Imputing earnings data into Hungarian LFS from administrative sources – An attempt

Judit Dobszay, HCSO

Introduction

Over the past five-six years, the Hungarian Central Statistical Office (hereinafter: HCSO) has begun to redesign one of its major household type surveys i.e. HU-LFS. As well as making significant changes to the data collection instrument and methods, HCSO is also considering how to better process the data once it is collected. In particular, imputation methods have come under scrutiny in case of - gross or net value of - monthly take home pay. The main reason was, that this is the only variable in the survey, which could be refused and the refusal rate is very high, almost 60%. Our aim was to obtain reliable gross value of monthly take home pay (INCDECIL) by developing a new pay-imputation method based on administrative sources.

In order to achieve this goal, the following tasks have been planned and performed:

- Studying literature concerning missing data and imputation methods
- Selection of potential data sources for imputation
- Determination of methods and variables used for imputation of gross value of monthly take home pay (INCDECIL)
- Imputation process based on data sources chosen
- Comparison and assessment of the results, decision on the "best method"
- Production of consistent time series of gross values of the monthly take home pay by back calculations using the best method
- Developing a conversion program for gross/net value calculation based on the actual personal taxation and national insurance contributions rules

The frames of our work were assured by Grant Project: **Quality improvement for the Labour Force Survey** – Action 1, ***Obtaining gross values for the revised INCDECIL variable with register information and/or net-gross conversion*** – Sub-action 1

1. Potential data sources and variables for imputation of INCDECIL variable

1.1. The Hungarian Labour Force Survey (HU-LFS)

The Hungarian Labour Force Survey, introduced in 1992, provides quarterly information on population living in private households. Socio-demographic data are collected without age limit, however, questions on the labour market position relate to persons aged 15-74. The concepts, definitions and procedures used are in harmony with the ILO definitions and ESS regulations. Only those household members are interviewed in the target population, who are contributing to the common income-consumption.

In spite of the fact that data series of earnings are not coherent with macro statistics, the HU-LFS remains the most regular, coherent, harmonised, reliable and used source of basic information on Hungarian labour market.

1.2. The Hungarian Structure of Earnings Survey (HU-SES)

The other possible data source to impute earning into HU-LFS is the annual Hungarian Survey on the Remuneration of Individuals, which has been conducted for more than 30 years and is similar to the Structure of Earnings Survey. The requirements of the SES were fulfilled by a few modifications and the development of the original survey in 2002 and since then all surveys have been conducted according to the EU requirements.

Businesses with legal entities and other businesses employing at least 5 persons, all budgetary organizations, and selected non-profit institutions are covered by the survey. Sample size is about 13.6 thousand institutions from the government sector, 3 thousand non-profit and 15 thousand business units, altogether 31.6 thousand units. The reference month is May, however – instead of annual earnings –, regular monthly earnings and bonus earnings of previous year are observed.

Data in HU-SES is available on individual level. Beyond detailed information on remuneration of a selected employee, information is gathered also for working time and for other work characteristics. Demographic criteria and data on educational attainment help to use them for imputation purposes into HU-LFS.

1.3. Monthly Contribution Declaration (HU-MCD)

Administrative records are data collected for the purpose of carrying out various non-statistical programs. They – as is the case of data of Monthly Contribution Declaration (HU-MCD) datasets of National Tax and Customs Administration of Hungary – are maintained to administer contribution payment obligations for individuals or for businesses. Using them for statistical purposes continues to grow and presents a number of advantages. They already exist, so do not incur additional costs for data collection and do not impose further burdens on respondents.

On the one hand HCSO intends to develop a sustainable method to replace data collection ”Monthly Labour Report” on earnings, i.e. to use in the future contribution declaration data for businesses in lieu of seeking survey data for them. On the other hand – and it was one of our task in the framework of this application – use earnings information in imputation process and survey evaluation.

Every employer in Hungary must produce a declaration on the taxes, contributions, and other relevant data related to the disbursements and benefits for individuals on a monthly basis in order to account for the advance payments of the social security services, of the social contribution taxes and of the personal income tax.

In the declaration files data are provided per individual on the taxes and contributions of the disbursements and benefits for the individual. Beyond the disbursements, important information are to be provided regarding the work as well.

