Incorporating administrative data in monthly LFS estimation of economic activity in Hungary

16th Labour Force Survey Methodology Workshop

Réka Bárkai, HCSO
Viktória Magera, HCSO

Introduction

Due to rising user demands, notably following the rapid economic changes brought on by the COVID-19 epidemic and their effects on the labour force, the question of providing monthly LFS data has become imperative for the Hungarian Central Statistical Office. With the addition of the EU legislation urging member countries to provide data on monthly unemployment rates to Eurostat, it has become clear that an accurate method must be found to estimate monthly economic activity data. The Hungarian LFS is a continuous data collection, the quarterly sample of which consists of three statistically independent monthly sub-samples with large-scale fluctuation. Because of that, the reliability of monthly data is low, so we had to create a method which has to follow the economically explicable monthly changes while not being distorted by the fluctuation of the independent monthly samples. In order to satisfy these needs, with the help of administrative data the HCSO started to develop a monthly estimate for employment and unemployment numbers in Hungary.

Administrative data sources

While exploring the available administrative data sources, we had to keep in mind that there were certain criteria to be fulfilled in accordance with the legislation regarding monthly unemployment rates: since the final data had to be available in certain sex and age subgroups, we needed administrative data available in these subgroups as well, for it to sufficiently guide the model estimation. The data used had to be available preferably in a long time-series, as we knew back-calculations were necessary for Eurostat, and it was obvious that observing a longer time period could help us ensure the lasting reliability of the estimation process.

Additionally, a challenging aspect that limited our choices was timeliness, as we needed monthly data that was also available on time for our first release (which usually falls near the end of the month following the reference month), so it can include the model estimation results. This narrowed down our options to the tax records of the National Tax and Customs Administration for the estimation of employed persons, and registered jobseeker records of the National Employment Service for the estimation of unemployed persons.

Employment

For the estimation of the number of employed persons we ended up using a combination of different tax records; employee records, self-employed records and so-called "small taxpayer" records.

In terms of the aforementioned demographic dimensions, the data we had was available in suitable age and sex subgroups, however the processing of the records takes relatively long (almost two months following the reference month). To combat this, another administrative data source had to be included, which records the monthly inflow and outflow of citizens into the Hungarian social security system through employment. We used these records as supplementary data to predict tax data in advance. This prediction is later revised with the actual tax data, once it is available, but through preliminary research we have found that the prediction of supplementary data on tax data is quite accurate.

While there is obviously a sizeable overlap in the LFS and the tax records, we can presume that there are certain limits to the two data moving together as there are differences in the populations and concepts covered between the two data sources. LFS only covers residents of Hungarian private households, so people living in certain institutions or abroad – but still working in Hungary – can appear in the tax records, while they are not part of the observed population of LFS. On the other hand, certain groups could be missing from administrative data but be observed by LFS, such as residents of private Hungarian households that work abroad.

Aside from the conceptual differences of the observed groups, we also had to take into consideration the conceptual differences of the indicators used, as LFS uses average monthly estimates while administrative data gives us the, exact number of people in the registry in the reference month. This can become particularly troublesome while accounting for short-term, seasonal jobs.

Unemployment

For unemployment data, the conceptual differences ended up being more prominent in proportion to the employed. Finally, we decided to use registered jobseekers' data to guide the unemployment estimate of LFS, however a considerable group of the people in the registry might not satisfy the ILO concept of unemployment. A prime example of this is people being able to do casual work, which the National Employment Agency allows people to do while being registered as jobseekers, however a couple hours of work can already constitute as employment by ILO terms. Additionally, one of the most prominent groups of this conceptual disconnect includes people merely being registered and thus showing up in the database, but not actively seeking work. There is also the case of people considered unemployed by LFS standards but simply not being registered as jobseekers, as it is not compulsory to do so in Hungary.

The lack of compulsory elements in unemployment data is certainly a problem that is hard to combat. While we are aware of even administrative data having limits, with tax records and employment data we had a certain stability provided by the fact that it is determined by law that records must be made. As a result, the differences between the two populations, LFS and tax records, could be defined and followed more easily. However, with our limited options, the number of registered jobseekers is still the most suitable supplementary data for our estimation of unemployed people.

