STATBEL
Belgium in figures

Conference paper:

Methodology for the derivation of the INCGROSS variable in the Belgian LFS, using administrative sources

Astrid Depickere

Statistics Belgium

16th Workshop on Labour Force Survey Methodology

Lisbon, 25-26 May 2023

Introduction

Since the entry into force of the Integrated European Social Statistics Framework Regulation (IESS FR), some changes were made to the variable 'Income from work' included in the Labour Force Survey from reference year 2021 on (Regulation (EU) 2019/1700).

The new definition of the concept and the adapted transmission period (within fifteen months of the end of the reference period) now allow for the possibility to use administrative data for this particular variable. Given that both the quality of administrative databases as well as the access procedures and delivery times have improved a lot in Belgium, Statbel decided to examine the possibility to derive the variable from these administrative records. With the support of a Eurostat Grant, a feasibility study was carried out in 2019. The main purpose was to develop a methodology for the construction of the INCGROSS variable on the basis of the available administrative sources, to detect possible problem areas and to get insight in the remaining number of missing values and see whether an additional imputation method was needed (Statistics Belgium 2020).

The available sources

Two sources are available for information on Income from Work: on the one hand Social security data (DMFA) and on the other Personal income tax data (Belcotax/IPCAL).

The **Social Security data (DmfA1)** database is based on the quarterly declaration employers need to introduce to the social security Administration with the purpose of calculating social security contributions and determining social rights such as unemployment, pensions, family allowances, annual holidays, etc… It contains detailed information on wages and working times of all employees employed by an employer during a given quarter. The data are also quite quickly available, a first version after 5 months (T04), the final version after one year (T07).

A second source comes from the **Tax registers** and contains **provisional income tax** data (Belcotax). The dataset is constructed by the tax authorities on the basis of tax fiches which all debtors of income components need to send to the tax authorities. These fiches all together constitute Belcotax, and are used as prefill for the online tax declaration with 'tax on web'. So it means this dataset is available at the time of the start of the tax declaration i.e. 6 months after the reference year. One year later, a more complete version of the personal income tax data becomes available (IPCAL), based on the actual declared data. This is however too late to be used for the LFS INCGROSS variable.

The possible use of both sources in surveys had been examined thoroughly by colleagues from the Structure of Earnings Survey (SES) and the Survey on Income and Living Conditions (SILC) survey, so we could benefit from their expertise and knowledge of the databases when elaborating an approach for LFS.

Of these two sources, the social security data are the best suited to be used for the construction of the LFS variable INCGROSS. The variables included in this dataset are more specific and allow to better reconstruct the definition of the Gross Monthly Wage as described in the instructions for INCGROSS. Furthermore, the reporting period is on a quarterly basis, which makes it easier to match the information with the main job as reported by the respondent in the survey. The Belcotax data are still useful though, mainly because of its higher coverage. Some disadvantages of this dataset include the reference period of one year and the fact that the income data do not allow to distinguish between multiple jobs.

[1] Déclaration multifonctionnelle/multifunctionele Aangifte

Coverage of the two sources

The starting point of the analysis consisted in checking whether administrative data was available for all respondents of the survey for which the INCGROSS variable is needed. This was done by linking the survey data to the two available sources by using a unique identifier for each citizen[2]. Unfortunately, not all records in the LFS data could be matched to the administrative data sources. There could be several reasons for this. One is when it concerns a respondent for which we do not have a national registry number. This is a rather exceptional situation, but it can occur when persons have joined a household but could not be matched to the national register (yet), either because they are not registered (yet) or because the information given at the time of data collection was insufficient to link the person to an existing national register number. A second situation occurs much more often and consists of persons for which we do have a national register number, but no matching record was found in the administrative databases.

To better understand why information in administrative records is missing, we checked the number of missing cases for some specific types of workers, using some other variables of the survey, such as whether a person works abroad or whether it concerns a student or a disabled person (table 1, data for 2020). This was done separately for the two sources (first two columns) and also for the combination of both sources (last column).

