Correcting for time series breaks in the Swedish Labour Force Survey at the micro level – combining models and calibration

Thomas Önskog[1], Frida Videll[2]

[1]*Statistics Sweden, Sweden, thomas.onskog@scb.se*

[2]*Statistics Sweden, Sweden, frida.videll@scb.se*

Abstract

The Swedish Labour Force Survey (LFS) is the foundation for official statistics on the Swedish labour market. As from January 1, 2021, the 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 and the questionnaire. In addition to the changes caused by the new framework regulation, a revision of the auxiliary information used in the estimation was implemented at the same time as the new framework regulation. To the serious concern of the users, the changes made to the procedure of the LFS has caused breaks in many of the time series.

To estimate and correct for the breaks in the LFS time series, Statistics Sweden performed a parallel run with both the old and the new questionnaires during 2021. The results of the parallel run have been complemented by data from a series of other sources, such as data on respondents that were affected by a change in the definition of employed, estimates based on both the old and the new auxiliary information, respectively, as well as flow data describing how the employment status of respondents vary from one quarter to the next.

In this paper, we describe how the available data has been analysed and combined to correct for the time series breaks and how we have derived new time series for the LFS that are comparable back to 2005. Using a combination of calibration and imputation models for the micro data, we have derived a set of recalibrated weights for all the respondents of the LFS during the time period 2005–2020. The recalibrated weights constitute a link at the micro level between the old and new procedures. Time series derived using the recalibrated weights and the imputed micro data during the time period 2005–2020 are fully comparable with time series calculated during 2021 and onwards using the new procedure.

Keywords: labour force survey, time series, calibration, imputation, user demands.

1. Introduction

The Swedish Labour Force Survey (LFS) is the only source that continuously provides a coherent picture of the Swedish labour market in terms of employment, unemployment, hours worked, etc. As from January 1, 2021, the LFS must comply with the new EU framework regulation on social statistics (EU, 2019). To do so, several changes have been made to the survey regarding the population, 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. All these changes can potentially cause breaks in the LFS time series.

In order to make the LFS time series comparable over time, Statistics Sweden decided to produce linked time series for the time period 2005–2020. Linking of time series can be carried out at the macro level or at the micro level. Within the framework of these two approaches, there are many different methods that can be employed to produce linked time series. We here give a short overview of the two approaches and their respective benefits and drawbacks.

A macro-approach means that linking is done at an aggregated level for a selection of time series. The selected time series are adjusted backwards in time based on an analysis of the sizes of the time series breaks. Linking with a macro-approach is less resource-intensive because only selected time series are linked. Sum-consistency of the linked time series is not obtained automatically, but must be ensured by a separate process.

A micro-approach refers to changing the calibrated weights of the sample individuals. Based on these recalibrated weights, linked time series can be produced. A micro-approach is more flexible as the linking is carried out at the individual level, making it unnecessary to decide in advance which time series that are to be linked. With a micro-approach it is theoretically possible to link all time series in the LFS and consistency between the time series is obtained automatically. However, this approach is more resource-intensive than the macro-approach. Moreover, due to the large number of linked time series that are produced, it is difficult to control the quality of all linked time series.

Statistics Sweden used a macro-approach to produce over 1200 linked LFS time series for the time period 2005–2020, see Önskog and Videll (2022) for details. After that, the project continued with the objective to produce time series that are linked at the micro level. In this article, we describe the methodology and results of the linking at the micro level.

1.1 Design of the LFS

The LFS is a sample survey based on individuals and is conducted by telephone interviews every month throughout the year. The monthly sample in 2021 was 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 month. These results also form the basis for quarterly and annual averages. The LFS is a panel survey with a rotating sample where every individual in the sample participates once every 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 every month.

2. Changes made to the LFS in January 2021

In this section, we shortly describe the changes made to the LFS in January 2021. Four categories of changes have been implemented: changes in the definition, population, questionnaire, and auxiliary information.

