Italian strategy to obtain gross monthly pay for LFS employees[1]

Laura Esposito, Livia Fioroni, Sara Gigante, Silvia Loriga, Silvia Montecolle, Federica Pintaldi, Roberta Rizzi

laesposi@istat, fioroni@istat.it, gigante@istat.it, siloriga@istat.it, montecol@istat.it, pintaldi@istat.it, rizzi@istat.it

Subject: INCGROSS methodology, collection, and results

1. Introduction

Following the adoption of the IESS framework regulation (Reg. 1700/2019), starting from 2021 the Labour force survey (LFS) must collect a variable on gross monthly pay for LFS employees (INCGROSS), which replaces the INCOME variable collected till 2020 on net monthly pay.

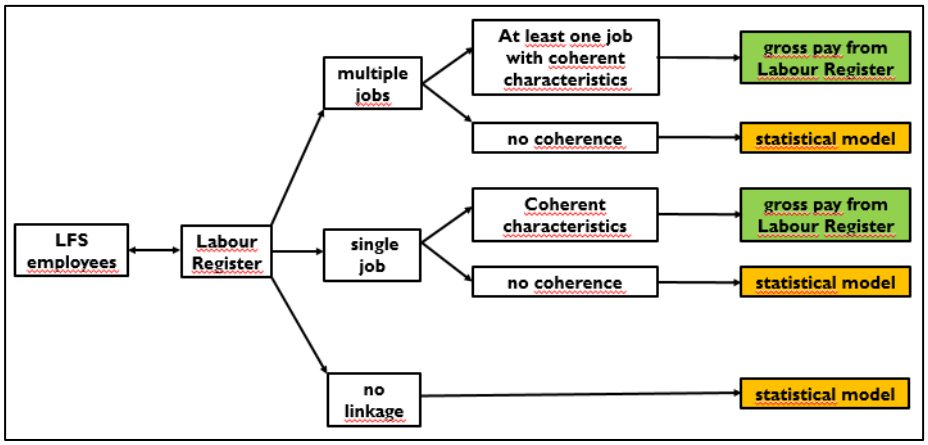
From previous tests, unlike net monthly pay, it turned out that asking directly to the interviewed person the gross monthly pay is not achievable (sensitive information, high non-response, low reliability). On the other hand, administrative information can be used to this aim. In recent years Istat is implementing the System of Integrated Registers (SIR), a harmonized set of statistical registers integrating administrative data. One of these registers, the **Labour Register** (LR) containing information on employment, including wages, may be used to obtain information on gross pay for employees in the LFS sample, through record linkage using a pseudo-anonymized individual key.

In any process, involving the integration between a survey and a register there are typical issues to deal with: missing linkage and linkage errors; under-coverage of the register; misalignment between the register and survey reference population. All these issues result in missing/incorrect linkage between survey and register have to be taken into account and properly treated. This is particularly important in case of register under-coverage and misalignment between the reference populations: in these cases, the characteristics of non-linked individuals may be different from those of the linked ones: non-linked individuals may be concentrated in certain jobs that are under-covered in the register, or may belong to the portion of LFS reference population outside the register reference population (for instance irregular work). Moreover, even successful linkage should be validated before taking the administrative variable and linking it to the LFS microdata, to be sure that there are not errors in the linkage and the job to which the gross pay in the register corresponds is the same that was declared as main job by the LFS respondent (this is particularly relevant in case of multiple jobs for the same individual).

The strategy to obtain gross monthly pay perceived by the employees has been designed after having conducted an extensive experimental study (financed by a Eurostat grant). In a nutshell, the variable INCGROSS is obtained adopting a **mixed strategy** based on **administrative data** and **model prediction**: i) LFS collected data are linked with the LR, and, subject to a validation step, information on the wage is imputed from the register (whether it is available and usable); ii) model prediction is applied in all cases for which information from the register is not available (missing linkage, missing value, irregular work) or unusable (not validated linkage, inconsistency between job characteristics in the LFS and in the LR, outliers). The main model is estimated on the LFS subsample in i), exploiting a very large set of covariates available in LFS sample, very predictive for the gross wage; specific models are used for the subsamples in which we assume a high probability of having an irregular or a "grey" job (that is a partially regular job, in which part of the wage is undeclared in administrative sources).

