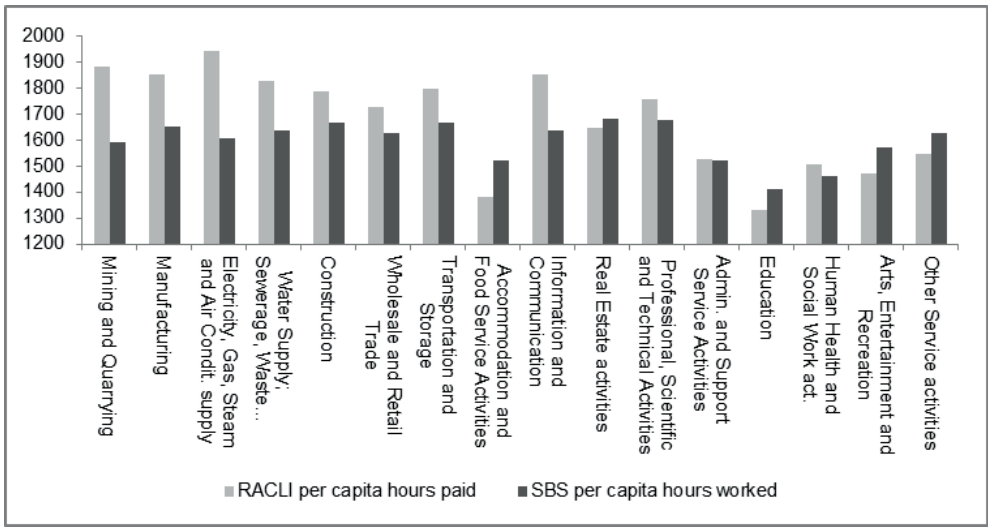
3.1 The comparison of hours worked estimates

In order to assess the coherence among the several sources, a comparison between register data and survey macro estimates officially released for SBS and STS regulations have been carried out. All results are shown in per capita terms.

The proxy variable of per capita hours paid has been compared to the per capita hours worked released by the surveys by economic activity, at the Nace Rev. 2 section level.

For the total population of enterprises with at least one employees, RACLI per capita data on hours paid may be compared with per capita hours worked from SBS estimates, see Figure 3.1.

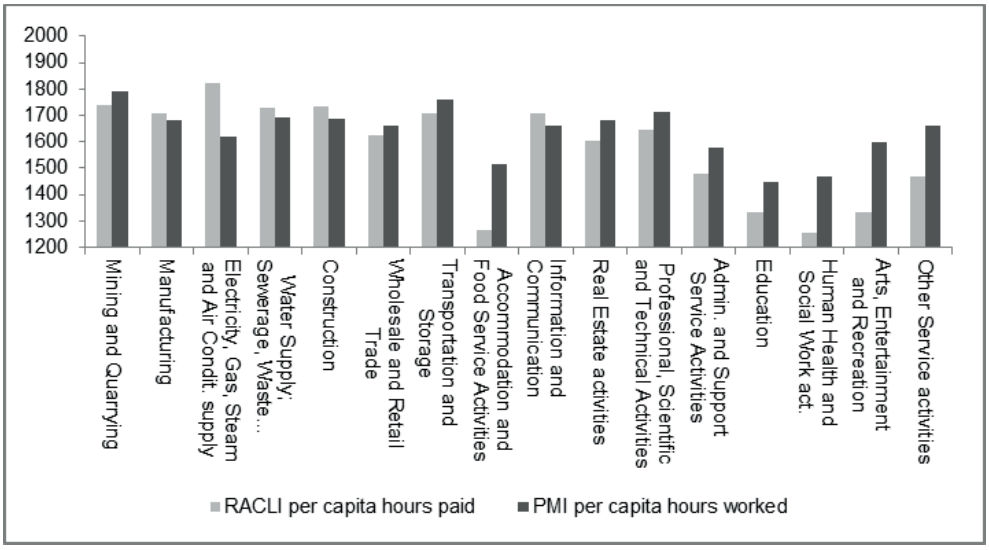
Figure 3.1 – SBS per capita hours worked and RACLI register hours paid in all enterprises with employees by economic activity. Year 2012



Source: Istat, RACLI Register, SBS disseminated data

Hours worked are expected to be lower or at most equal to hours paid, because the latter variable includes the first one, plus hours paid but not worked due to paid holidays, sickness, work permits, etc.[10]. Nevertheless, in some economic activities in the services sectors[11] the estimate of per capita hours worked calculated based on the structural PMI and SCI surveys is higher than the RACLI based estimate of per capita hours paid. To investigate this unexpected result, the population is divided in two groups of enterprises, with less and more than 10 employees, respectively covered only by PMI and covered by both PMI-SCI and GIVELA. As it is shown (Figure 3.2), the above-described result is mainly due to enterprises with less than 10 employees, where the estimate of hours worked based on PMI is higher than the RACLI based estimate of per capita hours paid for almost all of the services sections.

Figure 3.2 – PMI per capita hours worked and RACLI register hours paid in enterprises with less than 10 employees by economic activity. Year 2012



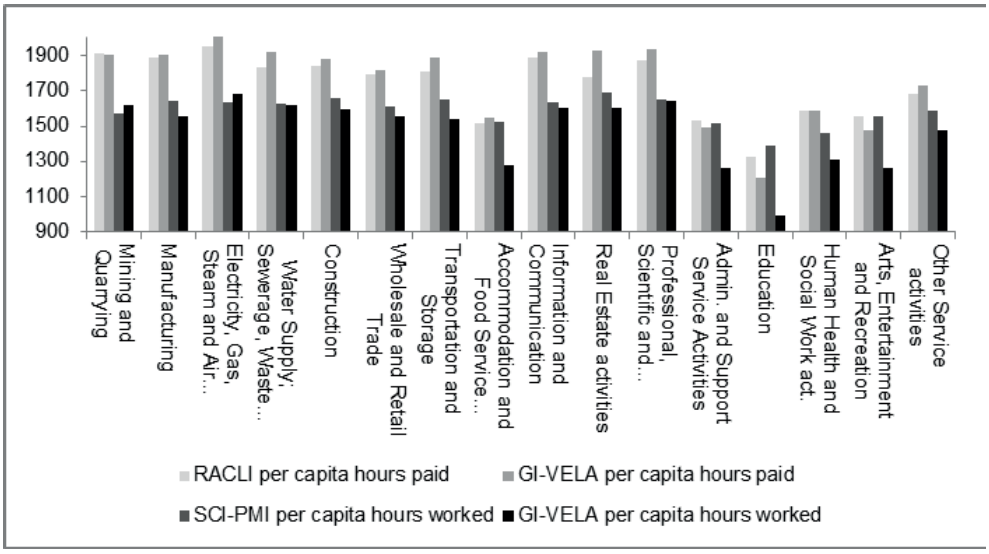
Source: Istat, RACLI Register and PMI Survey

