Estadística Española • Volumen 55, número 181 / 2013, pp. 231-254

How do we pursue "labormetrics"? An application using the MCVL*

José Mª Arranz
Departamento de Economía,
Universidad de Alcalá de Henares

Carlos García-Serrano
Departamento de Economía,
Universidad de Alcalá de Henares

Virginia Hernanz
Departamento de Economía,
Universidad de Alcalá de Henares

Abstract

The objective of this paper is to describe a process through which the original information of an administrative data source (the Spanish "Continuous Sample of Working Life" -"Muestra Continua de Vidas Laborales") can be organized in such a way as to permit the accurate study of work histories. It presents a comparison of the information obtained using the procedure we propose by linking the available samples (2004-2010 complete cumulative spells file) and that obtained using retrospectively only one sample (2010 data file) in three empirical applications. The aim of these applications is to demonstrate the sample loss and the potential biases which occur when data on spells and individuals are not processed correctly.

Keywords: administrative data, work histories, spells of unemployment benefits, employment

JEL Classification: C81, J60, J64

* Acknowledgements**:** Carlos García-Serrano acknowledges financial support from the Ministry of Science and Innovation (National Plan, ECO2010-19963) while José M. Arranz and Virginia Hernanz acknowledges support from the Ramón Areces Foundation. Thanks are extended to seminar and conference participants and one anonymous referee for their comments and suggestions. The authors also wish to thank Spanish Social Security for providing the data for this research. Obviously, the opinions and analyses are the responsibility of the authors. Copies of the computer programmes used to generate the results presented in the article are available, on request, to interested researchers from the first author at josem.arranz@uah.es.

¿Cómo hacemos "trabajometría"?. Una aplicación con la MCVL

Resumen

El objetivo de este artículo es describir el proceso por el cual los datos originales de la Muestra Continua de Vidas Laborales (MCVL) se pueden organizar de un modo que permite el estudio de las trayectorias laborales de las personas. Se presenta una comparación de la información obtenida utilizando el procedimiento que proponen los autores (fichero completo de episodios acumulados 2004-2010) y utilizando retrospectivamente sólo una edición (fichero de datos 2010) a partir de tres aplicaciones empíricas. El propósito de estas aplicaciones es demostrar la pérdida de muestra y los posibles sesgos en que se incurre si no se aplica un tratamiento correcto de los episodios y las personas que figuran en la base de datos.

Palabras clave: registros administrativos, historias laborales, episodios de prestaciones por desempleo, empleo

Clasificación JEL: C81, J60, J64

1. Introduction

In an article that every labour economist should read, Hamermesh (2000) described the difficulties and the dangers that await the researcher when carrying out an empirical analysis, illustrating them with examples taken from studies (some of them which he wrote himself) published in the most prestigious economics journals. A fundamental aspect of the subject in question is the initial choice and subsequent processing of a database in order to obtain answers to the research questions posed. Sometimes, a database may be employed which is unsuitable for analysing the key concept or relationship studied; on other occasions, insufficient attention may be paid to the existence of measurement errors or unusual data which a careful description would have detected; on still others, such a restrictive sample is selected that it contributes nothing to answering the question posed.

An additional element, which will be discussed in this article, concerns the form in which an already existing database, with a structure predetermined by the organisation generating it, is modified by the researcher in order to study a specific economic question, since without implementing such an alteration it would be very difficult to successfully conduct the analysis proposed. The danger in this case resides in the use of a modified database to explore questions, periods of time or groups for which the initial database was not designed, falling into the trap indicated by Hamermesh whereby the availability of the data (in this case, organising the data in a specific manner) does not imply that they can answer all the questions that we wish to address.

As with Hamermesh, our aim is to offer a vision of what can be done and what cannot in this field of empirical economics. To this end, we focus attention on the use of a database (Continuous Sample of Working Life or Muestra Continua de Vidas

Laborales, henceforth referred to as the MCVL) that, ever since the then Ministry of Labour and Social Affairs began to make it available to researchers in 2005, has given rise to an increasingly greater number of studies analysing diverse aspects of the Spanish labour market and of the Social Security system (employment, worker turnover, wages, pensions, unemployment benefits, etc.). In particular, the aim of this article is to present a process through which the original data can be organised in order to permit the study of work histories, to analyse what can be studied (and what cannot) with a database thus organised and to illustrate these questions with those studies which have been carried out to date in this area of labour economics (some of which are the result of our own research).