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

Change in population. The target population in the new framework regulation contains all people living in private households as compared to the target population in the old framework regulation which equalled 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 the answers collected in the LFS. As from 2023 there is, however, information from registers available so that the total number of people not living in private households can be extracted from registers (see Appendix A.2 for details).

Change in questionnaire. The new framework regulation enforces a new order in which the questions are asked. The classification of ILO status is now done in the beginning of the interview and this classification is also done in every interview. Before, the 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. There are also some changes in the questions that are used to determine if a person should be classified as 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. It is difficult to know from the outset what the effects of the changes in the questionnaire will be.

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

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. Moreover, information from the register of job seekers

provided by the Swedish Public Employment Service is used in the auxiliary information. As from January 2021, more categories in this register are used in the auxiliary information to better capture those who are unemployed in the LFS.

3. Data sources for analysis of time series breaks in the LFS

Six types of data have been used in the analysis of time series breaks in the LFS.

Parallel run. A parallel run was conducted during all months of 2021. During the parallel run, the monthly sample was divided into two parts. 80 percent of the sample was interviewed using the new questionnaire and the remaining 20 percent of the sample was interviewed using the old questionnaire. Hence, data based on both the old and the new questionnaires are available for all months of 2021.

Data with different types of auxiliary information. Monthly employer reports at individual level (AGI) are available since February 2019, although this information was not used in the production of the 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 questionnaires.

Data on the change in definition. To quantify the effect of the change in definition of employed, additional questions were introduced in the LFS between February 2020 and December 2020. The additional questions were given to people who were classified as employed but were at risk of being classified as unemployed or outside the labour force in the new framework regulation. The questions were asked at the end of the interview and were similar to the questions in the new questionnaire. Moreover, during 2021, we collected data on the respondents who are no longer classified as employed in the new questionnaire due to the change in definition but would have been classified as such in the old questionnaire.

Data on the change in population. Persons not living in private households generally belong to one of the following three main groups: (a) Persons living in residential care facilities, such as retirement homes or group homes, (b) Conscripts or contracted soldiers, and (c) Prison inmates. There are also other groups that do not live in private households, such as those living in monasteries, but these groups are considerably smaller in number than the three main groups and are therefore neglected here. The number of people in the three main groups can to some extent be estimated using existing registers (see Appendix A.2 for details).

Flow data. The 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 one quarter to the next. In particular, we can compare the flows in the labour market between Q4 2020 and Q1 2021 to the flows that are normally observed in the labour market. Any deviations from the normal flows are an indicator of differences between the old and the new procedures.

Time series analysis. In the seasonal adjustment of the LFS time series, the SAS macro Proc X13 is used to produce seasonally adjusted time series and trend series. With the help of the automatic outlier search in X13, it is possible to investigate whether there is evidence for a level shift in a certain time series starting at a certain month. The program also estimates the size of a possible level shift. By applying X13 to published LFS time series, it is thus possible to estimate the break in the time series in January 2021. Furthermore, we can apply X13 to the linked time series to ensure that no time series breaks remain in January 2021.

4. Methodology

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 (Van den Brakel, Smith, & Compton, 2008). The procedure was thoroughly described in Önskog and Videll (2022) and is hence omitted here.

The breaks due to the changes in definition and population, respectively, have been handled as follows. Using the methodology outlined in Appendices A.1–A.2, we have derived, for every month t of interest and every relevant LFS time series, estimates $\hat{\Delta}_t^d$ and $\hat{\Delta}_t^p$ of the impacts of the changes in definition and population, respectively, on that time series during month t.

In Figure 1 in Appendix B, we give an example of one of the linked time series (the total number of employed) that have been produced within the frame of this project. As is evident from the figure, one of the issues when handling this particular time series break was that it occured during the corona pandemic. For many time series it has not been straightforward to disentangle the effect of the new regulation from the effect of the pandemic.

