Statistical Office of the Republic of Serbia

Paper to the 16th workshop on Labour Force Survey methodology

INCGROSS variable in Serbian LFS

Milijana Smiljkovic

Abstract

The new redesigned Labour Force Survey (LFS) methodology, implemented from 2021, introduced several changes, including the inclusion of the variable INCGROSS as a mandatory annual variable. This paper aims to present the process of generating the INCGROSS variable in the Serbian LFS. It will provide details on data collection, including the source and imputation methods used to address item non-response. As the questionnaire primarily collects data on net income, the paper will also describe the net-to-gross conversion model applied. Lastly, preliminary data on gross monthly pay from the main job obtained from the LFS will be presented, along with a comparison to gross earnings data derived from the Regular Monthly Survey on Earnings, which is based on administrative data from the Tax Authority.

Introduction

Until 2021, net income was also collected through the LFS questionnaire; however, due to a significantly low item response rate, data on net income were never published. Starting from 2021, the practice of collecting net income data was continued because respondents found it easier to provide precise answers when reporting net values rather than gross values. The national taxation system facilitates the conversion of net to gross values, as the corresponding taxes and contributions can be calculated based on the type of employment contract.

What is collected in the LFS questionnaire?

Concept and definition used do not differ from the description in the regulation. Variable in the questionnaire refers to the actual net monthly payments, followed by detailed explanation by the interviewer to the respondents what else should they include in monthly pay. Monthly income includes regular overtime, extra compensation for shift work, seniority bonuses, regular travel allowances and per diem allowances, tips and commission and compensation for meals in cash, payments made on a higher than monthly periodicity (e.g. yearly or quarterly payments such as 13th month or holiday pay) proportionally included in the monthly pay. The variable collects payments received in the calendar month preceding the reference week.

In Serbian LFS questionnaire, there are two questions related to net monthly pay. The first question directly asks for the exact amount of the salary. If the respondent doesn't know or refuses to provide a direct answer, the second question offers income bands as option.

For the second question, nine income bands are proposed. There are also two open replies to cover lower and higher salaries. The first band is defined using the national minimum wage, rounded to the nearest thousand in national currency (which in Euro is hundred). This minimum wage level typically corresponds to the 3rd percentile based on the National Monthly Survey on Earnings, which utilizes Tax Administration data. The last band represents the rounded (to the nearest thousand in national currency) 99th percentile from the National Survey on Earnings.

In the National Earnings Survey, salaries and wages are calculated or expressed on a full-time equivalent (FTE) basis, and outliers are already treated. The proportion of outliers remains relatively stable over time, at approximately 3%. Bottom outliers are individuals with average hourly earnings lower than 35% of the average earnings for the previous 12 months, while upper outliers are those with average hourly earnings 7 times higher than the upper limit prescribed by the law for contribution payments. The highest basis for contribution payments is set at 5 times the average salary for the previous 12 months.

For closed intervals, the middle point of the interval is taken.

Response rate for direct question on exact amount of salary (INCSUM) in 2021 amounted 26.6%, and for the second question (INCDECIL) response rate amounted 27.3%. The response rate to these questions is very low, below 60%. For that reason, the imputation from the administrative sources is necessary, having in mind that according to the new Regulation imputation is needed when non-response is more than 5%.

Imputation methods in case of item non-response

Imputation of gross monthly pay is conducted using the Earnings Register, which is established based on income data from the Tax Authority (TA). The integration process between the Labour Force Survey (LFS) and TA data was implemented using the Python programming language. Record linkage techniques, with the assistance of libraries such as Record Linkage Toolkit, Pyodbc, numpy, and pandas, were utilized. Approximately 54% of missing earnings data in the LFS, which corresponds to around 25% of all employees, are imputed using gross earnings data obtained from the Tax Authority.

If it not feasible to impute the value from the Tax records, such as in cases of informal employment, an average value of earnings is imputed based on the occupation and activity at the two-digit level. This is done by considering data from both the tax records and the LFS, following specified criteria. This method allows for the imputation of 44% of missing data for the variable INCGROSS, which represents approximately 20% of all employees who need to respond to the salary question.