The article is organised as follows. In section 2, the characteristics of the MCVL are presented briefly. In section 3, we examine its use when conducting longitudinal research, illustrating this analysis with those studies which have used the MCVL for this purpose and assessing whether they fulfil the necessary prerequisites to answer the questions posed. In section 4, we propose a procedure for processing the sample in order to obtain a database organised so as to permit the analysis of individual work histories over time correctly. This same section also includes three applications that illustrate the sample loss and possible bias produced if the sample is not correctly processed. Lastly, we present our conclusions in section 5.

2. The continuous sample of working life: an overview

The MCVL provides information based on computerised Social Security system records, data from the Continuous Municipal Register of Inhabitants and, depending on the version, tax data held by the National Revenue Agency in Spain[1]. The population of reference used to create the MCVL is composed of individuals who have been paying contributions (either as registered workers, or as recipients of unemployment benefits) or receiving a contributory pension from Social Security (including pensions generated by compulsory old-age and disability insurance and state pensions for the surviving spouse or dependants) at some point in the year of reference, regardless of how long they have been in that situation.

Those registered with the Social Security system for health care, recipients of non-contributory benefits and those in receipt of national or Autonomous Community social benefits are not included in the reference population. Neither does the MCVL contain information on workers who are not registered with the Social Security system or who belong to a different insurance scheme, such as some civil servants.

Each year's sample comprises 4% of the people belonging to the reference population and is only representative of the population registered with the Social Security system in the year of reference. The size of the sample amounts to over one million people each

[1] The articles by Durán and Sevilla (2006), Argimón and González (2006), Durán (2007) and García-Segovia and Durán (2008) offer a good introduction to the characteristics and use of the sample. An exhaustive report on the information in the tax module is found in Arranz and García-Serrano (2011). Lapuerta (2010) set out some of the practical difficulties involved in handling the data.

year. Simple random sampling is used to generate the MCVL (without any kind of stratification), selecting people from the annual reference population whose personal identification code contains randomly selected figures in a determined order. These figures are identical every year. This method guarantees that the same people are selected, as long as they continue to be registered with the Social Security system, and also ensures that new entrants are representative of the registered population.

The information contained in the MCVL is organised into personal files, contributors' files and a tax data file obtained from three different sources, as indicated above. The personal files provide some characteristics of the individuals included in the sample, such as gender, age, place of residence and citizenship (these present the same distribution as that of the general population)[2]. Information on the contribution status includes some variables that merely concern identification (personal identification, national insurance contribution account code, etc.), some that correspond to the job (contributory scheme, start and termination dates, type of contract, full- or part-time, contribution group, cause for termination of contract, etc.), others that refer to the employer (industry affiliation, size of the firm, length of time as an employer, etc.) and others that concern pensioners. Lastly, tax information comes from the so-called 'Form 190', which contains a summary for each year of the sum of income tax deductions and contributions corresponding to salaries, certain economic activities, awards and attributed income of individual and corporate taxpayers.

In the subsequent sections, we focus on the form the database should take (and the information it should contain) in order to be able to study individual work histories appropriately.

3. Using the MCVL for longitudinal analyses

As mentioned earlier, the MCVL only contains representative information on individuals who have a connection with the Social Security system in the year of reference. Since the sample began in 2004 and is delivered to users each year, the information is representative of the annual cohorts from then on. By linking the data year by year together as a panel, it becomes possible to conduct a very rich and detailed analysis of individual work histories using both cross-sectional and longitudinal approaches.

A good example of cross-sectional use of the data can be seen in the work of Toharia et al. (2009). These authors estimated the number of unemployment benefit spells and recipients for each year from 2004 to 2007. In addition, they examined the personal and employment characteristics of the recipients of the different types of unemployment benefit for five points in time: on June 15th of each year from 2004 to 2007, and on December 15th of the last year. Thus, the authors used data which referred either to a year or to a specific date to examine certain aggregate questions related to employment and unemployment benefits.

[2] The Continuous Municipal Register of Inhabitants provides information on the individuals' educational level, but this should not be used since it is not up-dated correctly.

Another example of cross-sectional research using the MCVL is the study by Clemente et al. (2008), in which salary differences among employees on open-ended contracts were analysed according to whether their contracts included rebates on Social Security contributions or not. To this end they used the 2005 MCVL, selecting permanent, full-time salaried employees who were contributing under the General Tax Scheme and had been employed for at least a month in October 2005. Therefore, the analysis focused on the stock of open-ended contracts in force at that specific point in time rather than on contracts signed during that month or that year.