4.1 Linking time series with a macro-approach

Let X_t be one of the LFS time series and let $\hat{\Delta}^h$ be the break estimate due to the change in auxiliary information for this particular time series. Further, let $\hat{\Delta}^e$ be the possible additional break effect (detected from flow data, time series analysis and/or the parallel run) for this time series. Moreover, let $\hat{\Delta}_t^d$ and $\hat{\Delta}_t^p$ be the model estimates of the effect of the change in definition and change in population, respectively, for this time series during month t. We define a linked time series, X_t', at the macro level for this time series as

$$X_t' := X_t + \hat{\Delta}_t^d + \hat{\Delta}_t^p + Y_t^g\left(\hat{\Delta}^h + \hat{\Delta}^e\right),$$

where Y_t^g is the size of the population during month t in the age group that X_t refers to divided by the mean population during 2021 in the same age group. Linked time series for the ages 15–74 years are obtained by summing the linked time series for all sub-groups with respect to age. By this construction, the linked time series at a macro level will be sum-consistent.

4.2 Linking time series with a micro-approach

As mentioned in Section 2, a generalized regression estimator is used in the estimation of the LFS. To describe the methodology in more detail, we let $\boldsymbol{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)^T$ denote an auxiliary vector which is known from registers for the entire population U. Hence, the auxiliary totals $\boldsymbol{X} = \sum_{i \in U} \boldsymbol{x}_i$ are also known. To every respondent we assign a calibrated weight w_i, satisfying the calibration equation

$$\sum_{i \in r} w_i \boldsymbol{x}_i = \boldsymbol{X},$$

where r denotes the set of respondents. The calibrated weights can be used to estimate the total of every survey variable y_i in the LFS by means of $\hat{y} = \sum_{i \in r} w_i y_i$.

We here use a similar calibration approach to obtain linked LFS time series at the micro level. The idea is to use the linked LFS time series produced at the macro level as described in Section 4.1 as auxiliary totals in the calibration equation. In this calibration we use micro data from the respondents that has been slightly altered by certain imputation models, see Section 4.2.1 below. This procedure produces, for every month in the linking period, a set of recalibrated weights that are consistent with the linked LFS time series at the macro level. The set of recalibrated weights, which we henceforth refer to as micro-linked weights, can be used to produce linked estimates for all survey variables in the LFS.

A couple of important methodological issues must be considered when deriving micro-linked weights. We devote the following two subsections to two of these issues and to the solutions that we have found to the two issues.

4.2.1 Imputation models

One potential problem with the micro-approach is the following situation: we have no sample individuals in the survey data for a given month that satisfies the conditions of a given time series although the corresponding linked time series at the macro level has a strictly positive value that month. If this happens, the calibration equation will be violated and micro-linked weights cannot be derived. For the current time series break, this situation occurs for time

series on certain subgroups of people who are available for work, but not seeking. As shown in Figure 2 in Appendix B, this group has become considerably larger in the new questionnaire.

Here, we have solved this problem by means of an imputation model for the group of people available for work, but not seeking. The imputation model uses characteristics of the sample individuals (extracted from registers and the interview) to identify those who were likely to have been classified as belonging to this group if they had been interviewed with the new questionnaire in the past, but who were not classified in this group in the old questionnaire. With the help of the model, the survey data is adjusted so that individuals who are identified by the model are re-classified as available for work, but not seeking. For details of the imputation model, see Appendix C in SCB (2023).

In a similar way, we have created an imputation model for the change in definition. Using this model, we change the survey data for some sample individuals from being absent from work (and employed) to not belonging to the labour force, for details see Appendix A in SCB (2023).

In the derivation of micro-linked time series, the imputation models for the change in definition and for persons available for work, but not seeking, were used to adjust the survey data. In addition, persons who identified themselves as conscripts or contracted soldiers in the interview were removed from the set of respondents. In this context, it should also be noted that an advantage of using imputation models in the linking at micro level is that it reduces the difference between the micro-linked weights and the original calibrated weights.

