Use of administrative data for the compilation of the monthly gross income variable (INCGROSS) in the Maltese Labour Force Survey

by Charlene Abela, Tania Borg, Joslyn Magro

1 Introduction

The Labour Force Survey (LFS) is a household sample survey which provides quarterly and annual results on the employment situation of persons 15 years and over in accordance with the Integrated European Social Statistics (IESS) framework regulation (European Parliament, Council of the European Union, 2019). One of the aims of the LFS is to collect the gross monthly income (INCGROSS) for employees. The objective of this working paper is to develop a methodology for the compilation of the INCGROSS variable based on administrative sources.

2 Measuring the Gross monthly income (INCGROSS) in LFS

According to the IESS regulation, gross pay refers to the monetary component of the remuneration of employees in cash payable by an employer to an employee before deduction of income tax and national insurance contributions (European Commission Eurostat, 2021). Eurostat suggested several strategies for obtaining the gross remuneration: asking respondents for their gross income, collecting the net income, and applying a net/gross conversion model, obtaining the data from an administrative source or a combination of these strategies. The MT-LFS questionnaire collects information on employees' basic income for national purposes. Consequently, the gross monthly income will be obtained from administrative data to address the requirements of the regulation.

3. Rationale for using administrative data

The IESS regulation brought about a revised concept of the INCGROSS variable as well as a longer transmission period. For this reason, this variable could be obtained from administrative sources and the deadline would be respected. Besides, questions on income administered through CAPI or CATI tend to be more problematic to collect (Istituto Nazionale di Statistica, 2023). Although errors in surveys can be minimised through questionnaire design and careful fieldwork, errors in the collected data are inevitable (Kuhn, 2019).

There is also the underlying problem of the sensitiveness of the information being requested which is frequently present in the interview (Instituto Nacional de Estadística, 2017) and which results in high item non-response or an underestimation bias of income (Istituto Nazionale di Statistica, 2023). Another factor which influences the quality of the information on wages collected in household surveys, is the possibility of proxy respondents who can answer instead of the person directly concerned.

Linkage of survey data to administrative registries opens many research opportunities and may alter the way in which to conduct surveys (Kuhn, 2019). Despite the many challenges linked with administrative registers, there are clear advantages for using register information to measure income in the LFS rather than adding more questions to an already extensive questionnaire. The register information is more accurate when it comes to obtaining the exact amount in gross terms, compared to a telephone interview where people might be reluctant to answer questions on income or uncertain about the exact gross values (Statistics Norway, 2023). There is also a risk of people over-stating or under-reporting their income.

4 The CFR Administrative Source

For this study, administrative data from the Office of the Commissioner for Revenue (CFR) was used to compile the gross income for LFS employees. The variables included in this dataset allowed to better reconstruct the definition of the INCGROSS variable as described in the LFS explanatory notes.

Nevertheless, there were still discrepancies between the LFS and the CFR administrative data. To start with, the statistical unit in the administrative source was jobs not persons therefore a person can be included in the register multiple times either for different employments within the same year, for a single employer or for different employers, either consecutively or simultaneously (Statistics Belgium, 2023). A further inconsistency was that the register did not allow the possibility to distinguish between multiple jobs.

5 LFS record linkage with CFR administrative data

The quality of the INCGROSS variable depends on the coherence between the two sources. The starting point of the analysis consisted in linking LFS data to the CFR register using a personal identification number as the primary key. Over 90 per cent of all the records between 2018 and 2022 were linked with the register. Since the administrative source contains