Nevertheless, there is no doubt that the fundamental attraction of the MCVL resides in its longitudinal nature, since the information it contains can be linked to construct complete labour histories since 2004 onwards due to its annual representativeness and the inclusion of the corresponding personal identifiers that enable the data to be combined. However, care must be taken since the MCVL is not representative before 2004 because, although it contains information about membership for as long as computerised records have been maintained, it does not hold information for that date on people who died, emigrated or became inactive without receiving a pension. This is of particular importance for certain periods (for instance, during an economic crisis, when the probability of a transition to out-of-the labour force increases) or for groups with less stability in the labour market (young people, women who have quit to take care of their children, immigrants, etc.).

In other words, if the decision is made to extend the panel data backwards from 2004 to previous years, such data would not be longitudinally representative because the information for each year from 2003 back is not representative of the population on those years. Thus, the problem would arise of focusing attention on a cohort of people exhibiting atypical behaviour at a specific point in time in 2004. This would imply assuming that the individuals who lived and worked in that year presented identical conditions (for example, low unemployment rates) and behaviour as individuals in 2003 and previous years, when this is not necessarily the case; they might even have responded differently, under different conditions, to the individuals who do not appear in 2004[3].

A very simple example will suffice to illustrate the above. If we randomly select a group of athletes who have just finished running a marathon or of individuals who have just finished their university studies, these will be representative of the population of athletes who have crossed the finishing line or of people who have obtained a university qualification, but not of those who started the race or began to study in the educational system, since there will be some who having begun will abandon the race or their education at different times. This implies that the samples mentioned would be representative if we wished to study what happens to athletes when they have to run another mile, or to university students once they have entered the labour market, but not for extracting conclusions about what happened to all the athletes after beginning the marathon or to all people after entering the educational system.

[3] This error is known as the "fallacy of historical period" or the "fallacy of cohort-centrism" (see Blossfeld et al., 2007).

Similarly, information on the people who make up the 2004 MCVL sample is representative of the population in 2004 but not of the population before that year, for example, in 1997, since a proportion of the latter would not have been included in the 2004 sample, whilst another proportion (possibly having different characteristics to the total population in 1997) would have been included. Therefore, the 2004 MCVL (linked to that of successive years) can be used to investigate what happened to the individuals who were included in 2004, but not to study what happened to those that the 2004 sample indicates were present in 1997 (who are not all from the population of that year).

This feature of the MCVL is highly relevant and should be kept in mind when posing the research question to be resolved. Let us assume that we want to study people entering the labour market and to analyse transitions from the moment of initial entry into employment and subsequent employment episodes. One possibility would be to take the MCVL from any year and, using previous spells of employment recorded for the individuals included in that year, to organise the information beginning with their first spell of employment according to the year they started, thus obtaining information for individuals who entered the labour market in, say, 1995, 1996, 1997, and so on.

However, this method would not be correct since, as Cebrián and Toharia (2008) have indicated, "this configuration [of the MCVL] would permit an analysis of the people who entered the labour market in these years [2004, 2005 and 2006], but not of those who entered in previous years, since only those who continued to stay in the system in the years of reference would be visible, and it is not possible to determine the bias that this situation produces. This represents an important limitation of the MCVL, since it implies that it is only possible to conduct methodologically correct longitudinal studies from 2004 onwards, or retrospectively provided that it is clear that past cohorts are not being studied." García-Pérez (2008) arrived at the same conclusion, affirming that "for the study of population magnitudes and for certain life cycle studies that require the use of retrospective information, the way in which the database sample has been extracted will need to be considered with care", since "in these studies it will be necessary to keep in mind that it is not possible to study past cohorts using the MCVL, but rather only those who are represented since the first sample was conducted in 2004."

It was for this reason that Cebrián and Toharia (2008) selected individuals whose first work experience as employees registered with the Social Security system occurred at some point in the year 2004, and who were followed up through a period of 730 days (two complete years), thanks to the 2005 and 2006 samples. This enabled the authors to analyse the transitions made by individuals who had entered the labour market with a temporary contract, on the one hand, or with an open-ended contract, on the other hand.

In another study, Rebollo (2011) analysed transitions toward open-ended contracts, focusing on young workers (18-28 years old) who first appeared in Social Security records with temporary contracts, for the period 2000-2007. For this, she employed the samples from 2005, 2006 and 2007 backwards, that is to say, she used the spells of employment and unemployment benefits in a prior period (2000-2007) for the individuals included in 2005-2007, creating an age cohort variable (measured at the moment of entry into the labour market) which was used in the descriptive analysis and in the estimated duration model.

