Thomas Önskog, SSA/SU/MET
Frida Videll, SSA/SU/MET

Methodological summary of corrections to the main indicator series in the Swedish LFS

Background

As from January 1, 2021, the Swedish Labour Force Survey, LFS, must comply with the new EU framework regulation on social statistics. To do so, several changes has been made to the survey, both regarding the population, the definitions, and the questionnaire. In addition to the changes caused by the new framework regulation, the auxiliary information used in the estimation has also been revised. To reduce the number of breaks in the time series, the change in auxiliary information was made at the same time as the implementation of the new framework regulation.

Design of the Swedish LFS

The purpose of the LFS is to describe the current labour market conditions for the entire target population. The LFS is the only source that continuously provides a coherent picture of the labour market in terms of employment, unemployment, hours worked, etc.

The LFS is a sample survey based on individuals and is conducted by telephone interviews every month throughout the year. The monthly sample is approximately 18 200 individuals. The sample individuals answer questions about their situation on the labour market during a specific week, called the reference week, of the reference month. The structure is such that all weeks during the year are studied. The results of the monthly surveys are published shortly after the end of the reference period. These results also form the basis for estimations of quarterly and annual averages.

The LFS is a panel survey with a rotating sample where every individual in the sample participates once each quarter for two years. This means that 7/8 of the sample is repeated at a three-month interval and 1/8 of the sample is replaced with new sample individuals.

Changes made to the Swedish LFS in January 2021

Change in definition

The main change in the definition of status on the labour market concerns the class of employees, or more precisely employees who are absent from work. Some people who were previously classified as employees are no longer classified as such according to the new framework regulation. The changed status applies to people that are absent for three months or more due to some particular reasons. The change in definition will lead to a decrease in the number of employed.

There are also some changes in the questions that are used to determine if a person should be classified as unemployed, but it is difficult to know from the outset if these changes will increase or decrease the number of unemployed. The question about how the respondent has been looking for a job has been clarified. Before, this was an open question, but it now has fixed alternatives. The interviewer reads the alternatives until the respondent answers 'yes' on an alternative or until the alternatives run out.

Change in population

The target population in the new framework regulation consists of people living in private households as compared to the target population in the old framework regulation being the entire resident population. Due to this change, the target population will decrease. There is no prior information about the size of the population in private households, so this group is identified based on their answers to questions in the LFS.

There is also a change in the age of the individuals in the target population as the age group is widened from 15–74 to 15–89.

Change in the questionnaire

There is a change in the order in which the questions are asked. The classification of ILO[1] status will now take place in the beginning of the interview and this classification will also be done in every interview. Before, the Swedish LFS had dependent interviewing in which the respondent was asked if the labour situation had changed since the last interview. If the situation was the same, a shorter follow-up interview was conducted. It is no longer possible to use this approach for questions about classification of ILO status.

Change in auxiliary information

The estimation in the Swedish LFS is based on a generalized regression estimator with auxiliary information from administrative data. The auxiliary information comprises variables that identify important domains or that covary with the survey variables and/or the response propensity. Information about sex combined with age of the

[1] International Labour Organization

respondents, as well as information on region and county of birth, are taken from the Total Population Register (TPR).

Information from the Swedish Public Employment Service's (Af) register of job seekers are also used in the auxiliary information. One of the changes in the auxiliary information concerns the information used from this register. As from January 2021, more categories in the register are used in the auxiliary information to better capture those who are unemployed in the Swedish LFS.

Information from the Employment Register, which is updated yearly, has previously been used in the auxiliary information. As from January 2021, employment status in the auxiliary information is derived directly from monthly employer reports at individual level (AGI). This makes the auxiliary information timelier. The change in administrative data for employees does not only concern a change from yearly to monthly basis. The information from the Employment Register was divided into groups of industrial classification, but information on industrial classification was not in place for AGI when the Swedish LFS started to use this register so the information from AGI is not divided by industrial classification. Instead, it is divided by age group, since the non-response pattern is different for different age groups.

Available information

There are many ways to analyze and quantify a potential break in the Swedish LFS time series in January 2021. In our analysis, we have used four main sources of data, which we now describe shortly.

Parallel run

A parallel run has been conducted during the entire year 2021. To obtain the parallel run, the monthly sample of approximately 18 200 persons has been divided into two parts. One part of the sample, consisting of 80 percent of the sample, is approached using the new procedure whereas the other part, consisting of 20 percent of the sample, is approached using the old procedure. Hence, data collected using both the old and the new questionnaires will be available for all months of 2021.