[1] This work is carried out within an Istat Task Force whose members are: Ciro Baldi, Laura Esposito, Dario Ercolani, Livia Fioroni, Sara Gigante, Silvia Loriga, Alessandro Martini, Silvia Montecolle, Silvia Pacini, Nicoletta Pannuzzi, Federica Pintaldi, Roberta Rizzi.

Figure 1: **IT-LFS strategy for the gross monthly pay**



2. Operative definition of gross wage from administrative data

As mentioned in par. 1, the strategy adopted is a combination of two approaches: record linkage with LR and statistical model. The LR is focused on labour market information and covers all regular paid jobs, active in the national territory and in all sectors of economic activity. The statistical unit is the job position that, in the context of dependent employment, corresponds to the employment contract between an employer and an employee. The use of LR allows to obtain monthly gross pay for the main position of LFS employees using also information related to the reference week. The setting up of such Register in Italy is based on a plurality of administrative sources each with different coverage, variables, units, reference periods, metadata, timeliness etc. This implies that the Register is produced through a complex process of harmonization and integration of these data sources. The defining aspects of gross wage also raise relevant issues. In fact, since administrative data are not collected for statistical purposes, the information have been partially treated to derive a new statistical variable taking into account the requirements of regulations. In particular, additional monthly payments (e.g. yearly payments such as 13th month) in administrative data are included (in most cases) in the gross salary of the month in which they are paid according to a cash criterion; therefore they have been hived off and added in proportion to the monthly salaries received. Moreover, for public sector the employer replaces social security institutions, maternity and parental leave are included in estimated gross wages and cannot be separately and clearly identified as social benefits. For private sector, instead, gross wage includes payments made by the employer while those paid by social security institutions (sickness, maternity, parental leave and lay-off) are excluded. To harmonize the estimate of gross wage between private and public sector, the allowances payable by social security institutions were added to the gross salaries in the private sector.

3. Record linkage and validation

The job characteristics of each LFS employee, collected from the survey, refer to the main job position during the reference week. The aim of this work is to associate the main job position stated by the employee in the survey with that one deriving from the LR and consequently to attribute the gross monthly pay information derived from the LR itself. Therefore, the methodological approach is based on linking to each LFS employee all job positions in LR active during the month including the LFS reference week in order to pick the one that mostly resembles the main position chosen in the survey. This process consists of two steps:

1 - record linkage, possible thanks a unique pseudo-anonymized person ID associated to each individual in LR and in LFS, which accounts of privacy requirements (from the code it is not possible to directly obtain the personal information to identify the individual) but allows to link information regarding the same individual. Finally, each LFS employee is linked with one or more job positions active during the reference period in the LR.

2 – validation, the consistency between the LFS and LR job characteristics was analyzed, using common variables: working time (full time/part time), type of contract (temporary/permanent) and economic activity

sector (NACE). Comparing the linked records between LFS and LR, each job position could result fully validated (all variables are coherent), partially validated (two variables or only one is coherent) or not validated (no one variables are coherent).

The result of the record linkage between 2019 LFS sample (around 154 thousand employees) and 2019 LR job positions is: around 94% of the LFS employees have at least one LR job position (81% have only one job position and the remaining 13% have two or more job positions).

The working time is the variable considered more relevant because determines different pay levels. In cases of only one job position is linked to the LFS employee, this is considered validated if there is consistency in the working time (full time/part time). In cases of multiple job positions, the validation analysis helps to choose the job position that most closely resembles the LFS main position: i) if only one job position is fully validated, it is chosen; ii) if more than one job position is fully validated, the one with the higher number of worked hours is taken as main jobs; iii) if all the job positions are partially validated, the job position is chosen giving priority to the part time/full time consistency; iv) if all the job positions are partially validated and have part time/full time consistency the same criterion mentioned above is followed (higher number of worked hours).