On the population of enterprises with at least 10 employees, which is covered also by the short-term GI and VELA surveys, the estimate of per capita hours worked calculated on the basis of these two surveys is lower than the RACLI based estimate of per capita hours paid, as expected (see Figure 3.3). For this size class also SCI+PMI per capita hours worked are in almost all sections lower than the RACLI per capita hours paid even if they are always higher than the GI-VELA estimates.

10 The only case when hours worked can be higher than hours paid is in enterprises where there is a compensatory time or "time off in lieu" scheme and in the considered time period above average hours worked have not been compensated by an equal amount of time off.

11 Industry includes sections from B to F, while sections G to S are classified in services.

Figure 3.3 – SCI-PMI and GI-VELA (a) per capita hours worked and RACLI register hours paid in enterprises with 10+ employees by economic activity. Year 2012



Source: Istat, RACLI Register, SCI-PMI and GI-VELA Surveys
(a) GI-VELA hours paid includes overtime hours.

Many complex issues have been identified as underlying the unforeseen results in the services sector. This has driven further analyses aimed at investigating all the aspects of how different production processes can generate different estimates of the considered variables. Thus, an assessment of how the RACLI register proxy variable on hours paid is affected by errors of measurement (underestimation of the overtime hours and overestimation due to hours not paid in paid days) has been made. A more in depth exploration of hours paid has been carried out comparing RACLI and GI-VELA microdata at enterprise level (see § 3.2), that helps to focus on the specific component of overtime hours and which is the effect of not measuring it in the register data.

Beyond such issue, all the aspects described in the previous paragraphs have been taken into account in order to identify how the different production processes can explain the discrepancies in hours worked measurement and to delineate a quality assessment of the several sources, with respect to the construction of a consistent integrated dataset (see § 3.3).

3.2 Quality assessment of hours paid

The subset of linked respondent statistical units across all surveys has been built, to proceed with the necessary deeper analyses about the quality and characteristics of the relevant variables and their appropriateness for the model design.

To assess the suitability of the hours paid estimated in RACLI as auxiliary variable in a predictive model, this variable has been compared at micro level with

Table 3.1 – Difference between RACLI and GI-VELA hours paid, total and net of overtime hours, on GI-VELA respondents (net of managers) by economic activity. Year 2012

Economic activity	Num. of enterprises	Difference between RACLI and GI-VELA hours paid (%)				Difference between RACLI and GI-VELA hours paid net of overtime (%)			
		Mean	Q1	Median	Q3	Mean	Q1	Median	Q3
B - Mining and Quarrying	382	-0.9	-8.4	**-2.2**	3.5	1.9	-4.2	**1.0**	6.2
C - Manufacturing	4,625	0.2	-6.1	**-1.1**	4.0	2.5	-2.9	**1.5**	6.0
D - Electricity, Gas, Steam and Air Conditioning Supply	243	-2.7	-7.1	**-2.3**	1.4	0.6	-2.6	**0.9**	4.2
E - Water Supply; Sewerage, Waste Manag. and Remediation Activities	661	-3.4	-8.7	**-2.8**	2.0	0.3	-4.3	**1.2**	5.4
F - Construction	907	-0.4	-11.0	**-2.5**	4.2	2.7	-9.3	**0**	5.9
G - Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles	1,235	-1.5	-7.6	**-1.5**	3.2	0.9	-4.7	**0.6**	5.3
H - Transportation and Storage	734	-3.0	-12.3	**-3.2**	3.6	-0.3	-9.5	**-0.2**	6.0
I - Accommodation and Food Service Activities	1,625	2.8	-12.2	**-1.9**	7.6	4.6	-9.9	**-0.3**	9.4
J - Information and Communication	542	8.9	-6.3	**-0.3**	5.5	10.5	-4.1	**1.4**	6.6
K – Finance and Insurance Activities	477	-0.2	-5.3	**0.3**	4.6	1.2	-3.4	**1.6**	6.2
L - Real Estate activities	158	-1.8	-6.8	**-1.5**	3.3	-0.2	-5.5	**-0.1**	4.7
M - Professional, Scientific and Technical Activities	552	1.0	-6.2	**-0.8**	4.2	2.7	-4.4	**1.1**	6.0
N - Administrative and Support Service Activities	580	-1.8	-10.1	**-1.8**	3.6	1.6	-7.2	**0.9**	7.6
P - Education	218	10.4	-7.1	**1.9**	13.4	11.1	-5.5	**3.2**	14.2
Q - Human Health and Social Work act.	406	-1.8	-8.5	**-2.4**	3.5	-0.5	-7.0	**-1.0**	4.9
R - Arts, Entertainment and Recreation	262	20.5	-13.6	**-2.4**	8.7	26.0	12.4	**-0.2**	10.7
S - Other Service activities	178	-3.4	-11.1	**-3.6**	5.1	-1.2	-9.1	**-0.1**	6.8
Industry net of Construction	5,911	-0.4	-6.7	**-1.4**	3.7	2.1	-3.1	**1.3**	5.8
Industry	6,818	-0.4	-7.1	**-1.5**	3.7	2.2	-3.5	**1.2**	5.8
Services	6,967	1.5	-8.9	**-1.4**	5.0	3.7	-6.8	**0.4**	7.0
Total	**13,785**	**0.6**	**-8.0**	**-1.5**	**4.3**	**2.9**	**-5.1**	**0.9**	**6.3**

Source: Authors' calculation on RACLI Register data and GI-VELA Surveys

the hours paid measured by the surveys. This comparison has been possible only with GI and VELA data, because the PMI survey does not measure hours paid and both structural surveys measure total hours worked without distinguishing between normal time and overtime.