Table 1: Missing values after linking LFS 2020 data to administrative sources of Social Security (RSZ) and Tax register (BTAX)

		RSZ/DMFA		BTAX		BTAX & RSZ/DMFA	
		missing	not missing	missing	not missing	missing	not missing
Student (mainstat)	N	42	175	23	194	23	194
	Row Pct	19,4	80,7	10,6	89,4	10,6	89,4
Disabled (mainstat)	N	185	61	77	169	77	169
	Row Pct	75,2	24,8	31,3	68,7	31,3	68,7
Working abroad	N	666	53	668	51	666	53
	Row Pct	92,6	7,4	92,9	7,1	92,6	7,4
NACE U (extraterritorial organisations and bodies)	N	184	31	176	39	125	77
	Row Pct	85,6	14,4	81,9	18,1	61,9	38,1
Temporary job	N	190	1120	139	1171	120	1190
	Row Pct	14,5	85,5	10,61	89,39	9,16	90,84
Absent from work	N	603	1876	272	2207	268	2211
	Row Pct	24,3	75,7	11,0	89,0	10,8	89,2
All	N	1.659	12.656	1.215	13.100	1.172	13.143
	Row Pct	11,6	88,4	8,5	91,5	8,2	91,8

For 8,2% of all persons in the LFS sample, we could not find any data in one of the administrative datasets. This percentage is however, much larger for people working abroad (92,6%) and for people working for extraterritorial organisations and bodies (NACE=U) (61.9%). This is not very surprising as these are two categories of workers who usually pay social security contributions and taxes in another country. We also see a higher number of missing values on the social security data among those that

[2] the Belgian social security ID number (NISS) For people registered in Belgium, this is the same as the national register number).

consider themselves as disabled, among students and among those that are absent from work (any reason) or those who have a temporary job. For these categories the situation gets better when we also take the tax data into account. This makes sense, as the reference period of the last database is one year whereas for the social security data it is one quarter.

Deriving the INCGROSS variable: priority rules

As mentioned above, we decided to give priority to the social security data as the main source for deriving the INCGROSS variable. In 2021, for 88% of all cases, a value for INCGROSS was obtained from the social security database. Additionally, for 3% of the cases, a value was obtained from the Personal Income Tax database. Finally, an imputation method was developed and applied for the remaining cases (i.e. 4% of all cases), except for people working abroad, which were left missing. All together this leaves us with a little less than 5% of all cases that do not have a value on the INCGROSS variable.

Table 2: Source used for estimating the LFS INCGROSS variable

	2019		2020		2021	
Source	**N**	**%**	**N**	**%**	**N**	**%**
MISS	793	4.57	666	4.65	656	4.79
BTAX	474	2.73	487	3.40	433	3.16
DMFA	15573	89.84	12656	88.41	12082	88.27
IMPU	495	2.86	506	3.53	517	3.78

Method applied to social security data (DMFA)

To be able to use the social security data, we needed to match the information obtained from the quarterly social security dataset to the main job reported for the reference week in the survey. For this, we used a similar methodology as the one which is used by our colleagues from the SES, where a very similar concept of Gross Monthly Wage is derived. This is done in two steps.

First, because the social security database is composed of employment lines rather than individuals, a single person can have multiple employment lines, either for different employments within the same quarter, for a single employer or for different employers, either consecutively or simultaneously. In this case, we have to select the employment line which we think corresponds to the main job reported in the survey. Four different scenarios exist:

1) There is only one employment line and the period to which it applies contains the reference week. In this case, we can assume it applies to the main job reported in the survey.

2) There are multiple lines within a quarter and these do not overlap. In this case, we select the line that contains the reference week.

3) There are multiple lines that contain the reference week and these overlap. In this case, the line with the highest salary is assumed to apply to the main job.

4) None of the employment lines contain the reference week. In this case, we select the employment line for the period that is closest to the reference week.

The second step is then to recalculate the salary to a monthly basis. For full time workers, this is done on the basis of the number of working days. For part-time workers, we do the same using the working hours.