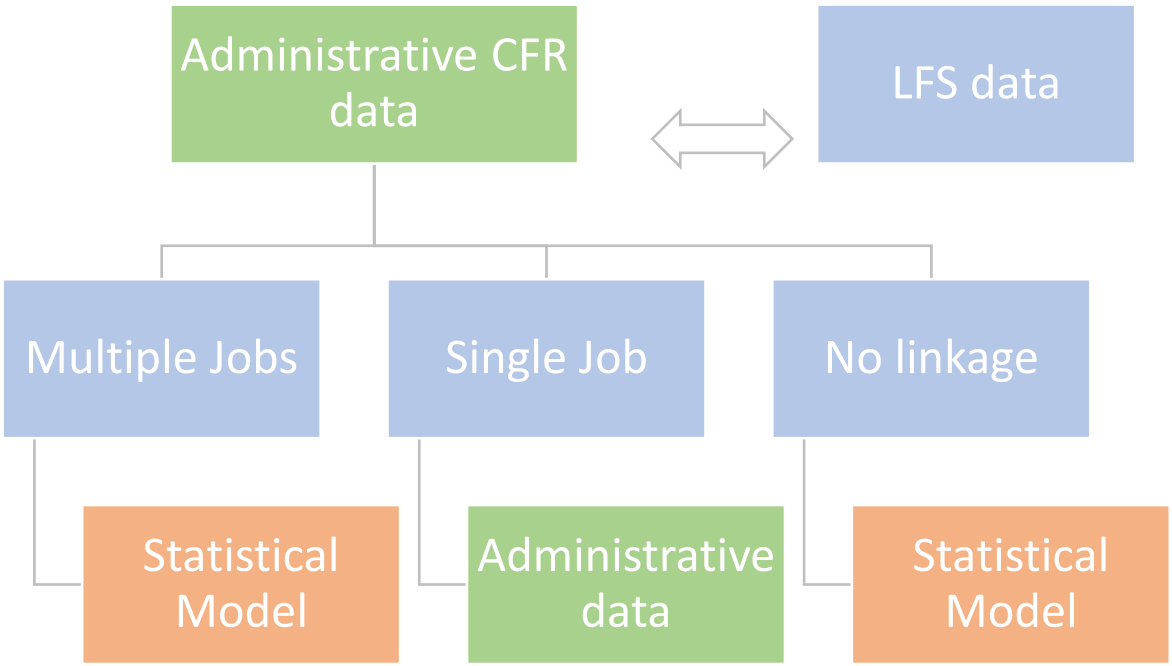
information on all jobs a person might have, between 10 to 15 per cent of the matched cases were linked with multiple jobs. Meanwhile, less than 10 per cent of all the records in the 2018 to 2022 LFS datasets were not linked with the CFR register due to a missing or incorrect identity card number in the LFS datasets, or the identity card number was not available in the administrative database.

6 First phase Imputation: Using the administrative source

The integration between a survey and a register typically generates issues such as missing linkage and linkage errors; under-coverage of the register; misalignment between the register and survey reference population (Statistical Office of the Republic of Serbia, 2023). Such issues must be considered and properly treated.

In this study, LFS records which linked once with the CFR data had the INCGROSS variable imputed using the figure available in the administrative source. Since the register did not allow for the distinction between different jobs, the INCGROSS variable for persons who were linked with the register multiple times, was imputed, using a statistical model. In addition, when no information was found in the CFR, the same imputation model was applied (Chart 1).

Chart 1: Imputation strategy for CFR and LFS linking scenarios



Adapted from "Italian strategy to obtain gross monthly pay for LFS employees" (Istituto Nazionale di Statistica, 2023, p. 2)

7 Second phase Imputation: Applying a statistical model

Initially, the INCGROSS data which was linked to the CFR register was analysed. A multiple linear regression was carried out in SPSS for the value of INCGROSS derived from CFR to detect outliers within the data. For this task, standardised residuals were used as diagnostics to examine extreme points that were dominating the regression and possibly distorting the results.

Linear regression assumes that the relationship between the response variable and the predictors is linear, therefore, the detection was carried out on the log of the INCGROSS variable. If residuals are normally distributed, then 95% of them should fall between -2 and 2. Residuals which fall above 2 or below -2, are considered unusual (UCLA: Statistical Consulting Group, 2021). A new variable 'Residual' was generated in SPSS indicating the vertical distance (or deviation) from the observation to the predicted regression line for each record. In this case, both low extreme values (< -2) and high extreme values (> 2), were removed from the INCGROSS variable and were imputed using a statistical model.

After removing all the outliers, the INCGROSS variable data was ready to be used in the statistical model for imputation. A generalised linear model with 5 imputations was applied. The imputation method is a maximum likelihood estimation with multiple imputation based on a Log linear Method using SPSS. The strength of this model derives from the large set of covariates (LFS variables), highly correlated with the dependent variable (INCGROSS). The covariates considered for the model were:

1. **Socio-demographic variables**: Age, Sex
2. **Background**: Highest level of education
3. **Job characteristics**: Occupation, Type of employment and Normal hours worked.

The independent variables selection was conducted through a stepwise procedure, where all the variables included in the model were significant. The imputation rate for the INCGROSS was consistent over the years, with an average of 20.2% from 2018 to 2022 (Table 1).