Conceptual differences between the indicators were present in unemployment data as well, as administrative data records the number of people in the registry at a given date. Although, in this case the date is the 20th of the reference month, not the end of the month, which could cause additional discrepancies in the movement of the two data. Fortunately, timeliness is not an issue in this case, as registered jobseeker data is available even before LFS data of the reference month is processed and the time-series available dates back to the early 2000s, so back-calculations were feasible as well.

Model estimation

The Employment Statistics Section and the Department of Methodology worked together on the development of the new estimation method. After the preliminary research the Department of Methodology decided that the best method to incorporate administrative data would be through the so-called “state space” models, which can be used for regularly measured data such as the LFS. The basis of such models is that there are constantly changing, but directly not observable events, for which we would like to have an estimation. In our case, the directly not observable events could actually mean observable events, however these observations might not be the most accurate, such as our monthly employment and unemployment statistics derived from LFS.

To get a stable, functional estimation, state space models had two main supports: the tendencies and movements of the raw estimate’s error in all the needed demographical subgroups and the previously described administrative data sources. As the sampling method and the survey were well-known to us, most of our preliminary research and work was focused on the administrative data sources. Beyond the conceptual differences, we examined the time series of all administrative data and LFS data available, checking for any further discrepancies and other unusual shifts in the time-series, as well as whether these shifts had any underlying economic explanations or were something we had to handle as outliers.

During the estimation process, the 15-74 age group for both employed and unemployed persons was calculated first. The Methodology Department found that it is more effective to give an estimation for the total population observed first and to calculate the demographical subgroups afterwards, instead of determining subpopulations’ data first and their sum making up the total population estimate. Following the estimation of the 15-74 age group, 15-24 and 65-74 age groups were estimated and other needed subpopulations were calculated from the differentials of these three given age groups. However, for unemployed persons the 65-74 age group was too small to model, so we ended up using raw LFS data for the subgroup.

With the addition of the sex variable being included for each age group, we ended up with model estimates for 9 subpopulations for the employed (Males, females and total for the 15-74, 15-24 and 65-74 age groups), and 6 subpopulations for the unemployed (Males, females and total for the 15-74 and 15-24 subgroups). A core principle while developing the model was that the three-month moving averages of the monthly estimates should be close to the three-month moving averages of the raw monthly data, even in the aforementioned demographical subgroups.

Results

The three-month results of the model confirm that, it is possible to produce better quality estimates by including administrative data.

With the help of state space models based on LFS data, tax records and registered jobseeker records, we ended up with results that were more consistent with other administrative data published by the HCSO and thus less confusing for the users. In addition, the model controlled the volatility of the monthly sub-samples better, while it was able to follow real-life changes, like the impact of the COVID-19 restrictions on the labour force market in the spring of 2020.

