Micro data integration for Labour Market Account

Senior Advisor Pernille Stender, psd@dst.dk and Senior Advisor Thomas Thorsen,
tst@dst.dk. Statistics Denmark, Copenhagen, Denmark

Abstract

During the last 15 years labour market statistics produced by Statistics Denmark have increasingly become more integrated. For example, the Statistics on People Receiving Public Benefits have been joined into an integrated statistical system. In this way, the quality of the statistics has been enhanced and the published figures have become logically consistent. However, the statistical users request a more cohesive statistical system covering the entire population's attachment to the labour market. The system should include volume information and provide the possibility of analyzing longitudinal labour market data.

Against this background, in the beginning of 2012 Statistics Denmark initiated work on developing an integrated statistical system for analyzing the entire population's attachment to the labour market. The statistical system is called Labour Market Account (LMA). It is intended to publish statistics from the LMA in 2014. In addition to being an important source of the future labour market statistics, the LMA must also be a direct or indirect input source to a number of other statistics within social, business and economic statistics.

This presentation gives a description of the new statistical system and of the user requirements with regard to the system. Data from the various source registers frequently contain non-permissible overlaps or inconsistent start and end dates concerning the individual states. Subsequently, the presentation describes the core of the new statistical system which is the harmonization of data from a great variety of input sources and the longitudinal data processing conducted by the rule driven engine developed for this purpose.

Key words: integrated statistics, data harmonization, data linkage, micro data.

Introduction

In 2012 Statistics Denmark began the work on developing a new integrated statistical system for labour market statistics. The integrated statistical system will make it possible to compile detailed structural statistics on the labour market in new ways.

In the new statistical system, micro data at individual level from various data sources are linked for the purpose of analyzing the population's labour market attachment (status and volume). In this context, a comprehensive data processing is performed, both of each data source separately and across the data sources. The new statistical system is called the LMA.

In the first part of this paper, a historic review is provided of the integrated register-based labour market statistics in Denmark. Subsequently, the LMA is presented and it is explained how the LMA will be able to enrich the existing labour market statistics. In the second part of this paper, a description is provided of how the LMA is built-up and which types of data processing are performed. The presentation at the ISIS Conference will primarily focus on the second part of the present paper.

1. The need for a LMA in Denmark

Statistics Denmark has a long tradition of developing integrated register-based statistical systems. The Register-Based Labour Force Statistics was developed within the labour market statistics at the beginning of the 1980´s. The statistics is compiled on the basis of an operationalization of the ILO’s guidelines and show the Danish population’s labour market status by the end of November each year. This implies that it is possible to conduct a socio-economic classification of the entire population at a comparatively detailed level. For persons in employment, the statistics is based on data linkages between each individual person and the establishment where the person works. Since 1981, the statistics has been published annually, and the statistics has been one of Statistics Denmark’s most widely used register during the last 30 years.

As a result of the labour market policy measures, which were launched as a consequence of the economic crisis in the 1990’s, there was need for further development of the labour market statistics, enabling the statistics to provide a description of the new conditions in the labour market. Some of these needs were fulfilled by refining the Register-Based Labour Force Statistics. Subsequently, new data sources concerning participants in the labour market policy measures were integrated in the statistics for the purpose of achieving a more detailed socio-economic classification.

As time passed, several separate statistics within the labour market statistics describing the development in employment and unemployment in full-time equivalents were developed. In some cases, these statistics contradicted each other with respect to levels and developments, which naturally gave rise to explanatory challenges for Statistics Denmark and interpretative problems for the users. Against this background, Statistics Denmark began by the end of the 1990´s to prepare the theoretical frameworks for an integrated statistical system, known as the LMA, which could, among other things, analyse the population’s labour market status in terms of full-time equivalents.

At the beginning, the LMA was intended as a statistical system based on micro data, but where the final product was aggregated data. The reason for this was that input data did not have the necessary content and adequacy to produce micro data statistics. It was especially a problem that the data quality concerning employee jobs was very low with respect to start and end dates for each job and with respect to volume information. This was also the reason why the Register-Based Labour Force Statistics was only compiled once every year. It was only at end-November that the quality of the time reference of each employee job was acceptable. The plans for the LMA were never realized, but by the end of the 1990´s Statistics Denmark published for the first time figures from the newly developed Working Time Account (WTA). The WTA is an integrated statistical system having common theoretical frameworks with the LMA, and producing statistics on employment, hours worked and total salary and wage costs intended for use in, e.g. the National Accounts.

In the 2000´s two important data sources were developed, making it possible to redefine the LMA into a complex micro data based statistic. Also, the great need for longitudinal data, which was present among researchers and analysts, could be accommodated.