Method applied to Belcotax data

As mentioned above, for a limited number of cases (3%), we did not find information in the social security database, although information on income from work was present in the personal income tax dataset. It mostly concerned persons that were absent from their job during most of the quarter in which the reference week fell, so we decided to use this information rather than applying an imputation method. An exception was made for persons working abroad, where the risk was considered too high to mistakenly use a value that applied to a small job rather than to the person's main job.

To construct the wage variable on the basis of the personal income tax data, we relied heavily on the work done by our SILC colleagues (De Schrijver A., 2020). First, a yearly income was determined by adding all income components. Next, we divided this by 12 in order to obtain a monthly figure.

A major drawback is that the dataset does not allow to distinguish between more than one job, so if a person has multiple jobs, then the calculated INCGROSS can be overestimated. At the same time, it seems rather unlikely that a person has multiple jobs for which no information was found in the social security database.

Imputation method

When no information could be found in any of the administrative data sources, we applied an imputation method using gender, level of education, age, region of work, profession, economic activity and parttime employment share as predicting variables. The imputation method is a maximum likelihood estimation with multiple imputation. We also used a log transformation to obtain only positive incomes. At first, we also applied this method to those working abroad, but some further analysis learned that this was problematic and lead to a systematic underestimation of wages of persons working abroad so we decided to leave these cases missing rather than imputing a value that is biased.

Results / Evaluation

In order to evaluate the result of our method, we tried to compare the obtained value with other sources that contain information on wages, such as the SILC survey and the SES survey. Another option was to look at the value of the 'old' wage variable in the LFS that was still collected in 2020. None of these comparisons are truly perfect, because either the population is different or the concept is not exactly the same. Furthermore, when differences are observed, it is difficult to make clear judgements about which source is closer to reality.

Nevertheless, we present some results here based on the comparison with the SES and the old LFS variable.

Comparison SES - LFS

We compared the new INCGROSS variable with some aggregated results of the Structure of Earnings Survey. For this purpose, we tried to delimit the sample to the one of the Structure of Earnings Survey Population, which only included enterprises with 10 or more employees and which does not include the NACE categories A, O-U.[3] Table 3 compares the median and mean values of the Gross Monthly Income in both sources, according to NACE sector and size class for employees working full-time.

[3] Some differences remain as for LFS we only have the size of the local unit instead of the enterprise. We therefore left out all local units with less than 10 employees, from both sources.

Overall, the estimated value of INCGROSS seems to be lower in LFS than in SES, both when we look at the mean as well as the median. If we look at the size class, we see that, the larger the local unit, the closer the two values. There are different explanations for this but the most obvious one is the fact that the results of the SES apply to October as a reference month, whereas LFS applies to the whole reference year 2020 and is therefore more affected by absences due to e.g. illness, temporary unemployment etc.

Table 3: Comparison gross monthly wage in LFS – SES 2020 (reduced sample, full time working employees)

Analysis Variable : INCGROSS_				SES			Diff SES vs INCGROSS	
(NACE Rev. 2, 1 digit)	**N Obs**	**Mean**	**Median**	**N**	**Mean**	**Median**	**mean**	**median**
Mining and Quarrying	13	**4.610**	3.339	220	4.367	3.900	106%	86%
Manufacturing	1.404	**3.988**	3.293	25.930	3.905	3.637	102%	91%
Electricity, Gas, Steam and Air Conditioning Supply	77	**5.538**	5.670	896	5.316	5.179	104%	109%
Water Supply; Sewerage, Waste Management and Remediation Activities	93	**3.430**	3.210	1.186	3.926	3.706	87%	87%
Construction	505	**3.170**	2.904	7.882	3.372	3.293	94%	88%
Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles	801	**3.278**	2.846	20.220	3.643	3.354	90%	85%
Transportation and Storage	603	**3.203**	2.892	10.287	3.437	3.195	93%	91%
Accommodation and Food Service Activities	138	**2.479**	2.371	1.374	2.789	2.789	89%	85%
Information and Communication	347	**4.322**	3.965	5.487	4.620	4.395	94%	90%
Financial and Insurance Activities	307	**5.304**	4.722	4.353	5.071	4.874	105%	97%
Real Estate Activities	35	**4.002**	3.469	515	4.501	4.116	89%	84%
Professional, Scientific and Technical Activities	311	**3.855**	3.411	6.855	5.260	4.909	73%	69%
Administrative and Support Service Activities	401	**2.713**	2.601	17.234	3.348	3.228	81%	81%
Size of local unit	**N Obs**	**Mean**	**50th Pctl**	**N**	**Mean**	**50th Pctl**	**mean**	**median**
10-19	640	**3.028**	2.791	6.239	**3.426**	3.239	88%	86%
20-49	980	**3.244**	2.902	17.216	**3.576**	3.361	91%	86%
50-249	1.843	**3.428**	3.008	41.470	**3.793**	3.551	90%	85%
250-499	541	**4.079**	3.465	19.011	**4.015**	3.671	102%	94%
500 +	1.031	**4.652**	3.877	18.503	**4.401**	4.083	106%	95%