For less than the 2% of missing earnings data imputation could not be performed, that makes 0.9% of all employees who receive a salary. This group, i.e. the nonresponse consists mostly of those employees employed abroad and for whom it was not possible to impute earnings or perform the conversion of net to gross in case they answered the question about net income.

Net-to-gross conversion

The national taxation system allows us to convert net to gross value quite easy, because the corresponding taxes and contributions can be calculated based on the type of employment contract. In the Serbian LFS starting from 2021, a new national variable called TEMPUG was introduced to assist in the conversion of net to gross earnings. By combining the responses from four questions: STAPRO, TEMP, PREDRAD (also national variable), and TEMPUG, it is possible to distinguish between four types of employment contracts that are taxed differently according to our taxation legislation. The fifth group consists of informal employees, for whom net earnings are equal to gross earnings (i.e., INGROSS).

Preliminary results - testing the method for detection of outliers

Data on INCGROSS variable shown in this paper are related to the 2021 and still be consider preliminary because are not yet published as well as because it not yet decided how outliers will be treated, and whether they should be treated at all?

For outlier detection for INCGROSS is used the formula for SPSS extreme outlier, also given and recommended to be tested on INCGROSS data in the document *LFS_CONVAL.*
Outlier detection for INCGROSS:

Q1 and Q3 are the 1st and 3rd quartiles

Lower boundary: EXP [log Q1 – 3*(logQ3 – logQ1)]

Upper boundary: EXP [log Q3 + 3*(logQ3 – logQ1)]

After the implementation of this formula for detection of outliers the following results are obtained. The lower boundary obtained from the formula resulted in a negative value, which is not meaningful since income cannot be negative. On the other hand, the upper boundary amounted to 210,824 RSD, which corresponds to the 98th percentile of the National Monthly Survey of Earnings based on administrative (Tax) data.

The preliminary results on INCGROSS variable from LFS in comparison with Earnings Survey show that data on gross monthly pay from the main job from LFS (both, treated and no treated outliers) is lower than the data from administrative data on Earnings statistics. The results are expected having in mind difference in the coverage (LFS cover informal employment) and the tendency of respondents to underestimate their earnings when reporting them in surveys.

Comparing the average earnings from the Earnings statistics and the LFS, it was found that the average gross salary from the Earnings statistics is 15% higher than the LFS data without treating outliers. When the outliers were treated, the difference increased to 25%. (Table 1).

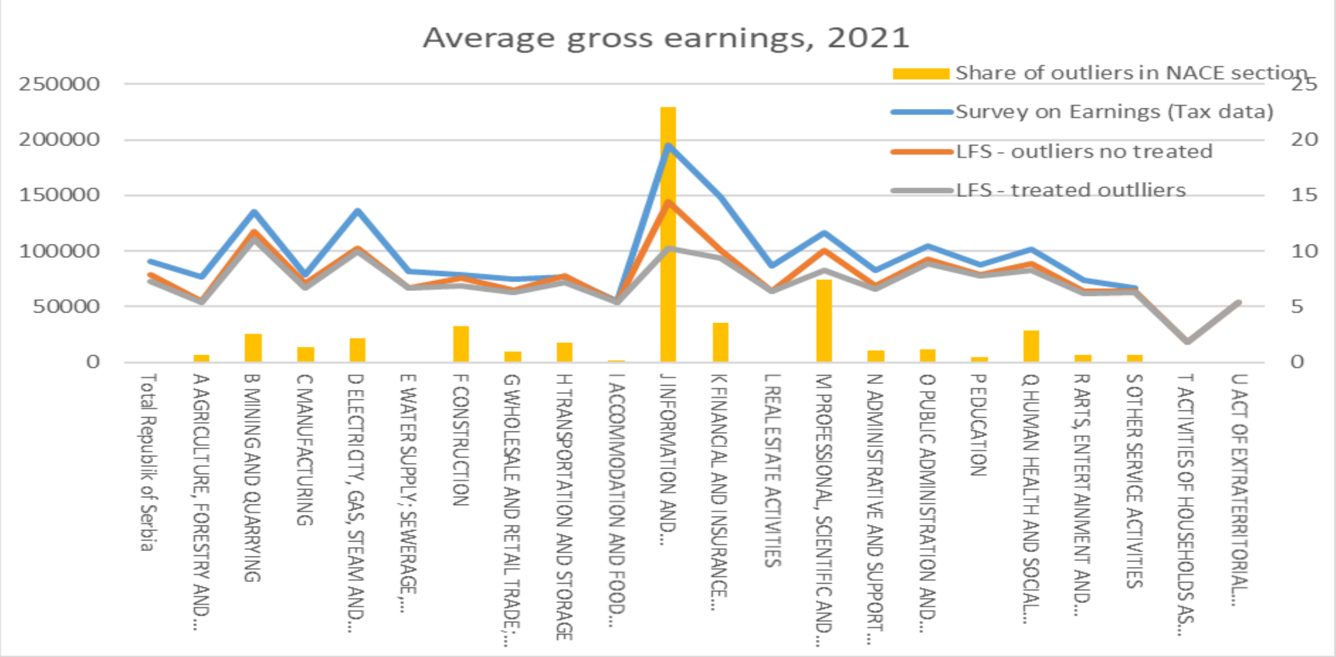
Table 1. Average earnings in national currency from Earnings statistic and LFS, 2021.