The first important data source was the Statistics on People Receiving Public Benefits, which were integrated in a statistical system by Statistics Denmark. In this statistical system, data are subject to processing of overlaps in the terms of time and volume, implying that a person can only be incorporated with 37 hours per week as maximum (corresponding to a full time standard). In addition to information on participation rates in each measure, the statistics also contains information on start and end dates.

The Statistics on People Receiving Public Benefits contains information on persons, who are unemployed, in job activation, on early retirement pension, on early retirement pay, etc.

Secondly, Statistics Denmark gained access to a new administrative register (eIncome) containing monthly information on payments of wages and salaries and transfer incomes. Thereby, the problem of poor dates of employee jobs was to a large extent solved. In reporting data to the new register, employers were also, for the first time, to report data on the number of paid hours worked by each employee and far better volume information was thus achieved. A major development project was initiated by Statistics Denmark which purpose was to establish a register of employee jobs on the basis of the eIncome register. The register was to contain monthly information on job relations between persons and establishments. The data processing involved, e.g. selection of employee jobs, imputation of missing information on paid hours worked, enrichment of data with information on occupation and placement of each individual employee at the correct establishment.

Statistics Denmark's new register containing employee jobs was fully developed by mid-2011. Thereby, Statistics Denmark now had high quality input data sources to the LMA. Consequently, the project of establishing a LMA was initiated at the beginning of 2012. The project is established as a Prince2 project, where the steering committee is composed of representatives from social statistics, business statistics, economic statistics and user service. The steering committee is chaired by the Director of social statistics. The aim is to publish figures on the basis of the LMA during the course of 2014.

In the business case for the project the overall reasons for establishing the LMA are set out, i.e. that the statistical system must, as something new for the labour market statistics, open up the possibility of:

- compiling consistent statistics on the population's labour market status in terms of full-time equivalents
- describing longitudinal labour market statistics
- conducting status observations of the population's labour market status at arbitrary points-in-time during the year, and enable the calculation of average figures for a given period
- compiling employment, jobs, number of establishments and total wage and salary costs

In the present project period, it is not intended to develop the LMA to incorporate information on job mobility.

Also the representatives from the three external departments in the steering committee have great interests in the LMA. With respect to the department for user service, the LMA will become a register that will open up unique possibilities for researchers and analysts of labour market statistics. In relation to the department for business statistics, the LMA will, e.g. become input source for the Business Register. With respect to the department for economic statistics, employment, jobs, hours worked and total wage and salary costs from the LMA via the WTA will be incorporated into the National Accounts. Also, the LMA can potentially become a valuable data input for the economic models.

2. Set-up of the LMA

The data processing in the LMA will be conducted by the steps below:

- data processing of the individual data sources and storing of data in a harmonized, respectively, non-harmonized source database
- processing of data between data sources subject to overlaps
- compressing, summation and classification. Linkages to other registers, final data editing, if required

Source database for the LMA

The LMA is flexible with respect to the selection of data sources. This implies, in practice, that the system is established so data sources can be replaced, when new or better data possibilities are available.

The LMA's database is divided into a part with harmonized data and a part with non-harmonized data. When data are processed with respect to overlaps, it is exclusively possible to make use of information from the harmonized part of the source data base.

During the period January 2012 to April 2013 a major part of the work was devoted to analyzing which data sources are to be used by the LMA. This includes specifying which kind of checks are to be carried out in order to reveal errors and problems of consistency and how the data are to be edited before data can be stored in the source database. The data editing is implemented using a source-specific program that will perform the necessary data editing and harmonization of that particular source. In addition to this, there is a common program, ensuring that the source data are subsequently stored in a standardized form in the source database. This program defines and handles, e.g. how rows of data are entered or deleted in the source database. The overall premise is that only new, deleted or altered rows are recorded in the database.

The status by the end of April 2013 is that the LMA source database contains information from all source registers relevant at this stage of the LMA. The database includes information on:

- Unemployment, labour market policy measures, early retirement and certain other social benefits. These data are based on the Statistics for People Receiving Public Benefits, containing data that have been processed for overlaps. An analysis has revealed that the processing of time and volume overlaps and the prioritizing of data sources are not problematic in respect to the LMA[1] specifications with one exception only: People temporarily absent from the labour market because of maternity or sickness leave. Such states are sometimes overwritten in the processing of overlaps carried out. In the LMA, temporary absence must be identified. As a consequence, source data regarding maternity and sickness benefits are based on data from before processing of overlaps in the Statistics for People Receiving Public Benefits.

- Employee jobs. These data are based on the Quarterly Employee Statistics. In this register, eIncome data are further edited regarding, e.g. the linkage of jobs to establishments. A separate version of the register has been created for the LMA to

[1] In LMA the guidelines for the International Classification of Status in Employment are to be followed.

include late reporting not included in the Quarterly Employee Statistics and to improve quality even further.