Figure 1: Number of monthly employed people

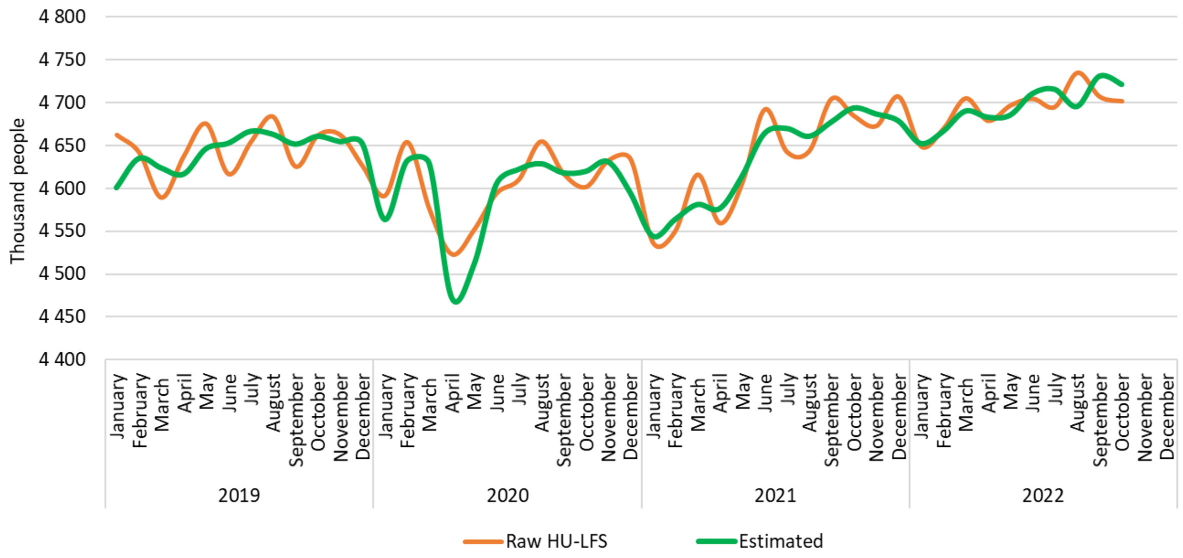
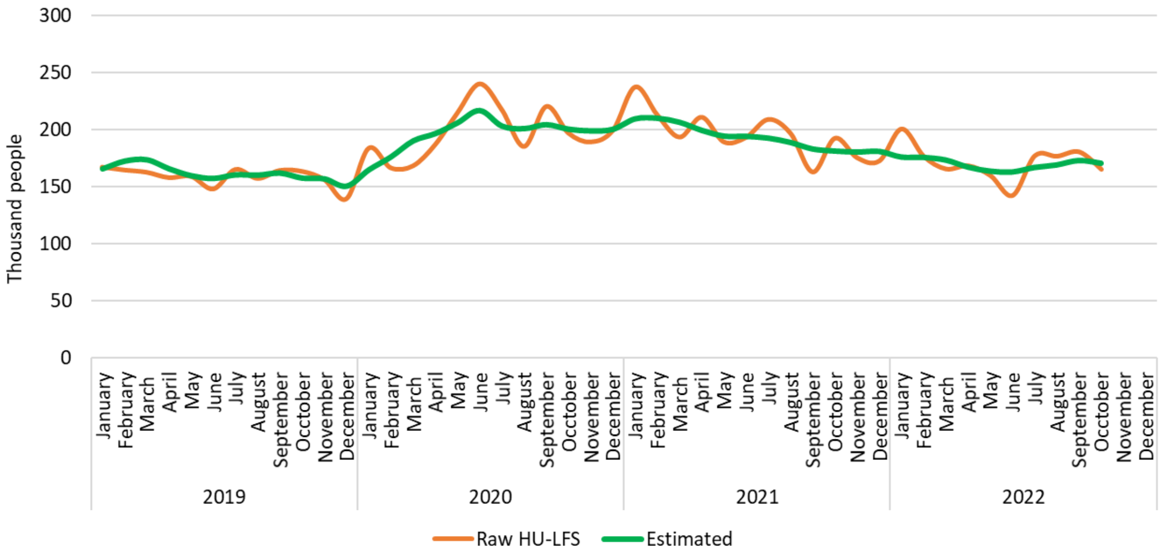


Figure 2: Number of monthly unemployed people



It is clear that, the estimation of both the employed and the unemployed not only effectively "smooths out" the fluctuation of the LFS subsamples, it also effectively follows seasonal monthly trends of the administrative data in a certain manner, that LFS three-month moving averages are not able to trace. By combining the two data sources, it becomes possible to examine labour market processes and their underlying economic causes in a more complex way.

Publication

The methodological change was introduced as part of a major revision on February 24th with the monthly first release of January 2023. The published monthly time series were replaced with the results of the model on the HCSO website, and the revised time series were sent to Eurostat as well. As a result, we managed to comply with our legal obligation regarding the monthly unemployment rate, while providing data necessary to create and monitor the European Union labour market policies.

We back-calculated the data until 2011, which caused a break in the time-series, as for the years before 2011 we only have had the raw monthly LFS data available. After the initial publication, we decided not to revise the time-series monthly (mostly to avoid confusing the users), we only extend it by adding the latest results as the time-series develops. It is important to highlight, the three-monthly moving average and the quarterly data remained the main indicator in HCSO's Employment Statistics publications, the monthly data is provided as complementary information with a higher degree of precision than before.

Summary

For the past couple of years, the HCSO had been in the process of developing the model estimation for monthly employment and unemployment data. As a result of the estimation, the available monthly data became suitable for temporal comparison parallel to the improvement of the other quality criteria (relevance, accuracy, accessibility, comparability).

Information from several data sources can provide a more precise picture of the labour market. The estimation procedure demonstrated to us that the combined use of a population survey and administrative data is a promising method in official statistics.