In a similar manner, Izquierdo et al. (2009) examined the evolution of immigrants' wages and their salary assimilation compared to Spaniards using the historical longitudinal information provided by the 2005 MCVL. They achieved this by selecting spells of employment (under the General Tax Scheme) for the period 1979-2005 among males aged 25-54 who were represented in the 2005 sample and had entered the labour market after 1979. They then grouped these individuals into cohorts according to time of entry into the labour market proxied by the year they were first registered with the Social Security system as being employed (as in the previous case, this variable was employed for econometric and descriptive analyses).

The way in which the MCVL data was employed in the latter two studies presents obvious limitations, since the past age cohorts that were constructed using information on individuals who were included in a MCVL sample for a given year (2005 and 2007, respectively) are incomplete: some of the individuals who belonged to the population of those cohorts were not present in 2005 or in 2007 (due to death, withdrawal from the Social Security system, emigration or other reasons).

Therefore, despite the potential the database offers for studying longitudinal aspects of the labour market, the MCVL imposes a limitation on researchers: it is only possible to analyse issues that affect individuals from the moment that information extraction began (2004 onwards), and issues that affected them in the past cannot be addressed, since the samples are not representative of past cohorts. This means that although it is possible to conduct a microeconomic assessment of some public policies using information on employees' labour transitions (for instance, to study the impact of the 2006 labour reform on the duration of the different types of open ended contracts, as did Cebrián et al., 2009), it would not be possible to study (at least not without biases in the results) the impact on labour transitions of open-ended contracts resulting from employment promotion policies or of subsidies for open-ended contracts, year by year from the moment at which both policies were established (in 1997), as did García-Pérez and Rebollo (2009a, 2009b) or Conde-Ruiz et al. (2010).

Likewise, whilst it would be possible to conduct an unbiased analysis of exit from unemployment among individuals who had lost their jobs in a specific period (for instance, in 2004 or 2005), following them up over a relatively long time period (until the end of 2007), as did Toharia et al. (2010), it would not be possible to study these transitions in a wider time window covering various years prior to the moment at which the sample began, as did Rebollo (2012), given that it would generate biases which would be difficult to predict beforehand[4].

These examples of research on employment and unemployment benefit trajectories using the MCVL bring about a clear conclusion. If samples for the period 2004 onwards are available, in the case of individuals registered with the Social Security system in, say, 2010, either in receipt of benefits or working, it will be possible to obtain information on them for each year

[4] Despite the limitation indicated, it is obvious that information about the past (of work history prior to 2004) can be used to construct variables, such as the number of jobs, the number of unemployment spells, etc., which can help to explain present behaviours or results.

prior to 2010 (until 2004), since the previous history of the randomly selected individuals is recorded for each available year. Therefore, the information recorded in 2010 for persons who were registered in 2004 and continued to be so in 2010 will provide a precise account of their complete employment history from 2010 back until 2004. For those individuals with a permanent or stable connection with the Social Security system, it would be sufficient to take the last file entry and record their past information (back until 2004).

However, what happens with individuals who have a sporadic or unstable connection with the Social Security system? Take those individuals who were registered in 2004 but never appeared again in the Social Security system, or those who were only registered in 2004 and 2006 but not in 2010. For these people, the 2010 file will be insufficient because it does not contain their information for those years. This implies that the information available in the 2004-2010 samples must be processed, creating a computer programme to collect representative information on the work histories not only of those individuals who have a stable connection with the Social Security system, but also of those who have an unstable connection.

In sum, the aim would be to have a database that for the 2004-2010 observation window -the period representative of the sample- contained all the work histories of those people who had had a connection with the Social Security system, whether for one, two or seven years. In the following section, we explain how to create a programme (using some standard statistical package) which collects the work history of all individuals in the representative observation period, and present three applications that illustrate the sample loss and the biases which occur if the information is not processed correctly.

4. Reconstructing information for a representative work history observation period

4.1 Description of the procedure

If the objective is to construct a database containing the work histories of those individuals who had some kind of connection with the Social Security system during one or all of the years available to date, it will be necessary to use a procedure whereby the annual MCVL files are obtained and the steps necessary to combine these files are taken. Given the particular features of the database (shown previously), a programme must be created that saves the information accurately and correctly. We propose the following method:

(1) Information is selected on the spells pertaining to people who were registered and had some connection with the Social Security system in the last year containing representative data, and the information this provides about past registration in the reference years is saved. For instance, if the last year is 2010, the information on those individuals with some connection with the Social Security system in 2010 is saved, together with all their information on past registration (from 2009 backwards).