For GI-VELA respondents both the values of the hours paid measured in the survey and those elaborated in the RACLI register are available. In Table 3.1 the main indicators of the distribution of the percentage difference between the two measures of the total amounts of hours paid are presented[12]. The median of the differences (1.5%) indicates a lower, albeit very near, level of the RACLI proxy of hours paid with respect to the GI-VELA measure. The size of the difference is almost equal in industry and services, although it shows more variation but nearly always the same sign across economic activity sections.

Furthermore, when the comparison is carried out on hours paid net of overtime, the GI-VELA measure and the RACLI proxy are even closer and the register measure is in median 0.9% higher than the survey one.

By enterprise size it seems that the overestimation in the RACLI hours paid data is negligible for the small enterprises (under 100 employees), while it is more relevant for largest ones (Table 3.2).

Table 3.2 – Difference between RACLI and GI-VELA hours paid, total and net of overtime hours paid by size, on GI-VELA respondents (net of managers). Year 2012

Employees size class	Enterprises	Difference between RACLI and GI-VELA total hours paid (%)				Difference between RACLI and GI-VELA hours paid net of overtime (%)			
		Mean	Q1	Median	Q3	Mean	Q1	Median	Q3
<=9	2,016	-1.1	-11.5	**-1.9**	4.5	0.3	-10.2	**-0.2**	5.7
10-20	4,741	0.7	-7.8	**-1.6**	3.6	2.9	-5.1	**0.4**	5.2
21-99	4,435	1.8	-7.6	**-1.7**	4.1	4.3	-4.5	**0.9**	6.2
100-499	1,249	4.8	-5.9	**0.0**	7.3	7.8	-2.6	**2.7**	9.9
500+	1,344	4.0	-4.5	**0.1**	5.3	7.7	-1.0	**3.2**	9.2
Total	**13,785**	**0.6**	**-8.0**	**-1.5**	**4.3**	**2.9**	**-5.1**	**0.9**	**6.3**

Source: Authors' calculation on RACLI Register data and GI-VELA Surveys

This evidence confirms that the RACLI proxy of hours paid is affected by both the underestimation due to the exclusion of not contracted overtime and the overestimation due to the inclusion of hours of unpaid absence in paid days and indicates that the size of each of these effects is not so relevant.

12 To this aim, managers are not included the hours paid in RACLI elaboration

Nevertheless, such effects are much characterized by the economic activity and by size, hence as far as it affects the relationship under study, it should be accounted for by the definition of a suitable stratification as part of the final model specification.

3.3 Issues to be considered on hours worked

Further analyses to understand the discrepancies occurring between the survey hours worked estimates has led to outline different issues that need to be tackled.

Firstly, the presence of a bias response effect in enterprises has been investigated considering that it has already emerged in earlier studies for the turnover variable (Oropallo, 2010; Casciano et al, 2011) and for hours worked (Baldi et al., 2013). To this aim, specifically designed analyses have been exploited, that compare main indicators of the correlated variable hours paid on the whole population, properly divided between respondents and non-respondents. For each survey the sample or the entire target population have been divided between the respondents and the not respondents to the survey[13]. In general, the survey respondents appear to have per capita hours paid higher than non-respondents average of hours paid based on the RACLI data. The set of respondents seems to be affected by a self-selection phenomenon negatively correlated to the enterprise size, as a consequence it is stronger for smaller enterprises. This phenomenon seems more relevant for structural surveys, in particular for PMI on the subpopulation of enterprises with less than 10 employees, characterized by lower response rate, the self-selection appears as more intense.

The calibration phase of the PMI survey shows that it cannot correct for the bias in per capita hours worked of the respondents. In fact, the grossing up to the reference population carried out by the PMI survey is based on a calibration on known totals on number of enterprises and jobs and it is not designed to solve any bias problem.

A similar consideration can be done for the editing and imputation procedures for SCI unit and item non-responses and PMI item non-responses: within these procedures, the imputation of hours worked is based on the respondents values (bias affected), observed on the previous year release. Hence, it can cause an overestimation of the final imputed data.

Hence, SBS estimates of hours worked, calculated based on the SCI and PMI surveys may be upwardly biased.

On the other hand, the short-term business surveys (GI and VELA) are much less affected by this self-selection phenomenon due to their much higher response rates. Furthermore, per capita hours worked enter in the sample design of the VELA survey. Its sample allocation is in fact carried out minimizing sample size under

13 This means that the entire target population has been divided into two sets: survey's respondents and all the other population units.

constraints on the maximum CVs on a set of variables which includes per capita hours worked.

Finally, for enterprises responding to both a structural and a short-term survey, the data on hours worked were compared to assess whether the questionnaire the sample unit was responding to and/or the data collection mode affected the measurement of the variable of interest. On average, in these cases hours worked were measured as lower by the short term surveys than by the structural ones. Several aspects of the production processes seems to cause differences in the variable measurement among surveys. We could assume that the timing of the survey could affect measurement accuracy, and result in more precise measures for monthly/quarterly data than for annual one. The discrepancies could also be a consequence of the different information system within the company used to calculate the requested information.

The much greater relevance of the hours worked data in GI and VELA with respect to PMI and SCI and the implications in terms of editing and imputation and validation procedures have led to give the priority to GI-VELA data in the building of an integrated data set of survey data. Therefore, when for a given enterprise and a given year a response is available both from a short term and a structural survey the GI-VELA response is selected. SCI and PMI respondents data (excluding the imputed ones) are used for enterprise non available in the STS surveys.