4. Prediction through statistical model

To develop the statistical models to predict gross wage in cases in which it is not possible to impute administrative data from LR, it is preventively necessary identifying the following four subsamples:

1. LFS employees linked with LR and validated (consistent working time); information on gross wage is available and usable.
2. LFS employees not linked with LR or linked but not validated; irregular job is excluded, i.e. public sector. In this group are also included:
   - LFS employees linked with LR and validated; information on gross wage is missing or outlier[2];
   - LFS part-time workers linked with LR but partially validated because full-time in LR; all activity sectors;
   - LFS full-time workers linked with LR but partially validated because part-time in LR; irregular job is excluded, i.e. public sector.
3. LFS employees not linked with LR or linked but not validated; irregular job is assumed.
4. LFS full-time workers linked with LR but partially validated because part-time in LR; “grey” job is assumed (partially regular job, in which part of the wage is undeclared in administrative sources)

The adopted strategy is different in each of these groups:

1. Administrative information on gross wage from LR is imputed to LFS employees.
2. Administrative information on gross wage from LR is unavailable or unusable; a statistical model (GROSS_MODEL) is estimated on the subsample in group 1 (linear regression, where Y=LR gross wage, X=a large set of LFS variables); this model is used to predict gross wage in group 2.
3. In this group irregular job is assumed (no social security contributions and no taxes are paid), so the amount of income assigned to these LFS employees reflects the net wage: if the LFS variable on net monthly pay (still collected by the Italian LFS) is available (no missing and no outlier), this amount is imputed in the INCGROSS variable; if the LFS net wage is not available (missing or outlier) a statistical model (NET_MODEL) is estimated on the previous group (linear regression, where Y=LFS net wage, X=a large set of LFS variables); this model is used to predict the amount to be imputed in the INCGROSS variable.
4. In this group, even if administrative information on gross wage is available it cannot be imputed, because it is necessary to verify if there could be part of the wage undeclared in administrative

[2] The outlier detection is conducted comparing the observed values with the predicted values through the statistical model; observations for which the relative difference between observed and predicted values is higher are considered outliers (the statistical models are re-estimated after their elimination).

sources. To this aim, individual LR gross and LFS net monthly pay are compared (if they are not available the predictions obtained with GROSS_MODEL and NET_MODEL are used) and a statistical function NET_GROSS is used to derive net from gross monthly pay, based on observed data and the income tax brackets. To estimate the part of the wage undeclared in administrative sources, the net wage (from LFS or NET_MODEL) referred to a LFS full-time job is compared with NET-GROSS function applied to the gross wage (from LR or GROSS_MODEL) referred to a LR part-time job; if the former is higher than the second, the difference represents the prediction of the undeclared income and it is added to the gross wage.

The statistical model GROSS_MODEL is rather simple, it implements a simple linear regression, but shows a good fit (the same is for NET_MODEL). The strength of this model derives from the very large set of covariates, highly correlated with the dependent variable (gross monthly pay). The covariates we considered (LFS variables) are:

- Geographical territory, socio-demographic variables and household composition:

  NUTS1, big municipality (more than 250,000 resident persons), degree of urbanization, gender, age, age-squared, citizenship, number of family members, household typology, family role (derived from the relationship with the reference person).

- Background:

  Highest level of education, years of work in this job, years of work in this job-squared.

- Job characteristics:

  Professional role (Italian variable reflecting contractual status), supervisor responsibility, occupation (ISCO1digit), temporary/permanent job, number of employees in the local unit, economic activity sector, full-time/part-time job, usual worked hours, actual worked hours, absence or reduced hours, overtime, net monthly pay (only in GROSS_MODEL).

The independent variables selection has been conducted through a stepwise procedure, most of the variables included in the model are significant. The fitness of the model is good: R^2>70-80% in the model excluding monthly net income, it becomes even higher including monthly net income.

5. Results

At the conclusion of the process (linkage, validation and identification of outlier gross wage) administrative information on the gross wage from LR is imputed for 84% of 2019 LFS employees. For the remaining 16%, gross wage is estimated using a model: in particular, 10% of 2019 LFS employees has job position that is not validated and the remaining 6% are not linked.

The percentages are different according to socio-demographic and job characteristics. The share of job positions for which gross wage can be directly imputed from the register is higher among men, those aged 35 years or older, nationals, employees with a higher level of education, permanent employees and full-time employees.