4.2.2 Optimal choice of constraints

Choosing the optimal set of linked time series to use as auxiliary totals, or constraints, when calculating micro-linked weights gives rise to the following trade-off. If the entire set of macro-linked time series is used as constraints, there is a risk of overadjustment in the calibration of micro-linked weights. This will result in too much variation in the micro-linked weights and perhaps even negative weights. On the other hand, if too few of the macro-linked time series are used as constraints, the linked time series that are not used as constraints might not be reproduced with sufficient precision by the micro-linked weights.

Here, the selection of constraints was done in the following iterative manner. First, an initial set of constraints was selected which included the most important LFS time series and the time series for which the largest breaks were detected. Based on this initial set of constraints, micro-linked weights were derived and these weights were used to create micro-linked time series for all time series that had previously been linked at the macro level.

The next part of the iterative process was to evaluate how well the micro-linked and macro-linked time series coincided for all months in 2020. Time series with large deviations between

the linking at macro and micro level were added to the set of constraints. The calibration was then repeated and time series with large deviations between the linking at macro and micro level were added to the set of constraints. This iterative process continued until the discrepancies between the micro-linked and macro-linked time series were deemed acceptable. Deviations with the same sign for most of the months were then at most about 2000 individuals. Deviations with different signs in different months, indicating random deviations, were at most 5000 individuals in certain months.

The final set of constraints consisted of six vectors, each of which divides the entire target population into mutually disjunct groups. For example one of the vectors consists of age groups crossed by gender and labour force status (at work, absent from work, unemployed, not in the labour force). Other vectors consist of, among others, regions and labour force status crossed with country of origin, respectively. The total number of constraints used was 184. Out of these, 25 constraints were regions and the remaining 159 constraints were macro-linked time series.

4.2.3 Quality control of linked time series

The quality of the micro-linked weights was evaluated in a number of different ways. We made sure that there were no negative weights and that the ratio between the micro-linked weights and the original calibrated weights did not deviate too far from the value one. For example, this ratio was found to be in the interval [0.5,1] for all but 704 of the more than 3 million LFS interviews conducted during the years 2005–2020.

For all LFS time series that are regularly published by Statistics Sweden, we have investigated the corresponding micro-linked time series using Proc X13 in SAS with automatic level shift detection in January 2021. Time series for which level shifts were detected was reviewed manually. Most time series passed the manual review, but for a small number of time series, for example regarding overtime and agreed-upon working time, respectively, the micro-linked time series contain breaks in January 2021. These time series have been flagged as unreliable in the dissemination of the micro-linked LFS time series.

5 Conclusions

Using a combination of various data sources and different methods for break estimation, we have created linked LFS time series at the macro level. By imputing some of the micro data and by using the macro-linked time series as constraints in a calibration equation, we have derived micro-linked weights for the LFS. By means of these micro-linked weights, we can derive micro-linked time series for all survey variables in the LFS. The quality of these micro-linked time series has been shown to be satisfactory for a vast majority of the survey variables in the LFS.

Acknowledgments

The authors would like to thank the other members of the project group that has developed the micro-linked time series: Stefan Andersson, Alexander Astlind, Mikael Lundsten and Michella Szukis.

References

EU. (2019). Regulation 2019/1700 of the European Parliament and of the Council of 10 October 2019 establishing a common framework for European statistics relating to persons and households. http://data.europa.eu/eli/reg/2019/1700/oj

Önskog, Thomas & Videll, Frida. (2022). Linking the Swedish Labour Force Survey to compensate for time series breaks caused by the new EU framework regulation. Paper presented at the conference Q 2022 in Vilnius, Lithuania. https://q2022.stat.gov.lt/scientific-information/papers-presentations/session-13

SCB. (2023). Metod för länkning av AKU:s tidsserier för perioden 2005–2020 (in Swedish). https://www.scb.se/contentassets/2aee1dd9405d41d0b9f99dfd012ba7f6/metod-for-lankning-av-akus-tidsserier.pdf

Van den Brakel, Jan. A., Smith, Paul A., & Compton, Simon. (2008). Quality procedures for survey transitions – experiments, time series and discontinuities, *Survey Research Methods*, 2(3), 123–141. https://doi.org/10.18148/srm/2008.v2i3.68

