Working Papers

08/2011

Integrating administrative data into the LFS data collection.

The Spanish experience obtaining the variable INDECIL from administrative sources.

Workshop on LFS Methodology, Weisbaden

Miguel Ángel García Martínez

Javier Orche Galindo

First draft: May 2011

This draft: May 2011

Integrating administrative data into the LFS data collection.

The Spanish experience obtaining the variable INDECIL from administrative sources. Workshop on LFS Methodology, Weisbaden

Abstract

Information on the level of wages of the main job is compulsory in the LFS since 2009 (year of reference). Asking for income in household surveys is a sensitive issue that can affect the response rates and the confidence of the respondents. It was decided to obtain information from administrative sources. The Spanish LFS does not ask the personal identification number of the respondents. The solution applied in the Spanish LFS was to incorporate the PIDN (personal identification number) from the register of population matching the information for both, personal and location variables and to use this PIDN to link through the Social Security and Tax databases and incorporate the data on salaries needed to calculate the variable requested in the LFS. A general view of all the processes involved, the difficulties that we had to overcome and the main findings obtained in the preparation of the information are described.

Keywords

Labour force survey, record linkage, microintegration, combination of administrative data, validation of data sources, best estimation method

Authors and Affiliations

Miguel Ángel García Martínez

S.G. of Labour Market Statistics, National Statistics Institute of Spain

Javier Orche Galindo

S.G. of Labour Market Statistics, National Statistics Institute of Spain

Integrating administrative data into the LFS data collection.

The Spanish experience obtaining the variable INDECIL from administrative sources.

Workshop on LFS Methodology, Wiesbaden

Miguel Ángel García Martínez
Javier Orche Galindo
National Statistics Institute of Spain

Requirements and conditions

The **Regulation (EC) No 1372 / 2007** of the European Parliament and of the Council establishes the mandatory inclusion of the variable "wages from the main job" in the Community Labour Force Survey, amending accordingly the basic LFS regulation Reg. 577/1998. This will improve the analytical potential of the survey, introducing the level of wages as a classification variable in the analysis of the characteristics of the main job for the group of employees.

The Commission **Regulation nº 377/2008** stated that the variable must be coded into **deciles** and allowed the capture of the information to be provided through interviews or by using administrative records.

In the Labour Force Survey of Spain, after conducting several qualitative tests (the last one was conducted in 2003 and was financed by the Commission under grant number 2002-32100015), it was considered very problematic to include additional questions in the survey to request information on wages. The main concerns were the lack of reliability of the information obtained by interview on this topic (the studies carried out and the experience on income surveys showed that the respondents did not provide good quality answer on income) and that eventually the reluctance of respondents when answering income questions spread to the rest of labour status questions. Additional concerns were detected in telephone and proxy interviews. Both characteristics are very frequent in the Spanish LFS. Taking these problems into account, we decided to look into the possibility of obtaining the salaries information from administrative sources.

The main advantages deemed for using administrative records to estimate the variable were, first, that it **would not increase the burden on informants** and secondly, that the survey **would not be affected by a lower response rate in whole**.

The principal drawback is that it takes **more time to capture the data** since it depends on when administrative records are available. Another possible drawback would be the necessity of adaptation in case of eventual **change in the characteristics of such records over time**.

Information availability

Unfortunately and not surprisingly (otherwise the exercise would have been undertaken before), there is no administrative source that meets a suitable definition that can be managed in a straightforward way.

What we found were several administrative sources having **different methodologies and limited coverage.** In trying to find the best estimate of the target variable, we had to obtain information from various administrative records, combining their data with the information of the LFS.

Therefore, the estimation of the wage of main job is what is termed in the statistical literature as a **"derived variable"**[1]. Since the need for information can not be filled immediately by direct reference to the information available, this variable is obtained by linking different sources that provide the required information, if not with absolute precision at least with good approximation. Following this methodology, two main sources of economic and labour information have been used to estimate the salary. On the one hand, the information on income from **annual statements of withholdings and advance payments on account of personal income tax** declared to the tax agencies[2] (Form 190). On the other hand, information about **affiliation** and **contribution bases to the Social Security System** from records of the General Treasury of Social Security – (Tesorería General de la Seguridad Social. Form TC-2).

Previously, it was necessary to make up a procedure that allowed us **to assign an (correct) identifier to each of the respondents in the survey** in order to transfer and cumulate the needed information across the different sources.