Data with different types of auxiliary information

Monthly reports at individual level (AGI) have been available since February 2019, although this information was not used in the production of the Swedish LFS until January 2021. Thanks to this, we can reproduce the new auxiliary information back to February 2019. For this time period, we can derive parallel time series based on the old and the new auxiliary information, respectively. In addition, as from January 2021, we can derive parallel time series with the old and the new auxiliary information, respectively, for both the old and the new questionnaire.

Data on the change of definition

To quantify the effect of the change of definition for people that have been absent from work for more than three months, a test was carried

out between February 2020 and December 2020. In this test, additional questions were asked to people who were classified as employed but at risk of being classified as unemployed or outside the labour force in the new framework regulation. These additional questions were asked at the end of the interview and were similar to the questions in the new questionnaire.

Moreover, during 2021 we have collected data on the respondents who are no longer classified as employed in the new questionnaire but would have been classified as employed in the old questionnaire.

Flow data

The Swedish LFS is a panel survey and respondents take part in the survey for eight consecutive quarters. This design provides us with a rich source of flow data describing how the employment status of the respondents changes from quarter to quarter. In particular, we can compare the flows in the labour market between Q4 2020 and Q1 2021 to the flows that we normally observe in the labour market. Any deviations from the normal flows are an indicator of differences between the old and the new framework regulation.

Methodology for break estimation

We have used weighted means to analyse the data based on the two different types of auxiliary information and the data from the parallel run. To describe the procedure in detail, we first introduce some notation. Let $X_t^{q,a}$ denote the value of time series X at time t. The indices q and a here correspond to questionnaire and auxiliary information, respectively. Both these indices take values in the set $\{o, n\}$, where o represents old and n represents new. This gives four versions in total of every time series. The total break in the times series is given by

$$X_t^{n,n} - X_t^{o,o},$$

where t is an arbitrary month after January 2021. At present, we have access to estimates of $X_t^{n,n}$ and $X_t^{o,o}$ for the first 10 months of 2021, but since we have data on the old and new auxiliary information during a longer period, we choose to decompose the break as

$$X_t^{n,n} - X_t^{o,o} = \underbrace{(X_t^{n,n} - X_t^{o,n})}_{change\ in\ questionnaire} + \underbrace{(X_t^{o,n} - X_t^{o,o})}_{change\ in\ auxiliary\ inf.}.$$

The first term on the right-hand side corresponds to the break due to change in questionnaire and it can be estimated using the data from the parallel run. The second term on the right-hand side corresponds to the break due to change in auxiliary information and it can be estimated using the data with different types of auxiliary information.

We first estimate the size $\hat{\Delta}^a$ of the break caused by the change in auxiliary information. The estimate of $\hat{\Delta}^a$ is given by

$$\hat{\Delta}^a = \sum_{t \in [T_0^a, T_1]} \alpha_t (X_t^{o,n} - X_t^{o,o}),$$

where T_0^a denotes February 2019 and T_1 October 2021. The coefficients α_t are non-negative numbers satisfying the constraint $\sum_{t\in[T_0^a,T_1]}\alpha_t = 1$. When calculating quarterly and yearly means for time series in the Swedish LFS, the first two months of every quarter have weight 4 and the last month weight 5. We use weights with these proportions in the estimate of $\hat{\Delta}^a$. However, we also note that the variance of the break estimate is minimized if the weight of a given month is inversely proportional to the variance of the estimate for that month. As the size of the sample in 2021 is $1/5$ of the size of the sample in 2019 and 2020, we multiply the weights of all months of 2021 by a factor $1/5$.

Since $X_t^{o,n}$ and $X_t^{o,o}$ are estimated based on the same sample, we can use the standard estimation procedure to estimate the mean error σ_t of $X_t^{o,n} - X_t^{o,o}$. The correlation between $X_t^{o,n} - X_t^{o,o}$ and $X_s^{o,n} - X_s^{o,o}$ for different times s and t appears to be very weak and this holds also when $s - t$ is a multiple of three months, that is when the samples at times s and t partly overlap. Consequently, we can estimate the variance of the break estimate $\hat{\Delta}^a$ as

$$\hat{V}^a \approx \sum_{t\in[T_0^a,T_1]} \alpha_t^2\sigma_t^2 \approx 0.0373\bar{V}_1 + 0.00063\bar{V}_2 \approx 0.0405\bar{V}_1,$$

where $\bar{V}_1$ is the sample mean of the variances of $X_t^{o,n} - X_t^{o,o}$ during 2019–2020 and $\bar{V}_2$ is the sample mean of the variances of $X_t^{o,n} - X_t^{o,o}$ during 2021. The right-most relation in the equation above holds since $\bar{V}_2 \approx 5\bar{V}_1$ according to the design of the parallel run.