Appendix A.1. Estimation of break due to change in definition

Regarding the break due to the change in definition, we have revised the approach adopted in Önskog and Videll (2022). The new methodology for estimating this break is based on the fact that the respondents‘ answers to other questions in the LFS, such as the reason for absence, the length of absence and the degree of attachment to the labour market, as well as the age of the respondent, strongly influence the likelihood of being affected by the change in definition. Based on the response data, we have therefore constructed a model that estimates the number of people affected by the change in definition. For details of the model, see Table 1 below and Appendix A in SCB (2023).

As in the previous methodology for the change in definition, response data from 2020 is used to determine the labour force status of people affected by the change in definition in the old questionnaire, and response data from 2021 to determine the labour force status they receive in the new questionnaire. It is worth noting that the probability of being affected by the change in definition given the response data is more or less constant over the year. Nevertheless, the model estimates of the number of people affected by the change in definition have a clear seasonal pattern, which is inherited from the seasonal pattern of some of the variables used in the model, such as reasons for absence. This is an advantage compared to the old methodology, which required separate estimates for summer and winter months.

For every month of interest and every relevant LFS time series, we can use the model and response data from that month to produce an estimate of the impact of the change in definition

on that time series during that month. In 2020, the new model and the previous methodology give similar estimates, but for the period 2005–2015, the new model estimates the total number of people affected by the change in definition to be approximately 10000 less than the previous methodology.

Table 1: Summary of the model for the change in definition. The table shows the estimated percentages of persons in different categories that are affected by the change in definition.

Reason for absence	Degree of attachment to the labour force	Length of absence			
		1–2 weeks	3–4 weeks	5–12 weeks	At least 13 weeks
Studies without salary	Permanently employed	20	70		
	Temporarily employed or entrepreneur	0		30	
Other leave of absence	All	10		30	70
Laid off, labour shortage or other reasons	Permanently employed	5	15	30	
	Temporarily employed	5	10	20	
	Entrepreneur	10	20	40	

Appendix A.2. Estimation of break due to change in population

Based on register data from the Property Register (address information for residential care facilities), the TPR, the Income and Tax Register (data on daily allowances for military service) and the Swedish Prison and Probation Service, we can estimate the number of individuals who do not live in private households as of January 2021. We can also estimate how many of these individuals that belong to the different categories in the auxiliary information. In other words, we estimate the distribution of those who do not live in private households by sex, age, region, country of birth, employment status according to AGI and unemployment status according to the Swedish Public Employment Service. With this information, we can adjust the totals in the auxiliary information so that the totals only correspond to those living in private households and then calibrate the LFS estimates according to these adjusted totals.

For the years 2005–2020, only part of the register information described above is available. For example, the address information for residential care facilities goes back to 2012. Based on the register information available during different years combined with models and extrapolations, we have been able to estimate the number of people who do not live in private households, for details see Appendix B in SCB (2023).

Appendix B. Two examples of linked LFS time series at the micro level

Figure 1: Monthly figures on the total number of employed during the time period January 2005 to March 2024 based on the old and the new regulation, respectively. The figures for the new regulation during 2005–2020 are given by the micro-linked time series.

