The Use of multi-sources for assessing the French labour cost

(Workshop on Labour Cost Data, Rome, Italy 5-6 May 2015)

Grégoire Borey, INSEE

Abstract:

The French national statistical institute uses multi-source data to complete the LCS (Labour Cost Survey).

Multisources are used in several steps during the process. This choice has been made as it allows us to

1. Improve timesavings for employers:
All information (wages and salaries) already available in administrative data is used to reduce the questionnaire burden

2. Extend the scope:
The scope of the Eurostat response should cover local units of firms with more than 10 employees in NACE B to S industries. However, the French LCS is designed to provide information on local units mainly in the private sector, that is, in NACE B to S, sector O and P excluded. The adaptation of the LCS to sectors O and P would require us to adapt the questionnaire to cover both public and private oriented sectors, which are significantly different in France. The strategy adopted by France has been to use data from a specific survey covering central government agents, rather than changing the LCS questionnaire.

3. Refresh data:
Second, the labour cost surveys are conducted every four years, with data collection covering two years. To assess the yearly variation in the labour cost and to update statistics to the year of LCS release, we use data from two administrative sources (social security declarations DADS and the system of information on civil servants SIASP, which is a dataset combining several administrative sources).

4. Validate survey information:
Some questions are asked even if we have similar information in other sources. Multisource confrontation enables us to to validate/correct our variables.

However, each of these uses has a cost, so in the following we present more precisely how we use these multisources, underlining the advantages and limits we face in practice.

The LCS

Since the early 2000s, members of the UE have committed to Eurostat to produce data on Labour Cost and Earning Structure (Eurostat's Council Regulation 530/1999, Commission Regulations 1916/2000 and 1738/2005 for SES format and Commission Regulation 1737/2005 for LCS format). Data on labour costs are released every four years, alternately with data on earning structure.

The French LCS

For this purpose, France uses an annuel survey on local units to obtain an accurate breakdown of Labour Cost components and an accurate number of worked hours for a sample of their employees. The French survey covered firms with than 10 employees in a part of the public and private sectors in metropolitan France, for cost reasons. The scope was progressively completed in order to match the Eurostat request of exhaustivity. Local Public Service agents and National Public Service agents were added in the SES2010 and LCS2012, and French Overseas Departments (Exept Mayotte) are added in the SES2014 and LCS2016.
The French LCS is collected over a period of two years. Each year about 150,000 employees and 17,000 local units are interviewed. That means the response to Eurostat is based on about 300,000 employees and 32,000 local units (about 2,000 local units are interviewed from one year to another).

Why we need multi sources

In order to reduce local units' response burden, we enriched LCS data with administrative data coming from the annual social security declaration (DADS) and the system of information on civil servants (SIASP). These administrative data give information on local units, their employees, their jobs and their wages. DADS are the INSEE reference source in order to produce statistics regarding wages and employment. They are exhaustive as each employer has the obligation to make these social statements to social security in France. For NPS agents, we instead rely on SIASP (System of Information on Civil servant). SIASP data have a higher quality than the DADS for the National Public Service. We complement this administrative information with a specific survey, which is lighter than the full LCS and with questions specifically adapted to civil servants.

Now we will see how and when we use DADS sources, SIASP sources, specific surveys and other sources in the different steps of the LCS process, from sampling to the elaboration of the Eurostat response.

1. Improve timesavings:

The main and simplest reason to use multi sources is to reduce the firms' questionnaire burden.
Studies show that response rate and response quality depend greatly on the questionnaire burden. The lighter it is the better the responses are. The main administrative source used is the DADS.

DADS presentation:

The annual declaration of social data (DADS) is a mandatory annual declaration that applies to any employer.

The DADS is used by a wide variety of social entities (pension funds, social security funds, etc.) in order to calculate social security contributions or payroll tax. Each year employers, including government, administrations and establishments, supply, for each establishment, the amount of salaries they have paid, the employed workforce and a nominative list of their employees, indicating the amount of wages received by each one. The scope of the DADS covers all employers and their employees, with the exception of employees of ministries, civil servants or not, domestic services and extra-territorial activities.

The DADS contains general information on the local unit, such as the employer's legal name, the establishment's identifier, establishment's turnover and total workforce, etc; and on employees, name, surname, id number, working period, total remuneration paid and benefits, etc.

The following data are the result of DADS published at an individual level in LCS:

- Region, department and town of work of the local unit and employee
- Region and town of residence
- Firm and local unit's workforce
- Total gross wages per firm
- Total hours paid per local unit
- Employees' sex and age
- Gross and net wages per employee
- Days worked and hours paid per year per employee
- Professional categories and manager or not (per employee)
- Firm's legal category

We use DADS for the private sector and the Hospital Public Service and Local Public Service sectors, but for the National Public Service sector, the SIASP file is more accurate.