2. Description of the workflow

2.1. General issues

Records of public workers and employees with minimum wage/guaranteed wage minimum should not be taken into account during the imputation process. The exact value of their earnings, as lump sums are known.

Other employees - regardless of whether the file contains earnings data for them or not - were labelled as “for imputation”.

We used real-donor imputation by statistical matching process. As the number of observations in the donor sets increased that of the HU-LFS more than tenfold we applied repeated selection for both cases. Key variables have been created based on common variables of both HU-LFS and the relevant administrative datasets.

The weighted average earnings of the actual key groups calculated in the donor dataset served as the imputed earning value for the matching receiver item. For those records of HU-LFS, non-matching in the actual iteration step, a next one with less strict filter conditions was applied. This iteration process was continued – step by step – until each of our recipient records has received a value, as gross monthly earning.

The actual imputation was executed by means of SAS programs.

2.2. Imputation from HU-MCD datasets

Monthly MCD datasets contain 4.5 to 5 million records, run time is significantly longer than in the case of quarterly HU-LFS or annual HU-SES ones. We decided to impute monthly HU-MCD data into the monthly HU-LFS dataset based both on the same reference period.

The key variables composed for linking were defined as follows:

key1= FM/PT || Occup_4digit || Econ_1digit || Age_5 || Gender

key2= FM/PT || Occup_4digit || Econ_sect || Age_5 || Gender

key3= FM/PT || Occup_4digit || Econ_sect || Age_10 || Gender

key4= FM/PT || Occup_3digit || Econ_sect || Age_10 || Gender

key5= FM/PT || Occup_3digit || Age_10 || Gender

key6= FM/PT || Occup_3digit || Gender

where:

FM/TM – full time / part time

Occup_4digit – occupation according to the Hungarian classification HSCO08 on 4 digits

Occup_3digit – occupation according to the Hungarian classification HSCO08 on 3 digits

Econ_1digit – industry according to NACE on the first digit

Econ_sect – industry according to NACE, grouped as follows: ‘A’, ‘B-F’, ‘G-N’, ‘O-U’

Age_5 – age groups as follows: 15-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-74

Age_10 – age groups as follows: 15-24, 25-29, 30-39, 40-49, 50-59, 60-64, 65-74

Almost 98% of the receiver records were matched by a donor record via the first key.

2.3. Imputation from HU-SES datasets

In the case of imputing from annual HU-SES data (including about 685 000 records) we created first quarterly datasets by adjusting gross earnings according to the relevant quarterly / annual ratios of official institutional earnings statistics.

Regarding the linking variables, two further information could be taken into account during the imputation process, compared to keys for HU-MCD. Educational attainment and region of place of work are factors affecting the level of earnings significantly.

The key variables composed for linking were defined as follows:

key1= FM/PT || Occup_4digit || Econ_sect || Region || Educ || Age_5 || Gender

key2= FM/PT || Occup_4digit || Econ_sect || Region || Educ || Age_10 || Gender

key3= FM/PT || Occup_3digit || Econ_sect || Region || Educ || Age_10 || Gender

key4= FM/PT || Occup_3digit || Econ_sect || Region || Age_10 || Gender

key5= FM/PT || Occup_3digit || Region || Age_10 || Gender

key6= FM/PT || Occup_3digit || Region || Gender

key7= FM/PT || Occup_3digit || Gender

where:

Region – counties grouped as follows: Budapest, Pest county + Western and Central Transdanubia, other counties