In total, we have 33 months of data on the change in auxiliary information. This is almost as much data that is required to detect a change in seasonal pattern. Therefore, we have, for all 12 months of the year, calculated a weighted mean of all differences $X_t^{o,n} - \left(X_t^{o,o} + \hat{\Delta}^a\right)$ that are available for that month. For the number of employed, there seems to be a linear relationship over the year of these weighted means. However, as this linear relationship cannot be seen in data from the new questionnaire, we have chosen not to take any change in seasonal pattern due to the change in auxiliary information into account in the construction of linked time series.

We next estimate the size $\hat{\Delta}^q$ of the break caused by the change in questionnaire. The change in questionnaire can, of course, also give rise to a new seasonal pattern. However, since the parallel run only provides us with one measurement per month, we cannot estimate this change in seasonal pattern and will therefore omit it. The estimate of $\hat{\Delta}^q$ is given by

$$\hat{\Delta}^q = \sum_{t\in[T_0^q,T_1]} \beta_t(X_t^{n,n} - X_t^{o,n}),$$

where T_0^q denotes January 2021 and T_1 October 2021. The coefficients β_t are non-negative numbers satisfying the constraint $\sum_{t\in[T_0^q,T_1]}\beta_t = 1$. We let the weights β_t have the same proportions as the weights α_t. The variance of the break estimate $\hat{\Delta}^q$ is then given by

$$\hat{V}^q = \sum_{t\in[T_0^q,T_1]} \beta_t^2\sigma_t^2 + \sum_{s,t\in[T_0^q,T_1]} \beta_s\beta_t\lambda_{s,t}\rho_{s,t}\,\sigma_s\sigma_t$$
$$\approx \begin{cases} 0.250\bar{V}, & \text{for employed,} \\ 0.156\bar{V}, & \text{for unemployed,} \end{cases}$$

where $\bar{V}$ is the sample mean of the variances of $\{X_t^{n,n} - X_t^{o,n}\}_{t\in[T_0^q,T_1]}$, that is the sample mean of $\{(\sigma_t^{n,n})^2 + (\sigma_t^{o,n})^2\}_{t\in[T_0^q,T_1]}$. Here $\lambda_{s,t}$ is the fraction of the sample that is common between times s and t and $\rho_{s,t}$ is the correlation between $X_s^{n,n} - X_s^{o,n}$ and $X_t^{n,n} - X_t^{o,n}$. For simplicity, we have replaced $\rho_{s,t}$ by historical values of the autocorrelation, that is

$$\rho_{s,t} = \begin{cases} 0.72^k, & \text{for employed} \\ 0.38^k, & \text{for unemployed} \end{cases}$$

if $k = |s-t|/3$ is an integer and $\rho_{s,t} = 0$ for all other choices of s, t.

As described above, the new definition of employed rule out some people who have been absent from work for at least three months. To estimate the size $\hat{\Delta}^d$ of this group, we use the monthly estimates of this group in the new questionnaire during 2021. We derive break estimates by calculating a weighted mean of the first ten months of 2021. For some of the time series, there is a clear seasonal pattern in the number of people affected by this change in definition, and, for that reason, we have calculated one weighted mean for the summer months (June to August) and another weighted mean for the remaining months. We also calculated similar weighted means based on the additional question posed in the old questionnaire in 2020, but as the results are similar, we have chosen to base the estimate of the break due to the change in definition solely on the 2021 data. We have estimated the variance $\hat{V}^d$ of $\hat{\Delta}^d$ using the same formula as for $\hat{V}^q$, but with $\bar{V}$ as the sample mean of the variances of the estimated number of people affected by the change in definition.

The data from the parallel run contains information about all changes in the questionnaire, including the change in definition but, potentially, also other effects that cannot be directly quantified. As the parallel run only gives 10 months of data with quite small sample sizes, the uncertainty in break estimates derived from the parallel run is high. As an effect, no statistically significant differences between the questionnaire can be seen in the parallel run, as shown in Table 2 below. On the other hand, the change in definition evidently causes a statistically significant change between the questionnaires. To ascertain that this change is incorporated in the linking, we have chosen to interpret the estimate $\hat{\Delta}^d$ as the break due change in questionnaire as long as

$$\hat{\Delta}^q - 1.96\sqrt{\hat{V}^q} \le \hat{\Delta}^d \le \hat{\Delta}^q + 1.96\sqrt{\hat{V}^q},$$

holds, that is if the estimated change in definition does not deviate significantly from the result of the parallel run.