Republic of Serbia - total RSD

Average gross salary (Survey on Earnings)	90 784
INCGROSS (LFS - no treated outliers)	78 441
INCGROSS (LFS - treated outliers)	72 572

In the Graph 1. is shown the comparison of average gross earnings from Survey on Earning, LFS – with treated outliers and LFS – without treated outliers, by section of activity (NACE), 2021. In three NACE sectors (E, T and U), before and after the detection and removal of outliers, there was no difference in the INCGROSS variable. In majority of sectors (A,B,C,D,G,F,H,I,N,O,P,Q,R,S) the difference between the INCGROSS variable with treated and without treated outliers ranges from 1 to 11%. However, in the sectors J and M, the difference is significantly greater. The greatest difference in INCGROSS variable was noted in the sector J Information and communication, and amounted 40%, the second sector with significant difference was M Professional, scientific and technical activities, 22%.

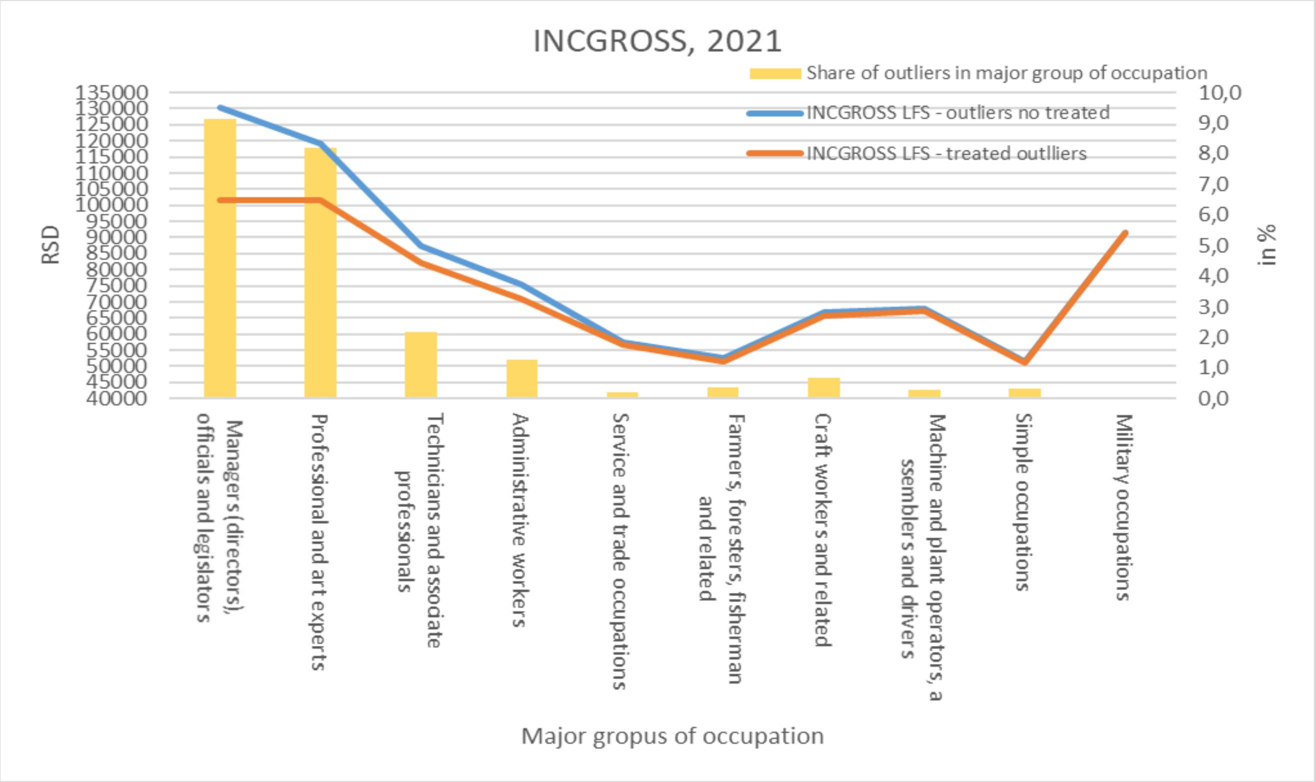
Graph 1. Comparison of average gross earnings from Survey on Earnings, LFS – with treated outliers and LFS – without treated outliers and share of detected outliers, by section of activity (NACE), 2021.