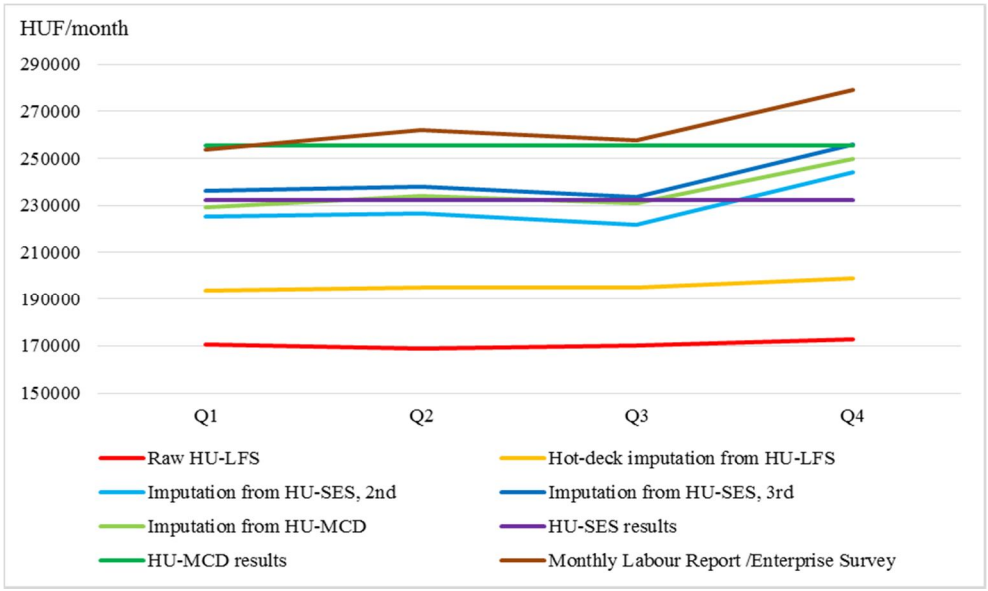
Educ – education level as follows: lower education, secondary education without final exam, secondary education with final exam, higher education

About 80% of the receiver records were matched by a donor record via the first key.

2.4. Results

Regarding the results we may state that the dataset imputed by data from own, i.e. HU-LFS dataset caused a small increase in the average earnings only, while by imputation from HU-SES and HU-MCD datasets the earnings level increased significantly. Results from these two sources ensured similarly big increase, and the seasonal effects showed similar patterns too. The highest level of average gross earnings was obtained in enterprise survey of HCSO (Monthly Labour Report) and the second highest average gross earnings derive from HU-MCD. Earnings based on imputation were the highest in case of imputation from the HU-SES, where looking for a donor the educational attainment was also taken into account (imputation from HU-SES, 3rd).

Figure 1. The level of gross earnings by different calculation methods, 2016 Q1–Q4



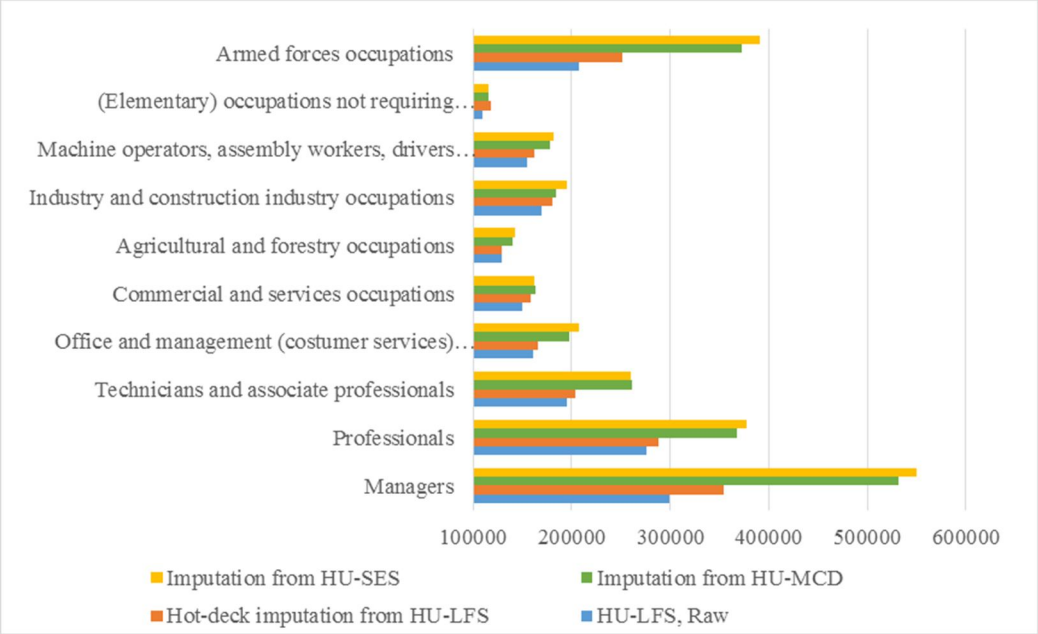
As the result of hot-deck imputation i.e. imputing from own (HU-LFS) dataset, the average earnings reached a higher level in each major occupational groups, except from agricultural and forestry occupations. Nevertheless in those occupational groups, where low, in high proportion minimum wage is typical, there is only a small difference between the original data and the imputed one.