All these analyses on respondents and non-respondents stressed the importance of defining a proper stratification strategy, to consider all the aspects that influence the relation between hours worked and paid, to optimize the model estimation in every strata. The availability of detailed structural information in the RACLI register allowed to identify specific labour or enterprises structures and economic events to deal with, such as a high share of part-time or job-on-call workers or a large share of employees in short-time working. Moreover, the analyses have shown that also the subpopulation of enterprises with less than one employee has to be considered separately. These very small enterprises are difficult to be surveyed, showing a lower response rate also in the PMI survey, and a peculiar relationship between hours worked and hours paid.

To this purpose different sub-populations have been identified for a more correct estimation of the model parameter (see § 4.2.1).

3.4 The final integrated dataset

The first step of the estimation process is the construction of an integrated dataset of microdata containing all available information on the target population, defined by the active enterprises with employees of the Business Register (BR Asia) belonging to the Nace economic activities covered by the SBS regulation. The final dataset would gather the auxiliary variable on every units, while the target variable would

be available only on the observed units properly chosen from the surveys as to be representative of the remaining part of the population. This set of units would be used to perform the estimation model.

The starting point for the construction of the dataset is the definition of a criterion to identify clearly the same unit in all the considered sources (RACLI, SBS and STS surveys). This identification can be non-trivial because the different processes of the sources can generate problems in the identification of the 'same enterprise' from the point of view of the economic significance. A record linkage operation has therefore been carried out, keeping an acceptance level of mismatch errors close to zero with the aim to avoid any misalignment (Zhang, 2012).

The key linking variable used for matching units is the Statistical Business Register code, available in all the sources, as unique business identification number (BIN). Despite an accurate pre-matching process, some problems in using the BIN equality function as unique key still remain, causing 'not matched' and 'false matched' pairs. The first event occurs very rarely and is due to surveys' coding errors (formal errors) or refers to codes that identify enterprises that have ceased to exist, while 'false matches' occur when codes are linked correctly but data refer to substantially different units. The main reasons explaining this phenomenon are the different rules about registration and statistical treatment of business longitudinal changes together with the different timing of the data collection across the various surveys and RACLI (Baldi et al., 2011).

For surveys 'not matched' units a match with units in the register is attempted using the company name, but only for large firms or for specific economic activities with small populations.

The detection of 'false matches' is done through an indicator function, based on the difference in the annual average of the number of employees between RACLI and a specific survey. These differences are calculated separately for each unit for which RACLI and survey data have been linked via the BIN and are based on a measure of the annual average of jobs in RACLI harmonized with how this variable is obtained in a specific survey. The table below illustrates the differences across surveys in the calculation of this variable.

Table 3.3 – Annual average of employees in the SBS and STS surveys

Source	Annual average of employees
PMI - SCI	average of the end of month stock for each month of the year
VELA	average of quarterly employees across the four quarters of the year (managers excluded) (quarterly employees are averages of end of previous quarter and current quarter stocks)
GI	average of monthly employees across the twelve months of the year (monthly employees are averages of end of previous month and current month stocks)

Source: Authors' depiction

Thanks to the availability in RACLI of monthly information on employment for each job within each enterprise, it has been possible to calculate the employment variable according to the different survey definitions above described.

The match is assessed by comparing the value of the indicator function with a threshold and is accepted if the difference in the annual average of the number of employees between RACLI and a specific survey is below the threshold. For large firms and for specific sectors, characterized by a high turnover or seasonal workers that can imply wider differences in the number of employees measured by different sources, a higher threshold is considered.

Furthermore, some large units with differences above the threshold are also assessed by experts on the basis of the information in the specific data base on enterprises demography and changes of the BR and in the GI data base. For most of these units (about 113 in 2012) survey data are, however, used for the final estimate of hours worked but not in the model estimation. The table below shows the difference between matched units by BIN and validated ones. This set of units will be the set over which the model will be performed.

Table 3.4 – Enterprises by class of employment and result of the match. Year 2012 (number and percentage)

Size class of employment	Target population RACLI (num. of enterpr.)	Observed units in surveys	Not matched units Racli-surveys (%)	Matched units Racli- surveys by BIN only (%)	Validated matched units Racli- surveys (%)
1-9	1,444,468	11,345	1.4	98.6	84.0
10-19	123,096	10,071	0.5	99.5	89.5
20-49	51,277	7,422	0.3	99.7	88.2
50-249	21,288	6,601	0.6	99.4	92.3
250-499	2,087	1,366	0.0	100.0	96.6
500+	1,510	1,429	0.0	100.0	97.3
Total	**1,643,726**	**38,234**	**0.7**	**99.3**	**88.7**

Source: Authors' calculation on RACLI Register data, GI-VELA and SCI-PMI Surveys

As there are overlaps in the target populations of the four surveys (GI - SCI/VELA - SCI/VELA- PMI), for some units data from more than one survey are available, leading to more than one possible record matching with RACLI. The much greater relevance of the hours worked data in GI and VELA with respect to PMI and SCI and the implications in terms of editing, imputation and validation procedures described in paragraph 2.1 have guided the choice to prefer GI and VELA microdata over PMI and SCI when for a given enterprise and a given year a response is available both from a short term and a structural survey.

The final dataset contains for all the enterprises in RACLI the information related to employees and paid hours from the register itself and, for the linked enterprises, the information on hours worked and STW hours coming from the selected survey.

The Table 3.5 shows the breakdown in the final dataset of the selected units by source in terms of employees that are available for the estimation model. The composition of the validated set of units, by source, put the light on the issue that the data from the short term surveys play an important role for the estimation model while the PMI is fundamental for the very small enterprises stratum.