Average earnings in sector Information and communication according to the data of the Survey on Earnings based on administrative sources is more than twice then the average earnings in Serbia, while the sector M Professional, scientific and technical activities is in the first five sectors with the highest salaries. The LFS variable INCGROSS in these two sectors have the largest share of the outliers within the sector. The share of outliers in sector J amounts 22.9%, which means that almost every fourth employee in this sector should be excluded because they receive wages that are considered as outlier. In sector, M 7.4% of employees should be excluded in the calculation of INCGROSS variable if we apply mentioned criteria for detection of outliers.

Further analysis revealed that the outlier detection method seemed biased and inappropriate. In the Graph 2. is shown the comparison INCGROSS variable – with treated outliers and without treated outliers, by major groups of occupation, for 2021. The major share of outliers are in the first two groups of occupation which naturally have the largest earnings. The greatest difference in INCGROSS variable before and after the detection and removal of outliers is in the first group (Managers/directors, officials and legislators) and amounts 28%, while in the second group (Professional and art experts) amounts 17%. In next two groups (Technicians and associate professionals and Administrative workers) the difference is 6%, while in other groups the difference is almost negligible, 1-2%. In Military occupation there is no difference.

Graph 2. INCGROSS variable (with treated outliers and without treated outliers) and the share of outliers, by major groups of occupation, for 2021



The greatest share of employees detected as outliers by groups of occupation is first two groups (Managers/directors, officials and legislators and Professional and art experts), 9.1% and 8.9, respectively. Total number of outliers is 50,000, which is 2.3% of total employees. Of that total number of outliers (50,000) 32 000 i.e. 64% outliers belongs to the group of occupation Professional and art experts which is the main reason for doubting the justification of the used criterion for the detection of outliers.

In the Table 2. are shown the data about the number of outliers by the type of response. The majority of outliers come from the data that are imputed from administrative sources where the outliers are already treated in some way. In addition, we can say that in case of interval response the outlier are treated by defining the lower and upper limits of the bands. The other type of imputation is also done by using the averages. Only 10.7% of outliers come from direct answer.

Table 2. Number of employees where INCGROSS > detected outlier (210 824 RSD)

	Count num. of employees	Share (%)
INCDECIL - interval response	5206	10,4
INCSUM - direct response	5354	10,7
Imputed net income for item non-response from the average (NACE and occupation at 2- digit level)	10198	20,4
Imputed gross income for item non-response from administrative sources (INCGROSS_F 25)	29340	58,6
Total	50098	100,0

Conclusion

Overall, the preliminary results highlight the challenges and dilemmas associated with outlier detection. It is important to consider the limitations of the methods used and further examine the appropriateness of outlier detection criteria.