Some details about the processes followed are described below.

Estimation phases

1 THE ADMINISTRATIVE PERSONAL IDENTIFICATION NUMBER (PIDN) OF THE WAGE EARNER

The LFS has the following personal data of employees, which are collected in the interview: name, sex, place and date of birth and place of residence. With this information, **the Personal Identification Number (PIDN) of the employee,** required to make successive links with administrative sources, **is obtained from the central administrative population register (Padrón de Habitantes)**. To search for personal identification numbers both deterministic and probabilistic techniques of record linkage are used.

[1] On this issue the approach described by JK Tonder *Register-based statistics in the Nordic countries. Review of best practices with focus on Population and social statistics* and A. Wallgren A. & B. Wallgren *Register-based Statistics. Administrative Data for Statistical Purposes* has been considered.

[2] State Tax Administration Agency (AEAT) and Navarre. In the period 2006-2009 it has not been possible to have data from the regional Basque Treasuries.

2 THE LINK TO THE MAIN REGISTRATION ON AFFILIATION AT SOCIAL SEGURITY IN THE SURVEY REFERENCE WEEK

The link to the General Treasury for Social Security files allows us to determinate the main job affiliation in the reference week of the survey and to get dual information:

• The **Business Identification Number (BIDN)** of the principal employer in the reference week. This BIDN will enable to continue linking with both, tax administration and social security contributions database

• The main characteristics of the contract(s). Particularly, it is crucial the **number of days worked to determine the monthly salary,** either on the whole reference year for annual totals or referred to the month of the reference week for monthly amounts.

To do this, primarily the affiliations under special schemes for self-employment or those belonging to special trading agreements are excluded (These people are affiliated but they are not really working). Then, it is assigned the affiliation corresponding to the reference week. If there are several affiliations for the same worker in the week, one must be chosen as the 'main one'. The job selected is that whose characteristics, i.e. activity of the establishment, duration of the contract, seniority in it, etc. resemble those declared in the LFS questionnaire.

Once it has been established the affiliation for the main job in the reference week, Both the Business Identification Number (BIDN) of the employer and the affiliated number of days in the year and on the month of the reference week are allotted. Other affiliation circumstances that may affect the estimation of monthly salary based on annual total are also considered.

Some **employees in the public sector**, which are not the object of holdbacks from the Social Security system but they contribute through their own mutual funds must be dealt with specifically. Given the expected stability of their employment, it is possible to estimate the number of days worked in the year by information derived from the LFS questionnaire, although the Business Identification Number (BIDN) of the principal employer may not be available in Social Security databases. This job is assumed to be unique and at least the largest in terms of revenue for the employee. This hypothesis is validated after crosschecking with tax agencies.

3 THE LINK WITH ANNUAL REGISTRATION STATEMENTS OF INCOME AND DEDUCTIONS AND INCOME TAX REVENUE ON ACCOUNT

The pair "Personal Identification Number" (PIDN) of the employee and the "Business Identification Number" (BIDN) of the employer is linked to the annual statements of income and deductions and income tax payments on account of tax agencies to get the "full annual performance ". This annual information (the only available in the Spanish Tax Administration) must be calculated in monthly estimates.

Once the link has been successfully achieved and the information obtained has been checked, a **first estimate of the monthly salary** can be made by dividing the annual full return by twelve and multiplying by the ratio between the number of days in the year and the number of days in the same year affiliated on Social Security with the principal

employer in the reference week. The following limitations must be noted in this first estimation method of the wage:

• We may have some **extra component pay** (severance payments outside the legally established, delays, etc.) included in the full work performance of the reference year that wouldn't correspond to the targeted 'monthly wage' variable.

• This is an estimate of the **wages for the whole year** and not for the month of the reference week, and the working conditions may have been changed during the year in the same company (part time to full time or vice versa, change of occupation, etc.) which may affect the wage in other months of the year.

• The tax administration in Spain is split into different agencies that must be dealt with independently (the main source of information is the national tax agency, but there are four so called 'foral' administrations.

4 THE LINK TO THE REGISTER OF BASES FROM SOCIAL SECURITY CONTRIBUTIONS

By now it must be clear that the data's journey is complex and incomplete in many cases, and that is necessary to make different assumptions in specific cases. But, on the other hand, a lot of information is gathered apart from the available data from LFS. The challenge was to transform all this information into a better estimate for each employee in the survey. In addition to the estimated wages derived from tax administration, it is also possible to calculate them from Social Security information. Doing this, both **coverage** and **edition** in the estimation of the monthly wage of the main job is improved.

Since contribution bases are recorded for each calendar month of the reference year, different calculations can be obtained:

> The annual 'average', by estimating the total salary base of contributions in the reference year divided by twelve and multiplied by the ratio between the number of days of the year and the number of days in the same year affiliated with the principal employer in the reference week, as in the previous case.
>
> Besides, an estimate can be obtained through the social security contribution base of the month of the reference week. In this case the base is multiplied by the ratio between the number of days in the month and the number of days affiliated in the month of reference.

Some limitations in the calculation of the wages by this method are:

• Contribution bases have both **maximum and minimum** limits, which makes the estimation difficult, especially in the case of the maximum limit.

• It is not applicable to employees in **mutual funds outside the General Social Security System**, for example, public servants.

• There can be two different monthly data contributions. The contribution base for common contingencies does not include the **wages for overtime** so, whenever possible,

we use the quota for work accidents and occupational diseases, which does incorporates the overtime.

5 INTEGRATION, EDITING AND IMPUTATION

As described above, in many cases it is possible to estimate salaries by **several methods** using the information available in administrative records and LFS. This enriches the possibilities for editing. In the rare event of **discrepancies** between different methods, we must first determine what is the more suitable estimate of income among all those available and validate it as the best one. Thus, the estimated final salary is obtained through a **combination of all sources** used and do not correspond exactly to the information received by any one of them.

Finally, for those employees for whom it has not been possible to establish their salary from administrative records or whose estimate was not considered sufficiently reliable, an **imputation** is made using the distribution of wages by type of time (i.e. full-time or part-time) and the occupation (three-digit standard classification according to ISCO).

6 ENCODING

Finally, the wages are sorted and **coded** into deciles from "01" to "10", corresponding to the decile "01" the group of 10 percent of employees receiving lower wages and to the decile "10" the group of 10 percent of employees who receive the highest salary.

7 SUMMARY

The LFS requests (2009 onwards) compulsory information of wage levels. This information can be obtained directly by asking respondents or through administrative sources.

We have strong doubts about the validity of the information obtained from direct questions in the questionnaire.

The variable INCDECIL in the Spanish LFS is obtained through linking the sample of employees in the LFS to the administrative sources (Social Security Affiliation and Contributions and Tax Administration).

The procedures are not straightforward. On the contrary, there are different situations that must be dealt specifically.

The estimation is calculated using all the information available gathered across the processes of linking the different sources. Key variables used for editing and imputation are the Part-Time / Full-Time working hours and the occupation (according to ISCO) of the job.

References

Official Journal of the European Union (2007). Regulation (EC) No 1372/2007 of the European Parliament and Council of 23 October 2007 amending Regulation (EC) No 577/98 on the organization of a sample survey the workforce in the Community.

National Statistics Institute of Spain (2008). Labour Force Survey. Methodology 2005. Description of the survey, definitions and instructions for completing the questionnaire.

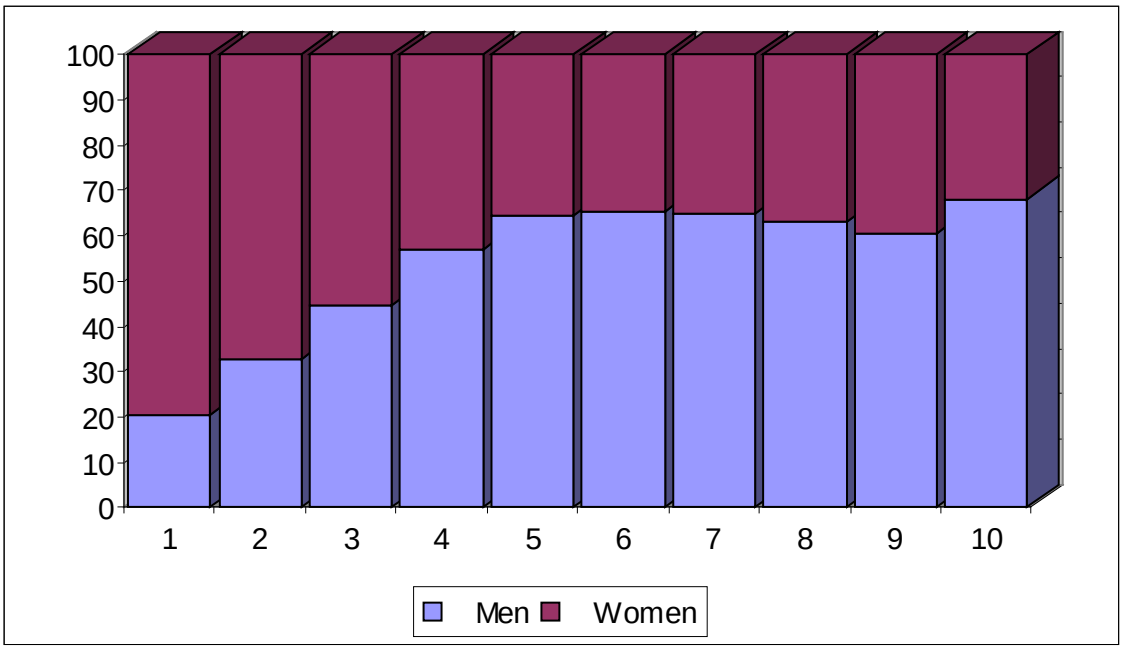
National Statistics Institute of Spain (2008). Labour Force Survey. Methodology 2005. Variables in the subsample.

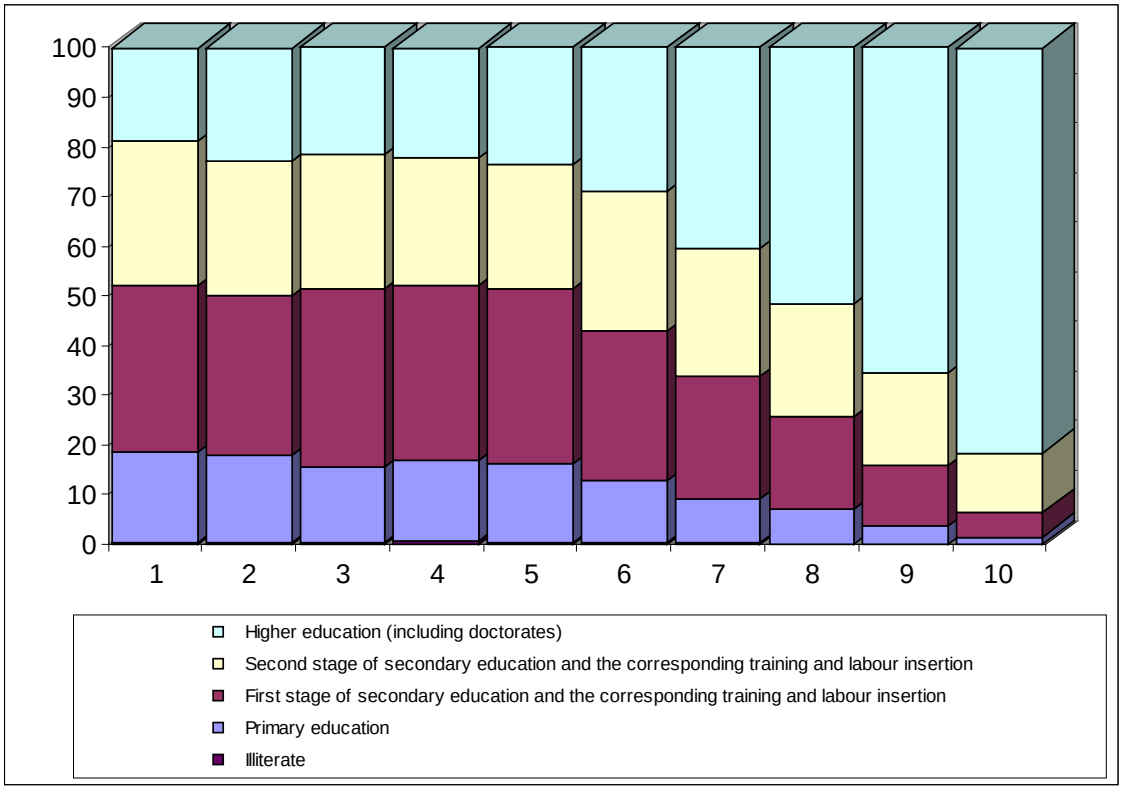
Tonder JK (Coordinator) - UNECE (2007): "Register-based statistics in the Nordic countries. Review of best practices with focus on Population and social statistics.

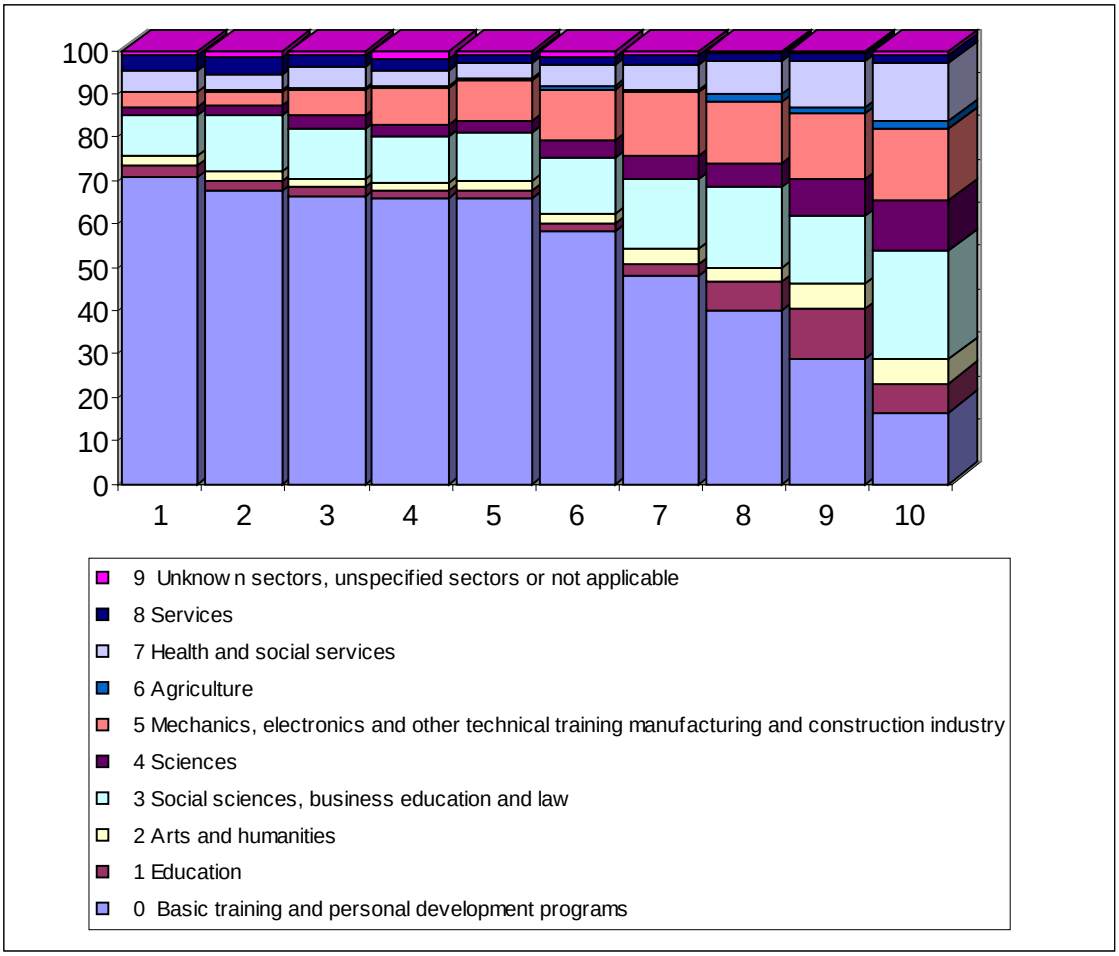
Wallgren A. - Wallgren B. (2007). Statistics Sweden. Register-based Statistics. Administrative Data for Statistical Purposes. Ed John Wiley & Sons, Ltd.

This paper was presented at Workshop on Labour Force Survey Methodology in Wiesbaden (May-2011:
http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/EN/Content/Events/LFS/PapersP/E2__IntegratingAdministrativeData__Martinez,property=file.pdf

ANNEX:
SELECTED GRAPHICS ON DECILE MAIN JOB WAGE. 2009 DATA FOR SPAIN







Average wages calculated from deciles by sex.
Gender Pay Gap calculation. 2006-2009 series.

	2006	2007	2008	2009
Total	1570,66	1635,89	1771,55	1811,48
Males	1724,31	1796,86	1961,31	2015,79
Females	1365,87	1420,11	1534,60	1576,09
Gender Pay Gap	79,21	79,03	78,24	78,19