SIASP presentation:

SIASP is a System for Information on Civil Servants that provides information about staff and payments of agents of the civil service for the three kinds of civil servants (national, local, health) since 2010. Prior to that date, we used different sources such as DADS and other administrative files. SIASP covers agents in metropolitan France and French oversea departments except Mayotte. It is a statistical file that combines several administrative sources such as the file on Central civil servants and the file originating from the General Directorate for Public Finances (DGFiP) for civil servants (excluding military personnel) of ministries and employees of certain public establishments, and the file concerning military personnel whose pay is managed by the Ministry for Defence. SIASP does not contain information on worked hours (information collected in LCS and NPS surveys).

Data from SIASP are the same (but for different individuals) as data from DADS.

Other sources

The DADS and SIASP are the main administrative sources that we use but not the only ones:
As an example, four different sources are used to estimate the Labour cost in the educational sector.

- SIASP (System of Information on Civil Servants) for workforces, wage bills, social contributions and hours paid.
- DGAFP and DGFIP data for the expenditure link to hiring, social welfare or training (D2 - Vocational training costs- must represent 0.5% of the wage and salary bill for Nace O agents and 1.5% for Nace P agents).
- National Public Service agents survey to estimate the ratio of hour worked to hour paid.
- National Accounts data for subsidies on remuneration. D391 (remuneration on subsidies) from the national accounts is used to calculate D5 (Subsidies).

2. Extend the scope:

We use both a specific survey and SIASP to extend the survey scope to NPS agents. The main source of information concerning National Public Service agents is administrative (SIASP). We use a specific survey for National civil servants to collect information on worked hours in order to complete the sector O and P response.

NPS survey presentation:

NPS survey (National Public Servants survey) is specific to National Public servants (except the Ministry of Defense). This survey was conducted for the first time in 2011 and will also be conducted in 2015 in line withthe second year of the SES survey (concerning data of 2010 or 2014). NPS survey data are then combined with SES survey data in the dataset sent to Eurostat.
This survey concerns 33,000 agents. The agents themselves are interviewed, not their employer. They can respond online or on paper. They are asked about present work, past career and working time in order to compute worked hours. The response rate of the 2010 release was about 45% versus 85% for the LCS. However, this was one of the first Internet surveys conducted by INSEE.
The NPS survey questionnaire is particularly well-suited to an assessment of the calculation of worked hours by teaching professions. In France, teachers (who are NPS agents) have a low number of class hours (around 15hrs to 18hrs per week) during school weeks. This greatly underestimates their actual working time. To calculate the hours worked by a teacher, we separately collect information on the hours paid (employment contract hours and overtime hours), on the hours spent preparing lessons/correcting work and on the hours spent doing extracurricular activities (such as parent-teacher meetings, etc.). This enables us for example to assess teacher holidays by subtracting the time spent preparing lessons and correcting work from school holidays.

3. Sampling biases:

Administrative sources (DADS and SIASP) are used upstream during the sampling process.
Samples are drawn from the administrative sources as they provide exhaustive sampling frames covering the population of interest (DADS for the private sector and Health and Local government agents; SIASP for NPS agents). We use a two-step stratified-sampling design. The first step concerns establishments and the second one concerns employees in these establishments. The sampling design aims to minimize the hourly wage variance in each stratum. The hourly wage is used here as a proxy of the hourly labour cost, which is not directly observed.

The strata chosen in the sampling design are obtained from the interaction between NACE, firm size, local unit size and location. We perform a constrained Neyman allocation of local units in the sample to minimise the variance in the estimator of the hourly wage in the strata, with a minimum number of local units per stratum. One to 24 employees per local unit are interviewed. The sampling procedure therefore consists, successively, of a calculation of the minimum number of local units in each stratum with a given precision objective, a Neyman allocation of the number of establishments (first step), a calculation of the minimum number of employees to achieve the accuracy objective for each stratum and, lastly, a constrained Neyman allocation subject to restrictions on the number of employees (second step). At the employee level, we interact each stratum defined previously with the management and non-management position.

SIASP provides the sampling frame of the NPS survey. The sampling design is stratified by sex * age group * hierarchical level (category) * professional status (civil servant or not) * ministry * location. The sampling design is obtained by minimising the hourly wage dispersion in each stratum. We also use a Neyman allocation to allocate numbers of agents interviewed in each stratum similarly to what is done for the LCS.

4. Validate survey information:

Administrative sources (DADS and SIASP) are also used for imputation/ correction of variables and calibration of the respondent sample.

Calibration:
To compute Labour Cost survey final weights for employees and local units, first we reweight the data to account for non-responses (initial sampling weights of respondents are divided by response rate in the stratum). Then we calibrate the structure of the sample of employees and local units to the margins of total population.