Comparison INCGROSS and the old Wage variable (INCDECIL)

Another interesting way of looking at the INCGROSS variable is by comparing the result to the old wage variable (INCDECIL & Q100 def being the net monthly wage). Unfortunately the concepts are no longer the same, as INCGROSS is a gross wage and the old variable was a net wage. We therefore had to convert the INCGROSS variable to a net value to be able to compare the two values.

Another method was to construct a categorical variable similar to the old INCDECIL variable and looking at the difference between both categorical variables. Table 4 contains both variables,

categorized into quintiles. By crossing both categorical variables, we can see how many individuals are classified differently according to both income variables.

About 50% of all observations is classified within the same quintile category. When we used deciles, this was only 30% of all observations. Furthermore, 35% of all persons fall into a cell with only one quintile difference between the two. And for 14%, the difference between the two is large, with a difference of more than 2 quintiles. When looking at the sign of the differences, we cannot say that there clearly is a under- or overestimation in one of the two variables. Overall, for 26% of the observations, the quintile_INCGROSS variables is higher, compared to 23% where the quintile_Q100def is higher.

Table 4. Comparison of old and new LFS wage variable, using a categorical variable (Quintiles)

	QUINTILE_q100def					
	1	**2**	**3**	**4**	**5**	**Total**
QUINTILE_INCGROSS						
1	**1741**	482	322	157	117	2819
2	633	**1139**	581	220	154	2727
3	141	709	**1136**	417	252	2655
4	69	229	823	**1081**	463	2665
5	20	56	168	650	**1889**	2783
	2604	2615	3030	2525	2875	13649

Frequency Missing = 666

Same quintile	**6986**	**51%**
Adjacent quintile, INCGROSS > Q100def	2815	21%
Adjacent quintile, INCGROSS < Q100def	1943	14%
More than 1 quintile difference, INCGROSS > Q100def	683	5%
More than 1 quintile difference, INCGROSS < Q100def	1222	9%
	13649	100%

Conclusion

As already mentioned, clear judgements about the quality of the variables cannot be done on the basis of these comparisons. The 'old' LFS variable certainly had several limitations, such as a high number of missing values (imputations were done on the basis of SES) and very imprecise measurement of the concept, especially in the case of proxy answers. Overall, we believe the new INCGROSS variable has the advantage of being more accurate, having a larger coverage and imposing less burden on the respondents.

The main limitation of the method lies in the high number of missing values for people working abroad and a potential bias for people working for extra-territorial organisations. An option that is still under investigation is to see whether a question would be added in the questionnaire for one or both of these specific populations.

References

De Schrijver A. (2020). Fiscal data in the Statistics on Income and Living Conditions (SILC) survey: a path for the future?. Brussels: Statbel - Statistics Belgium, FPS Economie (https://statbel.fgov.be/sites/default/files/files/documents/Analyse/EN/Analyse%20SILC-Fiscal%20data.pdf)

Statistics Belgium (2020). Report on Sub-action 4: methodological report. Calculating INCGROSS using administrative data. Grant ESTAT-PA7-S-2018/B2978-2018-LFS-Quality-Breaks