Table 3.5 – The final dataset in terms of employees: RACLI population and linked enterprises by survey and class of employees. Year 2012 (number and percentage)

Employees size class	RACLI population (number of employees)	Linked units (number of employees)	Employee coverage (percentage)	Linked units by survey (percentage)		
				Vela-GI	PMI	SCI
< 1 empl.	414.714	899	0.2	-	100.0	-
1-9 empl.	2.722.556	28.095	1.0	27.4	72.6	-
10-99 empl.	3.559.995	377.065	10.6	49.4	50.5	0.1
100-249 empl.	1.107.889	304.607	27.5	27.0	-	73.0
250-499 empl.	697.297	267.673	38.4	40.6	-	59.4
>=500 empl.	2.793.292	2.360.063	84.5	88.2	-	11.8
Total	**11.295.743**	**3.338.402**	**29.6**	**73.9**	**6.3**	**19.8**

Source: Authors' calculation on RACLI Register data, GI-VELA and SCI-PMI Surveys

4. Estimation model

The final aim is to estimate the total amount of hours worked for the domains required by the SBS regulation. Model-based sampling theory begins by recognizing that problems of estimation of finite population characteristics are naturally expressed as prediction problems (Valliant et al., 2000). To estimate a finite population total from a sample is equivalent to predict the total of the non-sample values. In this context, the values of the target variable are available on a subset of units, observed by several surveys. On the other side, the auxiliary variable, which is strongly correlated with the target one, is available for each enterprise in the target population from an administrative source.

In the following, the general scheme of the prediction approach is presented and the specifications tailored to this issue are described. The aspects that have been analysed in depth are about how the final target parameter is influenced by the description of the enterprise structure. Indeed, the definition of different sub-populations has resulted to be very significant, in order to tackle all the challenges

arisen during the preliminary studies on the estimation of hours worked in the presence of such a rich but complex informative context.

4.1 Prediction theory

Supposing that the number of units N in the finite population is known and that a number y_i is associated to each unit, the prediction approach treats the numbers $y_1,\ldots y_N$ as realized values of random variables $Y_1,\ldots, Y_N$. Once a sample $s<N$ is observed, the estimation of a function of the data h ($y_1, \ldots, y_N$) entails predicting a function of the unobserved y_r (Valliant et al., 2000). Hence, the whole population can be divided into two set of units: the first set s made by the units observed by the surveys $\{y_i, i\in s\}$, the second set r including all the unobserved units $\{y_i, i\in r\}$. The final aim is to learn about the second set by studying the first.

A further assumption is that an auxiliary variable is available on the whole population, for which the following general linear model is valid:

$$M: Y= \beta X + \varepsilon$$

where

$$E(Y_i) = \beta x_i$$
$$\text{var}(Yi) = \sigma^2 \gamma_i,$$
$$\text{cov}(Y_i, Y_j) = 0,\ i\neq j.$$

Under these assumptions, the model based method starts from the estimation of the relationship between the interest and auxiliary variables on the observed units. The variable of interest values on the non-observed units are then imputed applying the estimated relationship to those of the auxiliary variable. The best linear unbiased predictor (BLUP) $\hat{\beta}$ of β under model M is:

$$\hat{\beta} = \sum_s (x_i\ y_i/\gamma_i) / \sum_s (x_i^2/\gamma_i)$$

whose final expression depends on that for γ_i, that defines the variance shape.

The problem to estimate the population total for Y is solved as follows:

$$\hat{T} = \sum_s y_i + \sum_r \hat{y_i} = \sum_s y_i + \sum_r \hat{\beta}\ x_i$$

Therefore the estimate of the target variable in each of the study domains is the result of summing up the target variable values, both observed and imputed through the model, on the units belonging it.

In the case of the estimation of hours worked, it is possible to model the problem according to this scheme considering the proxy variable of hours paid as the auxiliary one. Indeed, the integrated dataset includes both the set s of observed units, in which both the target and auxiliary variable values are available, and the set r for which only the auxiliary variable values are available.

An extensive set of analyses on the two variables and the linking function have been carried out in order to obtain a robust formalization of the problem. Indeed, the form of the expression defining γ_i, that describes the variance of the target variable and the definition of a stratum design strategy allowing the estimation on homogenous groups of units resulted to be very important to tackle all the issues that arose.

The classification variables on which strata are defined need to be linked to the target variables. Usually, economic activity and size class are used, for their relevance and for their relatively easy availability on all population units. In this case, the rich set of information available for each enterprise in RACLI has been used to profile units according also to working time and the relation between hours worked and paid. In particular, this has allowed considering in the stratum design strategy variables related not only to STW hours but also to surveys' response bias.

In the following, the main results about which factors have resulted to affect the variable of hours worked per each enterprise are presented. Indeed, other variables besides economic activity and class size have been found to be very significant in properly estimating the target parameter.

4.2 Model specification

In this case, the target variable is the number of hours worked and the proxy variable of hours paid is the auxiliary one. The whole population is defined by the list of active enterprises belonging to the Business Register (BR Asia), the set *s* is given by the data observed by the surveys, suitably chosen to form the integrated dataset, and the remaining units of the population define the set *r*.

Therefore:

Y ≡ HW number of hours worked

X ≡ HP number of hours paid

Based on preliminary analyses, the following model specification has been found to better fit the data:

1. as target variable, the total amount of hours worked for each enterprise is considered (instead of a per capita value which was originally suggested because more easily interpretable);
2. the parameter β is estimated through a heteroscedastic model. More precisely, the parameter γ_i is a linear function of the auxiliary variable, describing the increase in the variance of hours worked with enterprise size. Thus, the expression for γ_i is the following:

$$\gamma_i = hp_i$$

This means that the BLUP estimator for β in model M can be written as:

$$\hat{\beta} = \sum_{s} hw_i \,/\, \sum_{s} hp_i$$

3. the stratification on the basis of economic activity and enterprise size allows to estimate hours worked including the overtime component on the basis of the proxy of hours paid, even if the auxiliary variable does not include this component. This happens because the incidence of overtime hours over total hours worked is strongly associated with the two stratification variables;
4. to identify sub-populations of enterprises, defined by specific characteristics relating to several events and the type of work remuneration, is important. Indeed, finding the proper strategy for defining the suitable strata is necessary to better represent the non-respondent units on the basis of what is observed on the respondent ones. To this aim, beyond the usual classification variables of economic activity and size class, many other aspects have been taken into account that can be analysed through the variety of information from administrative data.