Turning now to the change in population, we note that respondents affected by this change (respondents not living in private households)

almost exclusively do not belong to the labour force. As a result, the change in population will not affect the number of employed or unemployed.

Flow data have been analysed to see if there are any deviations in the quarterly flows that cannot be explained by the change in definition, auxiliary information, or questionnaire. As this was not the case for the time series of interest here, we do not consider the flow data further.

Methodology for linking

Using the analysis described in the previous section, we can estimate the sizes of the time series breaks arising due to the change in auxiliary information and questionnaire, respectively. To construct linked time series based on this information, we then proceed as follows.

First, we transform the break estimates (which are given in terms of thousands of people) into a factor of the total population. More precisely, we state all breaks in terms of percentage of the mean population Y_t^o (according to TPR) during January 2021 to October 2021.

Then, we obtain linked time series for the period 2009–2020 by adjusting the time series $X_t^{o,o}$ according to the formula

$$X_t^{link} := X_t^{o,o} + Y_t^o \frac{\hat{\Delta}^a + \hat{\Delta}^d}{\frac{1}{T_1 - T_0^q} \sum_{t \in [T_0^q, T_1]} Y_t^o}.$$

The resulting linked time series will remain consistent as all break estimates are weighted with the same amount Y_t^o. The adjustments are also proportional to the size of the population with the result that the adjustments decrease slightly as we go back to the beginning of the linking period 2009–2020.

Break estimates

We now present the analysis of the time series breaks. We begin with the break caused by the change in auxiliary information. As shown in Table 1 below, there are statistically significant breaks in all time series for employed but only for the age group 15–24 years for unemployed.

Table 1. Estimated size of time series breaks caused by the change in auxiliary information (new minus old auxiliary information) for the number of employed and number of unemployed, respectively, measured in thousands. Significant breaks at the 5% level are indicated by an asterisk.

Time series	Employed	Unemployed
Males, 15–24 years	4.7 (±2.1)*	-2.7 (±0.7)*
Males, 65–74 years	-4.7 (±0.8)*	0.0 (±0.1)
Males, 25–64 years	-15.6 (±3.0)*	0.4 (±1.7)
Males, 20–64 years	-10.5 (±3.0)*	N.A.
Females, 15–24 years	3.7 (±1.9)*	-1.3 (±0.5)*
Females, 65–74 years	-3.7 (±0.7)*	0.0 (±0.1)

Time series	Employed	Unemployed
Females, 25–64 years	-14.9 (±3.6)*	-1.4 (±1.5)
Females, 20–64 years	-11.0 (±3.6)*	N.A.

Next, we consider the break caused by the change in questionnaire and analyse data from the parallel run. As seen in Table 2 below, the parallel run gives no indication of statistically significant breaks in the time series of interest.

Table 2. Estimated size of time series breaks caused by the change in questionnaire (new minus old questionnaire) for the number of employed and number of unemployed, respectively, measured in thousands. Significant breaks at the 5% level are indicated by an asterisk.

Time series	Employed	Unemployed
Males, 15–24 years	-8.4 (±25.2)	10.6 (±16.0)
Males, 65–74 years	7.6 (±26.3)	0.5 (±4.6)
Males, 25–64 years	-6.3 (±40.7)	13.7 (±21.1)
Males, 20–64 years	-7.8 (±45.3)	N.A.
Females, 15–24 years	2.5 (±24.6)	-5.0 (±16.5)
Females, 65–74 years	5.9 (±19.9)	0.1 (±3.5)
Females, 25–64 years	-2.5 (±44.6)	-2.9 (±22.7)
Females, 20–64 years	-5.5 (±48.7)	N.A.

Since the uncertainty in the estimates from the parallel run are quite high, we cannot use these estimates directly to estimate the breaks caused by the change in questionnaire. Instead, we consider the direct estimates of the number of people affected by the change in definition. Tables 3 and 4 give estimates of the number of employed and unemployed, respectively, that are affected by the change in definition. For employed, we have statistically significant effects on all series. We also note that the difference between summer and non-summer means is statistically significant for the time series corresponding to age groups 20–64 years and 25–64 years, but not for the time series corresponding to age groups 15–24 years and 65–74 years. For unemployed, the break estimates corresponding to 25–64 years are statistically significant (with a statistically significant difference between summer and non-summer means), the break estimates corresponding to 15–24 years are almost statistically significant (but the break estimates are very small) and the break estimates corresponding to 65–74 years are not statistically significant.