In case of imputation from SES, the earnings as the outcome of the imputation differ more significantly from the original. In case of persons with higher status who represented a higher proportion in the non-response, the increase of the average earnings happened not only with a

higher rate but to a higher extent too. Similarly to data imputed by hot deck method, we experienced, that the increase in earnings of those with low earnings is much smaller than the average. Persons belonging to low earners are employed typically in occupational groups of services (catering, tourism), and in occupations not requiring qualifications (construction industry).

The differences in results coming from the two external data sources, i.e. imputed from HU-SES and HU-MCD ones are surprisingly negligible but we can state that the method taking the educational attainment into account causes somewhat higher earnings.

Figure 2. Average gross earnings by Major Groups of Occupation, 2016 Q1



Finally, it was important to carry out the assessment of changes in earnings of those whose reported earning values were replaced by imputed ones. The results of the imputation/replacing are as expected, i.e. the difference between the reported and the imputed values among managers is outstanding and for those working in occupations requiring higher educational attainment also significant. The use of imputed earnings in this group proved to be justifiable, with continuous monitoring.

For the graphical representation of the distribution of earnings based on different imputation methods and data sources, Lorenz Curves were made and Gini coefficients were measured. The HU-MCD imputation resulted in the highest inequality (Gini: 30%), while the raw data proved to be absolutely inappropriate to reflect the real inequalities in distribution of earnings mostly because of missing data, the value of Gini is 14%. Hot-deck imputation gave more plausible results (24%), but the value of Gini index based on imputed value from HU-MCD (30%) is closer to that of from EU-SILC which is around 28% for years (taking into consideration that EU-SILC results refer to the income and the total population).

2.5. Time series

Comparing imputed results based on three different data sources, and taking into account the important quality issues as accessibility and timeliness – not forgetting the other ones – we decided to continue the work using MCD datasets. We produced time series on earning for

years 2014-2017 which are in line with macro statistics and could serve as a good basis for the future data transmission concerning earnings in HU-LFS.

3. Conversion program for gross →net value calculation

A gross-net conversion model for the HU-LFS is of outstanding importance, because net incomes are important for analysing issues of income distribution.

The gross salary is what employers usually agree on with their employees, and what the labour contract includes. Employees have to pay taxes and contributions from that, which are deducted by employers and forwarded to the authorities for their employees. This way only the net salary is paid to the employees.

In Hungary, there exists a flat tax and contribution system. The personal income tax rate equals to 15% since 2016, whereas the figures of pension, health insurance and labour market contribution amount to 10, 7 and 1.5% respectively. Besides there is a tax allowance given via a family tax allowance. The family tax benefit is relevant for employers and employees, as it increases the net salary by reducing the employee's personal income tax and contributions. Based on the number of children, the parent's net salary can raise by HUF 10 000, HUF 17 500, HUF 33 000, or even more (ca. EUR 33, 60, 110, respectively). The allowance can be split between spouses or life partners.

4. Conclusions

Revised information on the take home pay from the main job, i.e. the revised INCDECIL variable of HU-LFS is expected to be a priority for many users not only in Hungary, but also at international level. This is primarily due to the fact that earnings data collected through the survey has never been published because of the poor quality. The item non response was continuously very high.

The aim i.e. to have the take home pay from the main job based on gross values, seems to be available.

According to our investigations, the Monthly Contribution Declaration (HU-MCD) datasets of National Tax and Customs Administration of Hungary proved to be the best source for imputing earning data.

Selection of imputation procedure and determination of matching variables (sex, age, occupation, full-time/part-time employment, regions) used for selection of donor were based on the relevant literature and preliminary investigations.

As a result of the imputation procedure carried out, appropriate datasets were available to calculate consistent and coherent time series back to 2014 which is in line with macro statistics.

As agreed in the application also a gross/net conversion program were developed for gross/net value calculation based on the actual personal taxation and national insurance contributions rules.

The results could serve as a good basis for the future data transmission concerning earning in HU-LFS. However, sustainability of the project's achievement depends on the availability and quality of the used data sources.