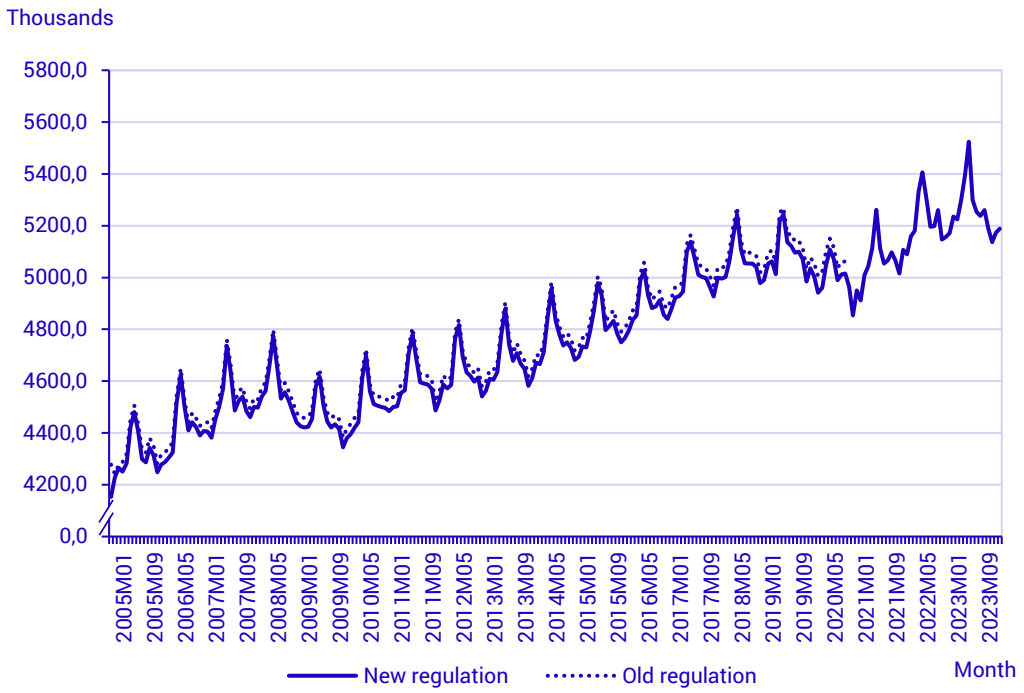
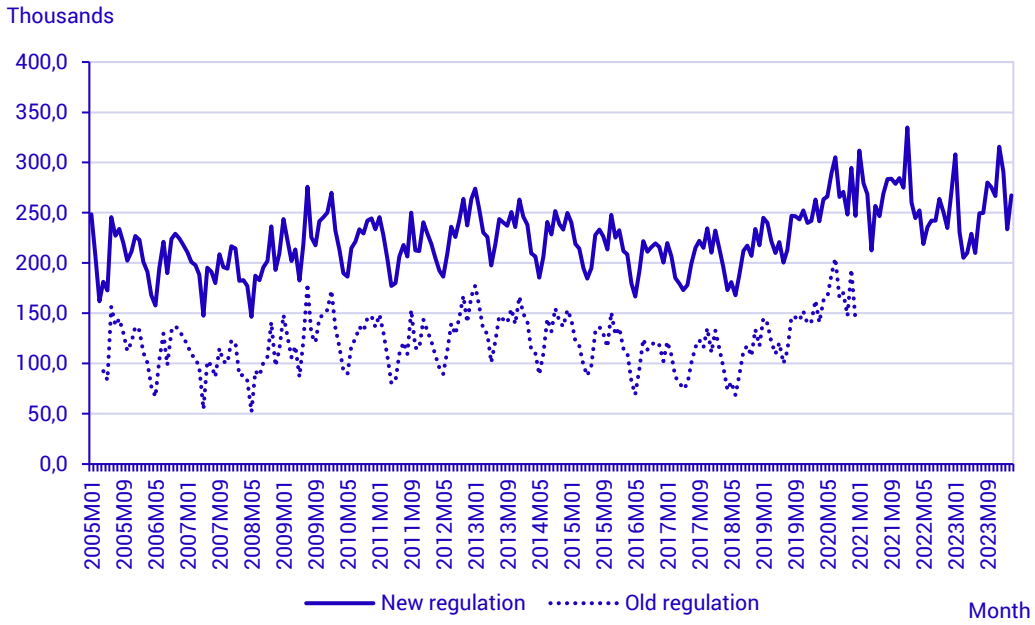


Figure 2: Monthly figures on the total number of persons available for work, but not seeking, during the time period January 2005 to March 2024 based on the old and the new regulation, respectively. The figures for the new regulation during 2005–2020 are given by the micro-linked time series.




EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL