Obtaining Statistical Information in Sampling Surveys from Administrative Sources: Case Study of Spanish LFS 'Wages from the Main Job'

Javier Orche Galindo and Honorio Bueno Maroto
National Statistics Institute of Spain
e-mail: javier.orche.galindo@ine.es

Abstract: In 2009 the variable 'wages from the main job' was added to the Spanish LFS. The information was taken from administrative registers (Social Security and Fiscal files), to avoid overburden on respondents and to improve the quality of the variable without increasing the costs. But the Spanish LFS does not ask the personal identification number of the respondents and the link with administrative registers is not immediate. The solution applied was to incorporate the PIDN (personal identification number) from the register of population matching the information for both, personal and location variables (names and surnames, addresses, date of birth, location of birth).

This PIDN assigned to the sample of employees in the LFS is used to link through the Social Security and Tax databases and incorporate the data on salaries needed to calculate the variable requested in the LFS. One intricate issue is to transform annual into monthly data referred to the reference week of the survey. Other problem is that no individual source has a complete coverage of all the employees. Consequently is was needed to combine the information from all the different sources in order to estimate the soundest wage.

Keywords: labour force survey, record linkage, population register, personal identification number, business identification number, micro integration, combination of sources, validation of sources, best estimation method.

1.The 3-dimensional record linkage strategy to assign PIDN in the LFS

1.1. The administrative personal identification number (PIDN) of wage-earner

The aim of this part is to describe in some detail the strategy followed to provide the correct personal identification number (PIDN) to the persons surveyed in the LFS in order to make possible the capture of administrative information in subsequent phases. Obviously, the PIDN can be used for linking to provide income information, as described in the second part of the paper or for any other administrative source of interest.

1.2. Persons to be search

The LFS in Spain interviews every year around 230,000 different persons with 16 and older years, distributed in 6 subsamples each of the four quarters of the year. This means that every quarter we have to obtain the PIDN of about 140,000 persons, even many of them have been interviewed in previous quarters. In the process described in this document we make a reduction of that number because we search every quarter only the 'new' persons that appears for first time, reducing to 28,000 the number of persons from whom must be obtained a Personal Identification Number (PIDN) in each quarter. These persons are searched in the Population Register database (called 'Padrón continuo').

1.3. The three dimensions

The searching in the Population Register is made selecting some criteria applied to the information, of every person, used to select candidates who could correspond to the LFS

person. Since 2005 different attempts have been testing trying to improve the output obtained (see figure 1).

D1 = Personal Data (Name, birth date, birthplace)
D2 = Residence Data (County, City, Street, building)
D3 = Human Group (Identity group: people surveyed in the same household)

The traditional variables for searching are the 'Personal variables': name-surname, birth data, birthplace, all of them are what we called as a Dimension 1 (D1). The second Dimension (D2) arose when we consider the advantages of using Street-codes that allows a fine-tuned searching by building, increasing the probability of find the 'correct' person. At the same time, the experience of the attempts showed that searching of easy-found people (with few mistakes) is quicker using a battery criteria ('waterfall searching mode'), and difficult-found people (with some mistakes in name, birth date or birthplace) are better-found using parallel mode. With moderates returns on both searching the improvement in the output is very high. In 1st quarter 2011 we obtained 99.1% of returned persons. The Criteria used to search candidates were as figure 2 shows. The return after confronting with the population register the 1st quarter was 31,550 candidates (out of 24,409 persons searched).

1.4. Record linkage process

With this file (1st quarter 2011) the process to assign the correct PIDN ran in six steps:

1. Distances: We calculate distances of four selected variables: Name, birth date, birthplace and residence. All the distances have been developed in our unit. The distances applied are based on 'coincidences' for letters (name, residence), or based on functions like 'birth codes-distance' and 'date-distance' (WMWEDD -'weighted match and weighted exact date difference') for birth date.

2. Select process: We select only one candidate, the best one within each criterion, for every person. That means the selection can be different if we have found a person in criteria1 or criteria4.

3. Segmentation: We make a segmentation of distances in order to separate the probability of being the correct person. For example, the group 100-100-100-100 (name, birth date, birthplace, residence distances) is the perfect probability, and we include in Level 10. But 100-100-100-80 has a lower probability, and then we include in level 9.

4. Iteration: After selecting the best choice in every level, we created de database with persons already linked and make another selection of candidates. This iterative process in each level is the best performance, and it avoids bad links of very close people like family members, relatives or neighbours.

When we finish this process, emerges a set of results (see figure 3). Levels 10 and 9 has a high probability of be the 'correct' person so we accept them as 'correctly' linked. The lower scores are the less coincidence of the information between the LFS and population register. Under level 6 we reject directly the link. When the scores are 8, 7 and 6 we apply additional criteria to confirm or reject these doubtful candidates. This is the moment when we use the Third dimension: the Human Group.

5. Confirmation: These doubtful candidates are solved using information of the human group, people whom has been interviewed in the same address (same dwelling). We have the identification number of the group (IDGROUP), and the selected address of the every member of the group already linked. Note that the interviewed address can be different to the address where they are currently living.

With the selection of these members we are using implicitly the probability of been part of that group (see figure 4). If the members of the human group (located by the IDGROUP) has been

selected in the same address, it means that this group live together with high probability. If we connect these two probabilities, namely the doubtful candidate and, at the same time, as being part of a human group (family, household, people living together), that live in the same address where all the members are already linked and confirmed, we can assume that we have found the correct candidate, and then we, after checking, confirm the link of this candidate as correct as well. As it can be guessed, this process is not free of exceptions, for this reason we repeat this process for different kind of human group, that means different probabilities (see table 5).

6. Manually searching: To end the process we search manually the rejected persons in the database of the Population Register.

1.5. Final results

The final output of this linkage process obtains 95.6% of automatically correct linked persons in 1st quarter 2011. But in coming quarters the aim is to reach 97-98%. At the same time the process described showed an important saving in process time, reducing the number of months involved to complete the process to 4 months, once the annual sample data is available (end of January year N+1 for survey in year N) (see figure 6). A main issue is the good quality indicators obtained. According to our analysis, we estimate in 0.9890 as a precision, with a match error of only 0.0111.

2. Case study of Spanish LFS variable "wages from the main job" (integrating administrative data into the data collection)

2.1. Requirements and conditions of the variable

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 in 2003, 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.

2.2. Information gathering

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, information about affiliation and contribution bases to the Social Security System from records of the **General Treasury of Social Security** (Form TC-2). On the other hand,. the information on income from annual statements of withholdings and advance payments on account of personal income tax declared to the **tax agencies** (Form 190)[2]. Previously (part 1 of this paper deals with this issue), 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 (see figure 7).

2.3. The link to the register of 'Affiliations and Social Security Contributions'

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

[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

in terms of revenue for the employee. This hypothesis is validated after crosschecking with tax agencies.

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

- 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.
- 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.

Some limitations in the calculation of the wages by this two methods 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.

2.4. 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 **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 third 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.

2.5. 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 (see figure 8), 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.

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).

2.6: Encoding

Finally, the wages are sorted and coded into deciles from "01" to "10", corresponding to the decila "01" the group of 10 percent of employees receiving lower wages and to the decila "10" the group of 10 percent of employees who receive the highest salary (see tables 9 and selected graphics in figure 10). From the results by deciles, some interesting indicators can be calculated (see table 11).

3. Figures and tables

Figure 1: *Historical view of searching and linking process in LFS-Spain*

	SEARCH			LINK					
	D1	D2	D3	D1	D2	D3		LINK METHOD	HIGHLIGHTS
	PERSONAL DATA	RESIDENCE	HUMAN GROUP	PERSONAL DATA	RESIDENCE	HUMAN GROUP			
2005	*			*	.		1D	Number matchs up to 8 personal variables	Deterministic link
2007	*			*	.		1D	Number matchs up to 8 personal variables	Probabilistic link
2008	*			*	.	o	1D	Number matchs up to 8 personal variables	Experimental confirmation by Human group
2009	*			*	o		1D	Sinthetic distance of 4 distances (Lebenshtein)	Experimental STREET distance
2010	*	*		*	*	o	2D	Segmentation of 4 distances (X3,WMWEDD)	1-Use of STREETCODE. 2-Experimental Quarterly. 3-Experimental confirmation by Human group.
2011	*	*		*	*	*	3D	Segmentation of 4 distances (X3,WMWEDD)	1-Confirmation by Human group. 2-Quarterly

Figure 2: *Searching criteria used in 1Q-2011 LFS-Spain*

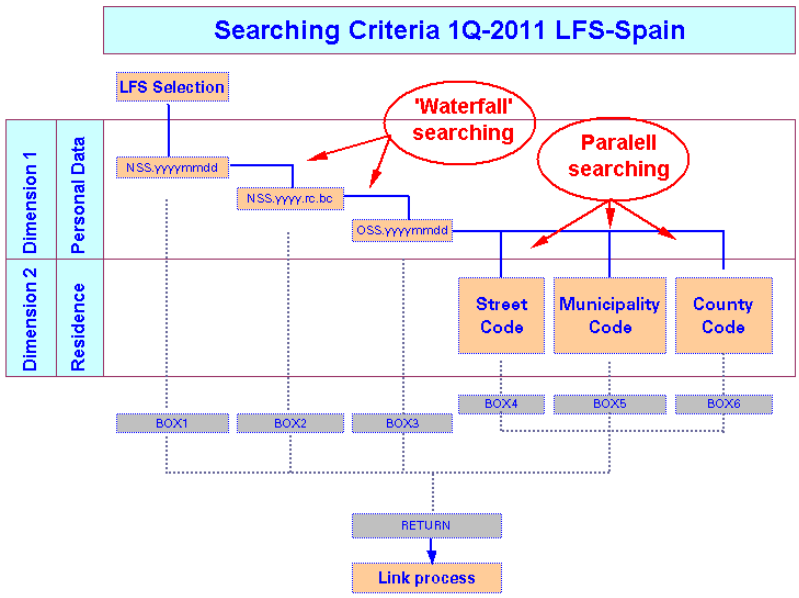


Figure 3: *Segmentation map of linkage process. LFS-Spain*

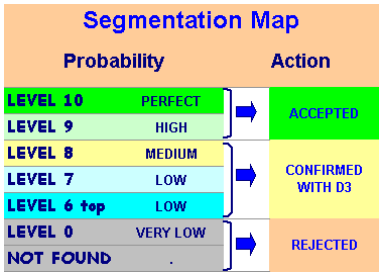


Figure 4: *Confirmation process of candidates selected using human groups variables (3rd dimension).*

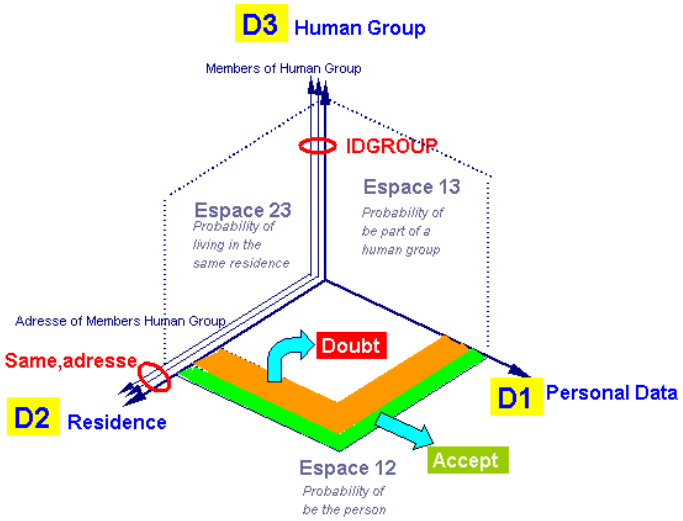


Table 5: *Typologies of human groups used*

CONFIRMATION BY DIMENSION 3 (HUMAN GORUP)		
HG1	Human Group 1	Human Groups of preselected persons (same IDGROUP, same adresse)
HG2	Human Group 2	Preselected persons with Human Groups in Selected persons (same IDGROUP, same adresse)
HG3	Human Group 3	Preselected persons without Human Group (yet) but same adresse searched and located
HG4	Human Group 4	Other

Figure 6: *Development in the process time and correct automatic ratio– LFS Spain*

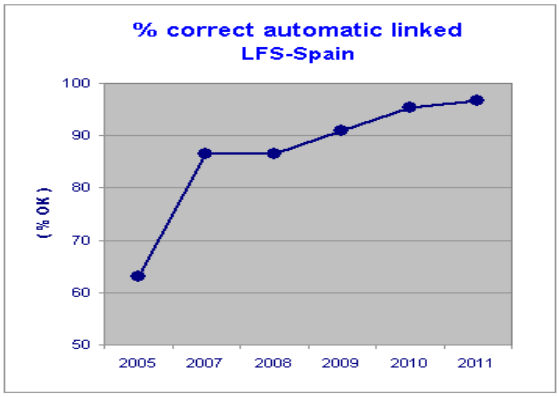
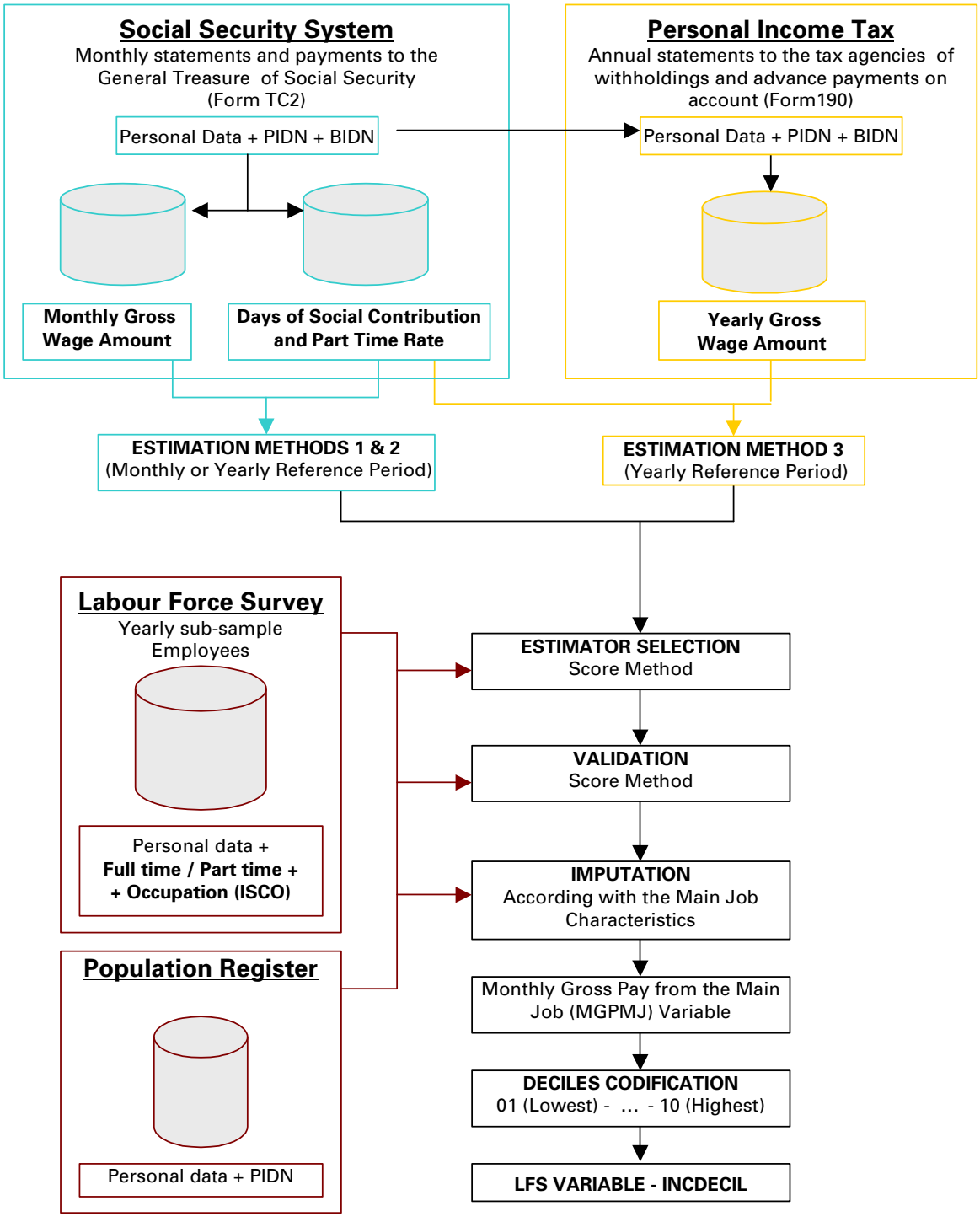


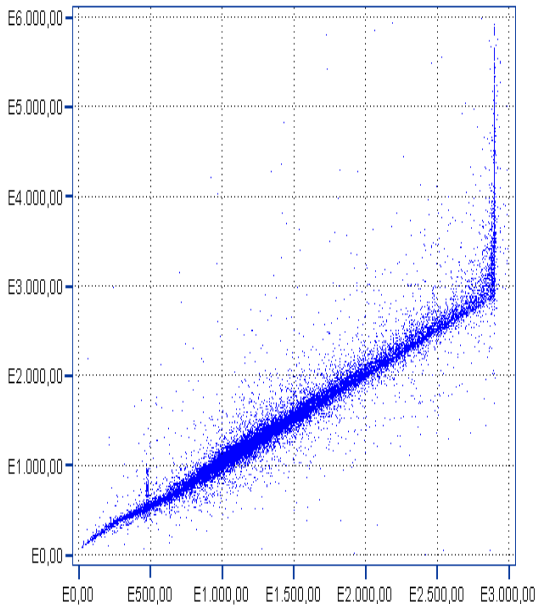
Figure 7: *Flow chart of the estimation process in the LFS variable: "Wages from the main job"*



Variables for Record linkages: (1) Personal data of the employee (Name and date of birth) + (2) Personal Identification Number (PIDN) of the employee + (3) Business Identification Number (BIDN) of the employer.
Variables for choose the best estimation method, validation and imputation:
Labour market characteristics from the LFS: (1) Full time / Part time + (2) Occupation (ISCO coded at 2 or if possible 3 digit level).

Figure 8: *Social security vs. Tax agencies estimation (2009 – employees LFS subsample)*
Monthly Gross Pay from the Main Job

Source 2: Tax agencies



Source 1: General Treasury of Social Security

Table 9: *Deciles by full time / part time and occupation (ISCO - coded at 1 digit level)*

2009 – employees LFS subsample		N	MONTHLY (GROSS) PAY FROM THE MAIN JOB (DECILE)									
			1	2	3	4	5	6	7	8	9	10
Total		**32.747**	**3.082**	**3.089**	**3.196**	**3.135**	**3.157**	**3.254**	**3.371**	**3.346**	**3.527**	**3.590**
	ISCO	**28.353**	332	2.162	2.846	3.039	3.099	3.189	3.290	3.302	3.504	3.590
Full time jobs	**0**	**236**	.	3	1	15	34	32	32	24	58	37
	1	**868**	.	13	12	27	18	28	49	93	145	483
	2	**4.681**	17	34	45	70	81	134	264	621	1.431	1.984
	3	**3.852**	20	151	256	247	312	394	656	695	603	518
	4	**3.040**	31	240	326	317	326	497	467	405	249	182
	5	**4.680**	98	606	820	740	564	490	502	386	350	124
	6	**334**	4	48	68	57	51	60	19	22	5	.
	7	**4.213**	16	159	367	668	819	664	581	487	305	147
	8	**3.066**	14	188	296	341	406	527	470	408	306	110
	9	**3.383**	132	720	655	557	488	363	250	161	52	5
Part time jobs	**ISCO**	**4.394**	**2.750**	**927**	**350**	**96**	**58**	**65**	**81**	**44**	**23**	**.**
	0	**2**	1	.	.	.	1	.	.	.	.	.
	1	**26**	16	3	1	1	.	2	3	.	.	.
	2	**447**	98	87	39	38	28	40	50	44	23	.
	3	**498**	263	111	43	13	17	23	28	.	.	.
	4	**495**	236	171	43	33	12	.	.	.	.	.
	5	**1.285**	949	235	101	.	.	.	.	.	.	.
	6	**18**	13	3	2	.	.	.	.	.	.	.
	7	**101**	63	31	4	3	.	.	.	.	.	.
	8	**124**	77	30	9	8	.	.	.	.	.	.
	9	**1.398**	1.034	256	108	.	.	.	.	.	.	.

Figure 10: *Selected graphics on decile main job wage. 2009 data for Spain*

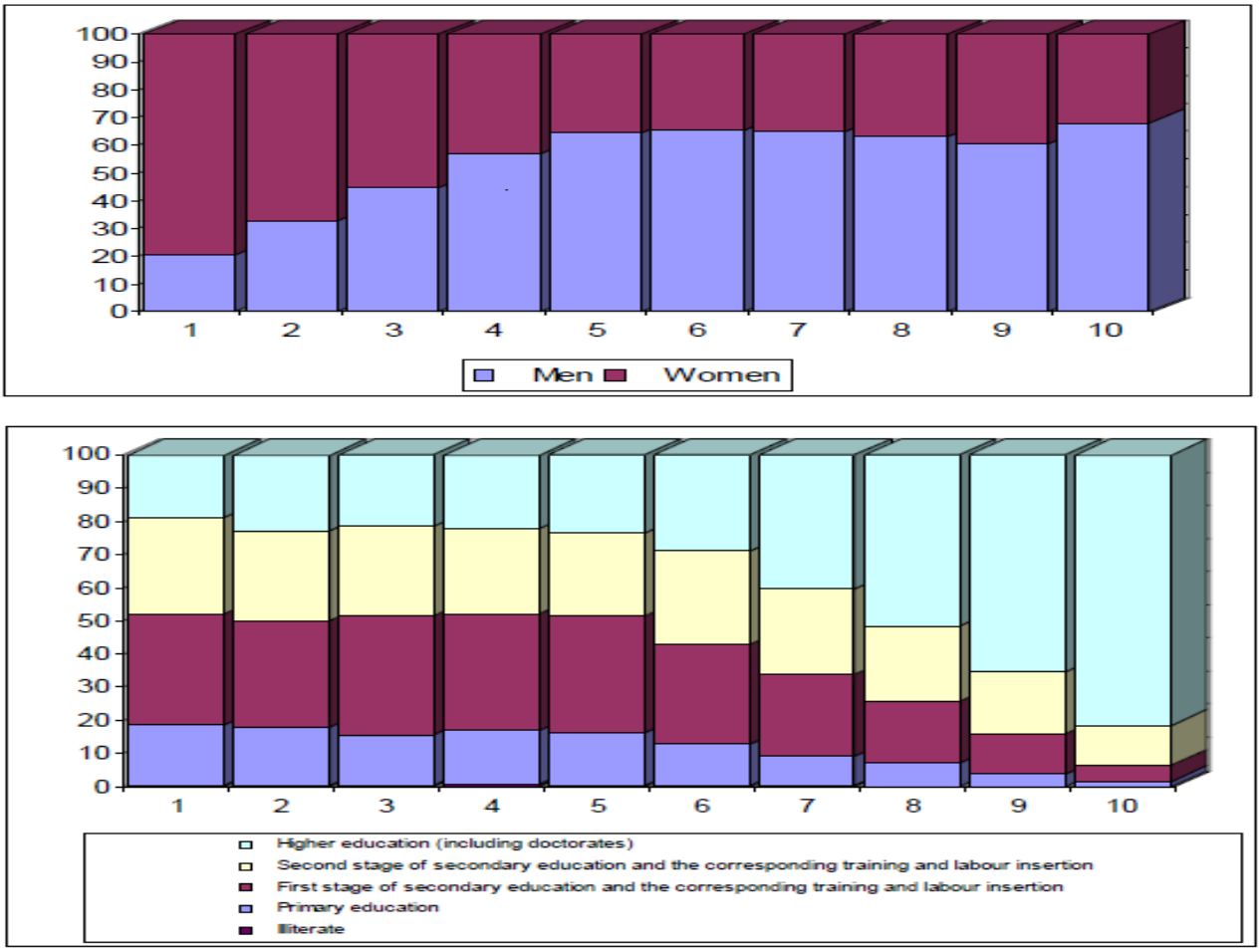


Figure 11: *Average wages calculated from deciles by sex. Gender Pay Gap calculation.*

2006-2009 series	**2006**	**2007**	**2008**	**2009**
Total	**1.570,66**	**1.635,89**	**1.771,55**	**1.811,48**
Males	1.724,31	1.796,86	1.961,31	2.015,79
Females	1.365,87	1.420,11	1.534,60	1.576,09
Gender Pay Gap	**79,21**	**79,03**	**78,24**	**78,19**

4. 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.