Hence, the whole scheme outline is as follows: at first, the whole target population is divided into the identified four sub-populations, defined as described below (see § 4.2.1). Afterwards, each sub-population is stratified according to economic activity and size. In this way, almost 250 different strata have been defined. They do not coincide with the study domains, for which the estimates of the total number of hours worked are obtained as sum of the observed and estimated values of hours worked on all enterprises in all the strata in the domain itself.

The hypothesized relationship is for each stratum C:

$\forall$ stratum C:

$$HW_C = \beta_C \cdot HP_C + \varepsilon$$

According to the classification made to build the integrated dataset, each stratum C can be partitioned into three sets:

s of observed units, for which data on hours worked are available from the surveys, used to estimate the model

z of observed units, accepted with a bigger threshold, so that are judged to be self-representative, not used to estimate the model

r of unobserved units, for which the values of this variable need to be estimated so that is $C \equiv (s \cup r \cup z)$.

The BLUP estimate of β_C is obtained as follows:

$$\hat{\beta}_C = \sum_{i \in S} hw_i \,/\, \sum_{i \in S} hp_i$$

Hours worked in the *j*-th not observed enterprise are then calculated as follows, based on the above indicated parameter estimate $\widehat{\beta}$ and of the number of hours paid available in RACLI, HP_j:

$\forall$ stratum *C* and $\forall$ unit $j \in r \subset C$:

$$\widehat{hw_j} = \hat{\beta}_C \, hp_j$$

Finally, the aggregated estimates of hours worked for each study domain *D* are calculated in the following way:

$\forall$ domain *D*:

$$\widehat{HW}_D = \sum_{i \in S} hw_i + \sum_{i \in Z} hw_i + \sum_{i \in R} \widehat{hw}_i$$

the sum of hours worked on all the units, observed or estimated, belonging to domain *D*. For each domain, the value of the total depends in different regards on the percentage of observed units or on the percentage of the estimated values.

4.2.1 Description of the sub-populations

Once the dataset has been built, according to standard quality requirements, a sample of observed data is available. To apply the predictive approach it has been important to assess whether the sample is representative of the remaining population, to be imputed, and which is the right classification in order to tackle all the issues that have arisen during the preliminary analyses.

A regression tree method has been used to test for the factors affecting more significantly the relation between hours worked and hours paid, represented by the parameter β. On every eventual classification so suggested, further assessment has been done studying the comparison between the hours paid on the two group of units used for the estimation for the model and the remaining units in the same set, to be imputed with the same model. These pervasive analyses have pointed out the necessity to go beyond the usual classification of size and economic activity, to weaken the effect of the bias response, that have always to be taken into account.

In this way, the relevant variables, their thresholds and the hierarchy with which the variables have to be considered, that is all the elements needed to identify the sub-populations, have been recognised. These sub-populations constitute a partition of the total population, due to the hierarchy established through the application of the regression tree method.

Four types of sub-populations have been identified as those to be considered before defining the strata on the basis of economic activity and size. They are defined as follows:

- enterprises with at most 1 employee: for such population the hypothesis under which the relation between the two variables is exactly equal to 1 has been

tested. For a consistent part of those enterprises such hypothesis has been accepted, for the remaining part a proper model has been estimated. The relevance of this size threshold has been shown by the regression tree method to dominate those of the other variables mentioned below (incidence of STW and job-on-call employees). Therefore, for enterprises with at most 1 employee these additional classification variables do not need to be considered;

- enterprises with STW incidence above a pre-defined threshold: the very detailed information on the phenomenon in the RACLI register has allowed to delineate four different profiles of STW use;
- enterprises with incidence of job-on-call employees above a pre-defined threshold: this kind of enterprises have been always under study, because for the workers on this type of contract the information about the day paid can be directly elaborated in terms of hours paid. Furthermore, they resulted to be affected by the response bias. For units with incidence of job-on-call employees above the threshold, hours worked are estimated as equal to hours paid. While for those below the threshold, hours worked are estimated through the usual model together with other units.
- the remaining enterprises (generally called no event).

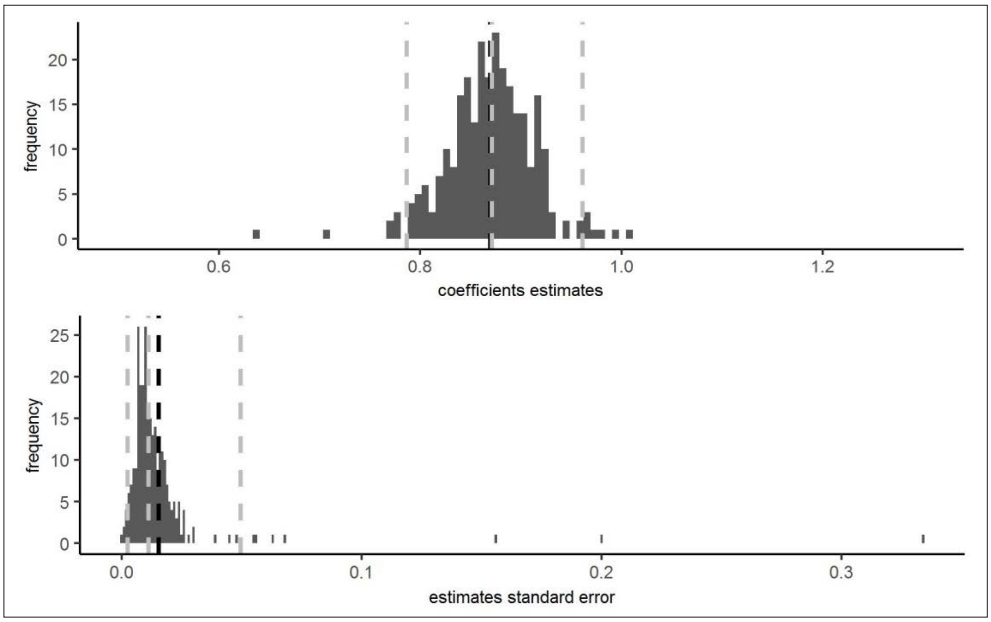
The stratification on each sub-population is carried out maximising the number of strata under the constraint that each stratum needs to include at least a minimum number of units. When this constraint requires to aggregate neighbour strata, the priority is given to keeping separate enterprises with different sizes, rather than with different economic activities as it is more common. In fact, all the analyses show that hours worked are much more influenced by the enterprise size rather than by its economic activity. This holds especially for the very small enterprises, where the sensitivity to size is very high. The level of disaggregation for the strata definition resulted to vary across sub-populations, the finest being based on 2-digit NACE and 6 size classes.