Table 3. Estimated size of time series breaks caused by the change in definition (new minus old definition) for the number of employed, measured in thousands. Significant breaks at the 5% level are indicated by an asterisk.

Time series	Employed summer	Employed non-sum.	Employed total
Males, 15–24 years	-0.9 (±1.0)	-3.2 (±2.0)*	-2.5 (±1.7)*
Males, 65–74 years	-2.2 (±1.6)*	-4.2 (±2.3)*	-3.6 (±2.1)*

Time series	Employed summer	Employed non-sum.	Employed total
Males, 25–64 years	-4.3 (±2.3)*	-13.9 (±4.1)*	-11.0 (±3.6)*
Males, 20–64 years	-5.1 (±2.5)*	-16.4 (±4.5)*	-13.0 (±4.0)*
Females, 15–24 years	-1.2 (±1.2)*	-3.3 (±2.0)*	-2.6 (±1.8)*
Females, 65–74 years	-2.8 (±1.8)*	-2.0 (±1.6)*	-2.2 (±1.7)*
Females, 25–64 years	-8.0 (±3.1)*	-20.5 (±5.0)*	-16.8 (±4.5)*
Females, 20–64 years	-8.0 (±3.1)*	-23.2 (±5.3)*	-18.6 (±4.7)*

Table 4. Estimated size of time series breaks caused by the change in definition (new minus old definition) for the number of unemployed, measured in thousands. Significant breaks at the 5% level are indicated by an asterisk.

Time series	Unemployed summer	Unemployed non-sum.	Unemployed total
Males, 15–24 years	0.0 (±0.0)	0.7 (±0.7)	0.5 (±0.6)
Males, 65–74 years	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)
Males, 25–64 years	1.5 (±1.1)*	4.5 (±1.8)*	3.6 (±1.6)*
Females, 15–24 years	1.2 (±1.0)*	0.6 (±0.7)*	0.8 (±0.8)*
Females, 65–74 years	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)
Females, 25–64 years	0.9 (±0.9)*	3.0 (±1.5)*	2.4 (±1.3)*

We note that all statistically significant break estimates in Tables 3 and 4 fall into the range of the confidence intervals from the parallel run shown in Table 2. In Tables 5 and 6, we have summarized the break estimates that are used in the formula at the bottom of page 7 to produce linked time series.

Table 5. Break estimates for the number of employed used in the linking of time series. Separate estimates for breaks caused by the change in auxiliary information ($\hat{\Delta}^a$) and breaks caused by the change in definition ($\hat{\Delta}^d$). Break estimates are measured in thousands.

Break type	M 15–24	M 65–74	M 25–64	M 20–64
$\hat{\Delta}^a$	4.7	-4.7	-15.6	-10.5
$\hat{\Delta}^d$ (summer)	-	-	-4.3	-5.1
$\hat{\Delta}^d$ (non-sum.)	-	-	-13.9	-16.4
$\hat{\Delta}^d$ (total)	-2.5	-3.6	-	-
Break type	**F 15–24**	**F 65–74**	**F 25–64**	**F 20–64**
$\hat{\Delta}^a$	3.7	-3.7	-14.9	-11.0
$\hat{\Delta}^d$ (summer)	-	-	-8.0	-8.0
$\hat{\Delta}^d$ (non-sum.)	-	-	-20.5	-23.2
$\hat{\Delta}^d$ (total)	-2.6	-2.2	-	-

Table 6. Break estimates for the number of unemployed used in the linking of time series. Separate estimates for breaks caused by the change in auxiliary information ($\hat{\Delta}^a$) and breaks caused by the change in definition ($\hat{\Delta}^d$). Break estimates are measured in thousands.

Break type	M 15–24	M 65–74	M 25–64
$\hat{\Delta}^a$	-2.7	-	-
$\hat{\Delta}^d$ (summer)	-	-	1.5
$\hat{\Delta}^d$ (non-sum.)	-	-	4.5
$\hat{\Delta}^d$ (total)	-	-	-
Break type	**F 15–24**	**F 65–74**	**F 25–64**
$\hat{\Delta}^a$	-1.3	-	-
$\hat{\Delta}^d$ (summer)	-	-	0.9
$\hat{\Delta}^d$ (non-sum.)	-	-	3.0
$\hat{\Delta}^d$ (total)	-	-	-

We refer to the separately submitted csv file for the linked time series.