Considering sector of economic activity, more than 90% of employees in Industry (excluding construction), Public administration, Education and Information and Communication has job positions validated. In contrast, Other personal and public services, Accommodation and food services, Agriculture and Household services have the highest percentage of model-estimated wages (around 30% or more).

As expected, gross wage has a higher level and greater variability than net wage due to the redistributive effect of the tax system, but also due to a greater precision in the estimate that is not affected by individuals rounding their wages. The coefficient of variation increases from 15 of net wage to 27 in gross wage and the ratio of the 90th to the 10th percentile increases from 2.9 to 4.6.

When analyzing gaps among socio-demographic characteristics the largest differences are between gross and net wage are by citizenship and educational qualification.

Figure 2: Gross and net remuneration density distribution for employees. Year 2019

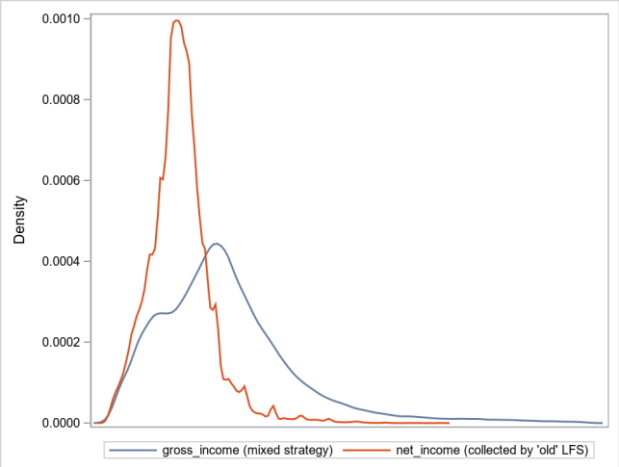
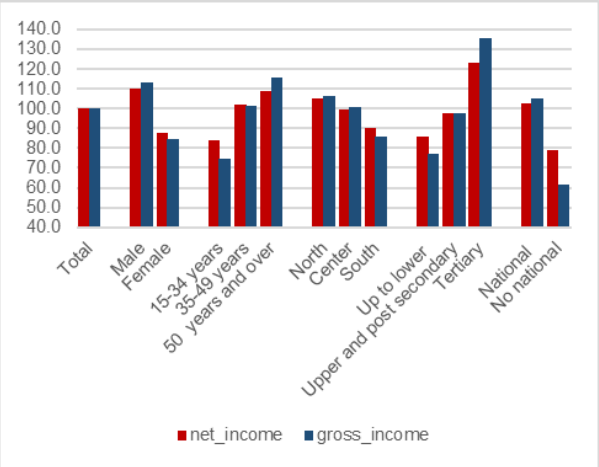


Figure 3: IT-LFS comparison of net and gross wage (index numbers). Year 2019



In conclusion, although the use of administrative data suffers from a delay in the dissemination of information due to the nature of these data, the strategy adopted for the estimation of gross income allows us to provide a high quality estimate.

References

Baldi C., Gigante S., and S. Pacini. (2018). *The development of the Italian Labour register: principles, issues and Perspectives*. Book of Short Papers SIS 2018.

Esposito l., Fioroni l., Guandalini a. 2019. Gross income projection in labour force survey data. *international journal of economic sciences*, vol iii, no. 3. pp. 50-65.

European Commission Eurostat 8 July 2021, Eu labour force survey explanatory notes (to be applied from 2021q1 onwards). Directorate f: social statistics, Unit f-3: labour market and lifelong learning.

Herzog, T.N. Scheuren, F.J. a Winkler, W.E. (2007). *Data Quality and Record Linkage Tehniques*. Springer Science+Business Media, New York.

ISTAT, Gigante S., Pacini S. and Rossetti F, (2019) *I differenziali retributivi nel settore privato, anno 2017*. Statistiche Report, 9 Dicember 2019.

ISTAT (2020). *Condizioni di vita, reddito e carico fiscale delle famiglie*, Statistiche Report.

Wallgren A. and Wallgren B. (2014). *Register-based Statistics: Administrative Data for Statistical Purposes*. John Wiley & Sons, Chichester, UK.

Winler E.W. (2007). *Data quality and record linkage techniques*, Springer.