Table 1: Imputation Rate INCGROSS: 2018 – 2022 (%)

Year	2018	2019	2020	2021	2022
%	21.0	21.7	17.4	19.3	21.8

Subsequently, the gross annual income derived either from the CFR data or from the statistical model was divided by 12 to obtain a gross monthly figure.

8 Results

The IESS regulation enforces the transmission of the INCGROSS variable from 2021 onwards. Yet, in Malta the gross annual income was derived from 2018 with the intention to build a robust timeseries for comparability and consistency checks. Results indicated that the average gross annual salary in 2022 was estimated at €27,970, which is 8.2 percentage points higher when compared to the average of 2021 (Table 2).

Table 2: Gross average annual income: 2018 – 2022 (€)

	2018	**2019**	**2020**	**2021**	**2022**
Annual Gross Income (€)	21,834	23,641	23,745	25,842	27,970
% Change	-	*8.3*	*0.4*	*8.8*	*8.2*
Annual Basic Income (€)	18,773	19,594	18,907	19,755	20,947
% Change	-	*4.4*	*-3.5*	*4.5*	*6.0*

The result of this method was evaluated through comparisons between the obtained value from the CFR data source and the basic income variable collected in the MT-LFS questionnaire. The objective of such a comparison was to assess the coherence of both variables and to be able to explain divergences where applicable. As expected, the annual gross income was higher than the annual basic salary due to the difference in the definition as well as the tendency of respondents to underestimate their earnings when reporting them in surveys. Despite these differences, similar trends in growth rates were observed from 2018 to 2022 for both variables (Table 2). In addition, consistency checks by sex, occupational group, economic activity, and type of employment were carried out to ensure comparability over time. Significant year on year growth was observed for each occupational group, except for 2019 – 2020, due to the impact of the COVID-19 pandemic.

9 Conclusion

The new legal requirements for income information in the LFS provided a good opportunity to produce more data which meet user needs. The National Statistics Office (NSO) decided to keep collecting the basic income from the LFS CAPI/CATI interviews to satisfy national

requirements. Simultaneously, gross income was derived from the CFR administrative data and was made available from 2018 onwards. As a result, users have the possibility to choose the more appropriate income variable, depending on their research question (Statistics Austria, 2023). With this methodology in place, the INCGROSS variable has the advantage of being more accurate, has a larger coverage and imposes less burden on the respondents.

References

Esposito, L., Fioroni, L., & Guandalini, A. (2019). Gross Income Projection in Labour Force Survey Data. *Rivista Italiana di Economia Demografia e Statistica, LXXIII*(4), 41-52.

European Commission Eurostat. (2021). *EU Labour Force Survey Explanatory Notes (to be applied from 2021q1 onwards.* Luxembourg: Directorate F: Social statistics, Unit F-3: Labour Market and Lifelong Learning.

European Parliament, Council of the European Union. (2019). Regulation (EU) 2019/1700 of the European Parliament and of the Council. *Official Journal of the European Union*, 1-32.

Instituto Nacional de Estadística. (2017). *Dissemination of wage data from the Spanish LFS: anonymized microdata files.* Copenhagen: 12th Workshop on Labour Force Survey Methodology.

Istituto Nazionale di Statistica. (2023). *Italian strategy to obtain gross monthly pay for LFS employees.* Lisbon: 16th Workshop on Labour Force Survey Methodology.

Kuhn, U. (2019). Measurement of income in surveys. *FORS Guide No. 02, Version 1.0*, 1-13. doi:10.24449/FG-2019-00002

Statistical Office of the Republic of Serbia. (2023). *INCGROSS variable in Serbian LFS.* Lisbon: 16th Workshop on Labour Force Survey Methodology.

Statistics Austria. (2023). *Gross Monthly Pay in the Austrian LFS.* Lisbon: 16th Workshop on Labour Force Survey Methodology.

Statistics Belgium. (2023). *Methodology for the derivation of the INCGROSS variable in the Belgian LFS, using administrative sources.* Lisbon: 16th Workshop on Labour Force Survey Methodology.

Statistics Norway. (2023). *Income from main job: Expanding the use of register information in the Norwegian Labour Force Survey.* Lisbon: 16th Workshop on Labour Force Survey Methodology.

UCLA: Statistical Consulting Group. (2021, 12). *SPSS regression diagnostics*. Retrieved from Advanced Research Computing Statistical Methods and Data Analytics: https://stats.oarc.ucla.edu/spss/seminars/introduction-to-regression-with-spss/introreg-lesson2/#s0