4.3 Model estimates' assessment

For all the strata, the null hypothesis of the estimate of the coefficient β_C being equal to zero is rejected. In the following, some graphs are presented (Figure 4.1) to give an idea of the overall behaviour of such model across all the strata (more than 200) used for the final estimation.

Both distributions are concentrated, especially the one of the estimates' standard error. The strata for which the coefficients' estimates are very small are those for very specific group of enterprises as the ones that use short-time working (STW).

Figure 4.1 – Histogram of coefficient estimates and their standard errors (quantiles in light grey, mean in black)

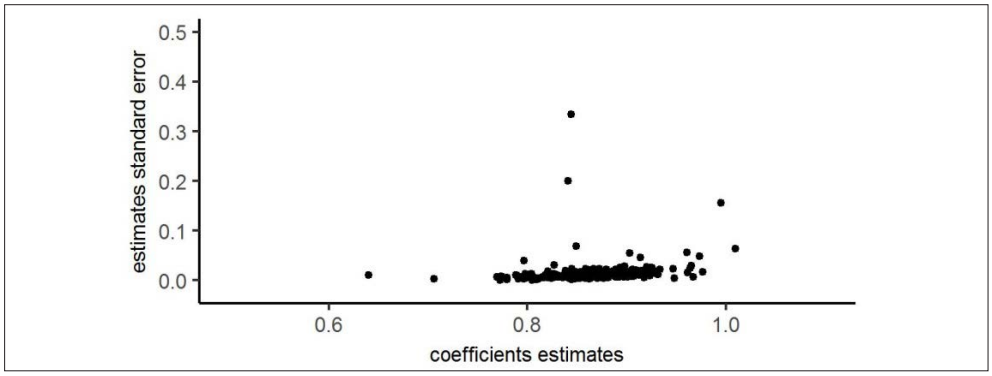


Source: Authors' estimates based on RACLI Register data, GI-VELA and SCI-PMI Surveys

The distribution of the standard errors of the estimates has also a concentration over a specific range: all of them are less than 0.07, only three of them reach values close to 0.2.

Finally, a scatter plot between the coefficients' estimates and the estimates' standard errors is analysed, to assess whether there is a relationship between the two.

Figure 4.2 – Scatter plot of coefficients' estimates and their standard errors



Source: Authors' estimates based on RACLI Register data, GI-VELA and SCI-PMI Surveys

A test on the correlation coefficient between the coefficients' estimates and their standard errors across all strata has confirmed that there is no significant linear relationship between the two of them.

5. Results and concluding remarks

5.1 Results and comparisons with previous hours worked estimates

The methodology to estimate hours worked described above has been tested on two years, 2012 and 2013. This has allowed validating the estimation's results both in level and in dynamics. In this way, the robustness of the criteria used for the identification of sub-populations and strata and the related model's parameters could be assessed.

As described above, the use of an integrated dataset as result of a mixed-source approach offers the opportunity to take into account many different kinds of information related to all the aspects of the working time structure. Indeed, a large amount of information is provided by the administrative source, not only with respect to hours paid but also to all the features characterizing different patterns of working time.

Table 5.1 – Annual hours worked per employee by economic activity and employment size class: comparison with SBS official data. Year 2013 (percentage differences between new and previous SBS data)

Employment size class	Economic activity[b]			
	Industry net of construction	Construction	Services (a)	Total Economy (a)
0-9	-9.2	-6.0	-15.9	-13.3
10-19	-4.4	-3.4	-8.6	-6.5
20-49	-3.7	-4.5	-7.7	-5.7
50-249	-2.6	-4.2	-7.8	-5.3
250 +	-3.3	-4.3	-6.1	-5.2
Total	**-4.2**	**-4.9**	**-9.9**	**-7.6**

Source: Authors' estimates based on Istat SBS data
(a) Excluding financial and insurance activities and Public Sector.

The comparison of the results of the new methodology and the old one is shown in Table 5.1 for the year 2013 (the results confirmed what was registered on the previous year 2012). The new estimates of hours worked are consistently lower than those previously disseminated for the SBS regulation and based only on the two SBS surveys. More specifically, the new estimates indicate that for the entire economy

hours worked are lower than the previously calculated figure by 7.6 per cent , respectively -4.2 per cent in industry net of construction , -4.9 per cent in construction and -9.9 per cent in services. It can be noted that the differences are strongly related to size and economic activity: the largest ones are recorded for smaller enterprises and services. It is worth pointing out that, for every sector, enterprises with less than 10 employees present a far bigger difference with regard the difference registered for the other size classes.

To explain such differences, it is useful to recall all the aspects regarding the statistical processes that have been considered and the composition of the final sample, used both for the model estimation and for the final vector on which the estimates for the domain total are built. As highlighted in Table 3.5, the sample coverage of the target population is increasing with size, together with the use of the short term data. This means that for the bigger size class, the differences can be explained mostly as a substitution effect between the SBS surveys and the STS ones.