- Jobs by self-employed and assisting spouses. These data are based on a variety of sources. There are four main challenges relating to data on self-employed and assisting spouses: Firstly, the information on start dates and end dates of the jobs is relatively uncertain. Secondly, it is uncertain how the activities are distributed across a given period of time. Thirdly, the activities are in many cases very small and it is therefore doubtful whether it really is a job, but on the other hand it typically cannot be ruled out in advance. Fourthly, there is no administrative information on hours worked. In the formation of the LMA sources for this group, data are compared at the individual level with information on persons included in the Labour Force Survey (LFS). On the basis of these analyses, persons with various characteristics in the LMA source data[2] and LFS information are joined and a likelihood is established for the jobs to reveal an actual activity as self-employed in the reference period. In the formation of source data an imputation of working hours by self-employed and assisting spouses is also conducted.

- Persons receiving maternity/paternity or sickness benefits, i.e. persons temporarily absent from the labour market. These data are based on data from the Statistical Register for Persons Receiving Maternity/Sickness benefits. In this register, an algorithm has been developed which – with a varying degree of certainty – determines if the person in question is absent from employment or from unemployment. This information is carried forward as a data input to the LMA, implying that jobs can be imputed in situations where there are no payments of wages or salaries, but the persons are still employed.

- Participants in education. These data are based on the Register of Educational Statistics which is a longitudinal register for this population.

In the Register-Based Labour Force Statistics there is a major group of the population which cannot be assigned any socio-economic status. In connection to the formation of the LMA source data, analysis has been carried out with the aim of reducing the size of this group. The analysis has gained only limited results, but nevertheless it has been revealed that data on participants in educational courses and students receiving public grants can reduce the size of the group. Therefore, these data and also data on retirement benefits will be used in the LMA. As these data are not to be used in the processing of overlaps, they will be added at a later stage in the data processing.

Processing of overlaps across sources

The rule engine is constructed according to the same principles as the rule driven engine used in the Statistics for People Receiving Public Benefits. The rule driven engine ensures that overlapping source data are processed according to a set of specified rules. The rules serve the purpose of:

- deleting erroneous states
- reducing the number of hours in one or more states when the states of each person is summing to a total of more than the allowed number of hours (the full-time standard)
- editing information on start- and end-dates for better longitudinal coherence

[2] The characteristics could be, e.g. purchases and sales by firms ("VAT statistics"), surpluses for self-employed in the year, whether the self-employed is employing any employees, or whether the person is insured in an unemployment fund for self-employed.

When the complex rules are set up handling overlapping states from various sources, the quality of the dates and hourly information must be taken into account. Another criterion to take into account is whether the different states are expected to, or expected not to, exist simultaneously. For the specification of rules, a number of analysis programs have been developed, mapping overlaps between different states and calculating consequences of implementing different rules for the processing of overlaps.

The processing of overlaps is the core of the LMA. Analyzing overlaps and setting up rules for the data processing have been initiated by the end of April 2013 and are expected to last until the end of 2013. The processing will be based on a weekly standard of 37 hours, so that the sum of hours for one person will not exceed this norm at any time. In addition, an alternative set of rules not limited to the 37-hour norm will be defined in order to take into account the needs of the WTA, being a future user of data from the LMA[3].

Subsequent Data Processing

After the data have been processed for overlaps, further data sources will be added with the aim of working out a detailed socio-economic classification of persons outside the labour force or in order to define the population.

On the basis of the data processed, one or more statistics registers must be defined. The registers have not yet been specified. The objective is that the registers are flexible, in order to allow for, e.g. different socio-economic classifications of the population. These classifications may be either static or dynamic.

Of course, the LMA contains unique identifications of the individual person and establishment. By the use of these identifications, background information on persons and establishments can be found in other registers. Information on persons, such as residential address, education and ancestry, is gathered from a database in the social statistics, where standardized versions of social statistics are stored. Information on establishments, e.g. address, line of industry and sector information, is gathered from frozen versions of the Business Register ensuring that the various statistics are as comparable as possible.

Conclusion

With the development of the LMA, Statistics Denmark will have a micro data register making it possible to describe the population’s labour market attachment defined in different ways with detailed background information. In addition to measuring status and volume, it will also be possible to analyze flows between different groupings, i.e. longitudinal analyses. It can be used as a direct source for publishing statistics as well as a tool for improving the quality of other statistics and registers. Furthermore, the LMA will enable Statistics Denmark to produce tailor-made solutions for its customers according to their specific needs. Finally, micro data can be offered to the research community. Consequently, the LMA will make important contributions to the description of the population, the labour market and the economy.

[3] The WTA produce information on, e.g. hours actually worked for the National Accounts as a measure of labour input into the production process. All hours worked, including those beyond 37 hours a week, have contributed to the production and must therefore be included in the WTA.