The population margins for the main LCS survey are:

- Social category * sex * full time/ part time (for the employee file)
- - Firm size * sector of activity * location (Ile-de-France/Other regions);

The population margins come from the DADS file corresponding to the year of the survey. The calibration variables of the employee file are:

- The unit variable (to be adapted in terms of employees)
- The total gross remuneration over the year
- The duration of pay in days
- The number of paid hours

The population margins for the NPS survey are:

- Hierarchical categories of public agents (A, B and C) * sex * civil servant or not
- Ministry * location
- Age group
- Full time / part time;

The margins in the population are calculated using the SIASP file corresponding to the year of the survey. The calibration variables are:

- Paid hours
- Gross wages
- Job duration

Imputation:

The DADS and SIASP are also used as reliable external data sources to check survey data. As an example, wages are available both in DADS and in LCS survey data. Therefore we can correct wages in the survey when they appear to be outliers or inconsistent. Each year, between 3% and 4% of the employee wage data from the survey are corrected by the employee wage data from the DADS.

Further, using multisources allows us to ask questions at a more appropriate level than the local unit level of the survey. Indeed, for some questions in the LCS, employers are more likely to have an answer at the firm level than at the local unit level. As an example, apprenticeship tax questions and training participation questions are particularly concerned. In the LCS questionnaire for these questions, the employer can choose between responding at the local unit level or at the firm level. If the employer responds for the firm, we need to reallocate the response at the local unit level. To do this, we also use information from the DADS. More precisely, we reallocate the total amount given for the firm proportionally between local units according to their numbers of employees, information we get from the DADS.
Finally, information coming from administrative sources enters into the calculation and the validation of the worked hours. The worked hours variable is crucial to compute the hourly labour cost but difficult to collect. As employers are not expected to be able to report the hours actually worked by their employees, they are asked to provide all the required information for us to do so (contractual paid time, holidays, paid leave and other absences, overtime, usual weekly days and hours, etc.). But even if those are considered reliable, some difficulties remain as in some cases we have difficulty distinguishing between paid or unpaid absences, or between real leave and compensation leave (in which there is no reduction of worktime but only compensation of unpaid overtime). So we use these data collected in the LCS to calculate hours paid and worked hours. Then, we confront the value of hours paid we calculated and those from the DADS, enabling us to validate or not the survey responses. Thus, the DADS are used both to corroborate survey data and correct them (when needed). It helps to obtain annual numbers of worked hours that are internally and externally consistent.

Concerning worked hours we use another data source for workers who have “requirement in days” contracts. Work contracts can be fixed freely between employer and employee, this is the so-called “requirement in days” contract. It is an agreement on a commitment to a number of worked days calculated on a weekly or monthly or annual basis between employee and employer. The agreement must specify the rate of remuneration. For this type of worker, we use another data source: the Labour Force Survey. Then we use the number of hours worked, usually by day (in LFS), and multiply it by the number of days in requirement (in LCS). Thus we get annualized worked hours for workers who have “requirement in days” contracts.

5. Refresh data:

To respond to Eurostat we aggregate data coming from two years of the survey. So we first need to correct data from the first year of the survey to account for inflation and legislative changes that occurred during these two years. We rely again on the DADS and SIASP for this step. Indeed, we calculate the variation (between the first and the second year) in the hourly gross wage which we then apply to cost components collected during the first year of the survey. This calculation is done on the Eurostat level data (Nace sector * firm size * location). This means that two firms of the same strata (Eurostat level data) will have the same variation rate.

6. Limitations:

Using multiple sources represents three mains challenges.

First, multiplying the sources means multiplying the time restrictions. As an example, we need to send LCS data to Eurostat before late June. We get raw LCS data in early January but to validate these LCS data we need the DADS file that is availabable only at the beginning of April.

Second, administrative data are slightly different from statistical concepts. Statistical concepts are clearly defined but still theoretical. Administrative data are supposed to be similar in definition but may vary. As an example, in France, a part of the contribution to health care (employers' contributions) was taxed after January 1st, 2013 (and not before). So in 2013, employers' contributions are measurably higher than in previous years without any practical reason. Administrative data are modified by legislative changes and we permanently need to adjust them to statistical concepts.

Third, administrative sources may evolve. In France in 2016, the DSN will replace the DADS. One of the main changes is that the file will contain monthly data instead of annual data.

Moreover when we use administrative sources we cannot modify the scope. In France, the LCS excludes military personnel (we do not have their address in administrative sources) and Mayotte (Mayotte is a recent French department and the data concerning this department are still delicate).

7. Conclusion

In a nutshell, multi sources allow us:

- To reduce the questionnaire burden on employers … but increase that on the national institute
- To extend the survey, adjusting it to some specific parts of the labour market … but sometimes the use of administrative sources forces us to exclude some of the scope.
- To consolidate our results by verifying information given in several steps (upstream and downstream from the survey process) … but our temporal leeway may be reduced
- To spread the survey workload over two years and then update the data collected a year before …. but add calculations (and so potential sources of error).