On the other side, for the smaller size class enterprises, the coverage is consistently due to the SBS surveys, so the differences can be ascribed to the model estimation scheme. In this regards, it is important to underline that the imputation is done on units whose values can be measured only through the register data. In this regard, the knowledge of the factors influencing actual working hours in Italian enterprises has proven very relevant and it has been enhanced significantly. In particular, relatively small sub-populations of enterprises with specific characteristics implying peculiar patterns of hours worked have been identified. The measurement of hours worked in these sub-populations can present additional difficulties, but even when this is not the case the small sizes of the sub-populations make it difficult to represent them adequately through a sample survey. This concerns in particular small and micro-enterprises, firms with a significant share of low labour input employment contracts (e.g. jobs-on-call) or of absence events, or units operating in specific economic activities such as arts, entertainment, recreation and other service activities (sections R and S of the Nace rev.2 classification).

Hence, as final consideration, among the many reasons behind the relevant differences between the previously released SBS estimates and the newly produced ones, it can be recalled that the sample data used for the new estimates include where possible GI or VELA data rather than SCI or PMI ones and that in the aggregate estimates of per employee hours worked produced by the STS sources are lower than those produced by the SBS ones (see § 3.1 Figure 3.3). Moreover, SBS estimates of per employees hours worked are high also when compared with the RACLI measure of hours paid used as independent variable in the models (see § 3.1 Figures 3.1-3.3).

Finally, the prediction approach provide the chance to take into account every kind of units, also the ones that in any regards are more difficult to be reached from

the direct surveys. In these terms, the identification of specific sub-populations at first helps in avoiding the distortion effect of the bias-response on the parameter estimation. Furthermore, the final imputation on the remaining part of the population, identified through the register data, entails the final estimation to represent also the type of units considered to be very elusive.

5.2 Concluding remarks

In general, the new methodology has produced lower estimates of hours worked in comparison with those based on the SCI-PMI surveys. Furthermore, the most relevant result is that the estimates show a far greater variability across economic activity and size class of enterprises. In particular, these differences increase as the enterprise size decreases.

The break in the SBS series is not negligible, it has been deeply studied, in order to understand whether the reasons are structural or not. These differences are mostly due to the fact that the additional sources used with respect to those of the previous SBS estimates measure hours worked quite differently from SCI and PMI surveys. Furthermore, an extensive use has been made of the detailed information available for the entire population of enterprises in RACLI. The register coverage with regards to the target population allows to appropriately represent specific sub-populations that are more difficult to measure via sample surveys. Therefore, the use of administrative data as auxiliary information has helped in shedding light on phenomena that tend to be very elusive.

The results, tested on two years, have been considered statistically reliable in terms of the basic assumptions, choice of the models and coherence with the labour cost variable of FRAME. In particular, the production of the new estimates on consecutive years has allowed to test all the issues raised by the consulted experts.

Hence, starting from the reference year 2014, the estimates produced with the method described here have been disseminated officially at national level and transmitted to Eurostat to fulfil the SBS EU Regulation for the variable “hours worked”.

Despite the stability of the model across the considered period, it is suggested that in the next years the definition of the sub-populations and their threshold values are tested as a preliminary step before carrying out the estimates. Moreover the future evolution of the informative contents of the social security source on working time, the enlargement of the target population of VELA, to cover enterprises with less than 10 employees starting from 2016 and the inclusion, from the same year, of managers among the employees whose hours worked are measured by GI and VELA will provide new opportunities for improving the estimation.

References

AA.VV. 2016. *Rivista di Statistica Ufficiale*. N. 1/2016. Roma: Istat.

Baldi, C., C. De Gregorio, A. Giordano, S. Pacini, F. Solari, and M.Sorrentino. 2013. Joint use of survey and administrative sources to estimate the hours actually worked. *1st Southern European Conference on Survey Methodology (SESM) and VI Congreso de Metodología de Encuestas*, Barcelona, 12-14 December.

Baldi, C., D. Bellisai, F. Ceccato, S. Pacini, L. Serbassi, M. Sorrentino, and D. Tuzi. 2011. The system of short term business statistics on labour in Italy. The challenges of data integration. *ESSnet Data Integration Workshop*, Madrid, 24-25 November. http://www.ine.es/e/essnetdi_ws2011/ppts/Baldi_et_al.pdf.

Casciano, M.C., V. De Giorgi, F. Oropallo, and G. Siesto. 2011. Estimation of Structural Business Statistics for Small Firms by Using Administrative Data. *Rivista di statistica ufficiale*. 2-3: 55-74. Roma: Istat.

Congia, M.C., and S. Pacini. 2012. La stima da fonti amministrative di indicatori retributivi congiunturali al netto della cassa integrazione guadagni. *Rivista di Statistica Ufficiale*. 2-3: 19-40. Roma: Istat.

Congia, M.C, and S. Pacini. 2010. L'utilizzo del lavoro a chiamata da parte delle imprese italiane. *Approfondimenti Istat*. Roma: Istat.

Ilo. 2005. *General Survey of the reports concerning the Hours of Work (Industry) Convention, 1919 (No. 1), and the Hours of Work (Commerce and Offices) Convention, 1930 (No.30)*. Genève: International Labour Organization. http://www.ilo.org/public/english/standards/relm/ilc/ilc93/pdf/rep-iii-1b.pdf.

Istat. 2016. Il censimento delle imprese. *Atti del 9° Censimento Generale dell'Industria e dei Servizi e Censimento delle Istituzioni Non Profit*. Roma: Istat.

Oropallo, F. 2010. Analisi delle differenze strutturali nella performance economica tra unità rispondenti e unità non rispondenti nella rilevazione dei risultati economici delle piccole e medie imprese (PMI). *Contributi Istat*. N.7/2010. Roma: Istat.

Rocci, F., and L. Serbassi. 2008. The process of Editing and Imputation on Large Firms survey: between experience on field and computational standardization. *Proceedings of 2008 European Conference on Quality in Official Statistic*, Roma, 8-11 July.

Valliant, R., A. H. Dorfman, and R. M. Royall. 2000. *Finite Population Sampling and Inference, a prediction approach.* Hoboken, New Jersey, U.S.: Wiley, Series in Probability and statistics.

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