(2) The registration file for 2009 is examined (linked to that of the individuals) to select spells pertaining to those people who had some connection with the Social Security

system in 2009 but not in 2010 (and who are not, therefore, included in step (1)) and obtain the corresponding registration information for these individuals for previous years (from 2008 backwards).

(3) The registration file for 2008 is examined (linked to that of the individuals) to select spells pertaining to those people who had some connection with the Social Security system in 2008 but not in 2009-2010 (and who are not, therefore, included in steps (1) and (2)) and obtain the corresponding registration information for these individuals for previous years (from 2007 backwards).

Similar steps (steps (4), (5), (6) and (7)) are to be taken for years 2007, 2006, 2005 and 2004. Through examining and saving the data in this way, it is possible to collect information relating both to those people who are always part of the MCVL because their connection with the Social Security system is stable and to those whose connection is interrupted and who thus do not appear again in the period of observation from 2004 onwards. The procedure which follows all the steps in the programme generates a file containing present and past information from all the registration files from 2004 to 2010. This file will be called the complete cumulative spells file (CCSF).

Having presented this programme, three empirical applications are described below to illustrate its advantages. One shows how the work histories of individuals with information concerning registration with the Social Security system (whatever their contributory status in a given year) are lost when the procedure described above is not implemented. Another reflects the sample selection bias produced in spells of employment or unemployment with benefit when that procedure is not implemented. The third focuses on an analysis of labour turnover and the duration of employment and unemployment in receipt of benefits.

4.2 First application: an analysis of patterns

Using the CCSF, Table 1 provides the patterns of those individuals for whom information on registration is available in one or more of the seven MCVL samples, identified as a result of applying the programme described above. The trajectories have seven positions, one for each year between 2004 and 2010: the value 1 indicates that the sample has information on an individual's registration in that year, while the value 0 indicates that information is not available and, therefore, does not appear in the sample for that year. The patterns are organised in descending order according to their weight in the total. For the sake of space, we do not show all possible patterns in the Table but only the first 42 types of trajectories.

Table 1

Patterns of individuals' registers in the MCVL (2004-2010).
Total number of individuals: 1,117,226

Frequency	*Percent*	*Cumulative (%)*	*Patterns*	*Type*
651,609	58.32	58.32	1111111	A
54,705	4.90	63.22	0111111	A
32,170	2.88	66.10	0001111	A
32,018	2.87	68.97	0011111	A
27,431	2.46	71.42	0000111	A
26,866	2.40	73.83	0000001	A
24,256	2.17	76.00	1111110	B
23,995	2.15	78.14	0000011	A
22,889	2.05	80.19	1111100	B
20,731	1.86	82.05	1100000	B
20,540	1.84	83.89	1000000	B
19,791	1.77	85.66	1111000	B
18,879	1.69	87.35	1110000	B
8,115	0.73	88.07	0000100	B
6,856	0.61	88.69	0001100	B
6,839	0.61	89.30	0001000	B
6,068	0.54	89.84	1011111	A
5,956	0.53	90.38	0000010	B
5,386	0.48	90.86	0000110	B
4,897	0.44	91.30	0010000	B
4,529	0.41	91.70	0111100	B
4,438	0.40	92.10	1101111	A
4,424	0.40	92.50	1111101	A
4,095	0.37	92.86	0111110	B
4,053	0.36	93.23	0100000	B
3,963	0.35	93.58	0001110	B
3,890	0.35	93.93	0011000	B
3,732	0.33	94.26	0011100	B
3,697	0.33	94.59	0110000	B
3,335	0.30	94.89	0111000	B
3,145	0.28	95.17	1110111	A
2,962	0.27	95.44	1111011	A
2,906	0.26	95.70	0011110	B
2,741	0.25	95.94	1001111	A
1,953	0.17	96.12	0000101	A
1,897	0.17	96.29	0001101	A
1,862	0.17	96.45	0101111	A
1,744	0.16	96.61	1100111	A
1,646	0.15	96.76	1111001	A
1,498	0.13	96.89	0001011	A
1,463	0.13	97.02	1000111	A
1,357	0.12	97.14	1110011	A
19,964	1.80	98.95	XXXXXXX	*Other patterns type A*
11,935	1.06	100.00	XXXXXXX	*Other patterns type B*

Note: Total type A=83.3%; Total type B =16.7 %.

There are 127 possible patterns representing 1,117,226 different individuals. The patterns are designated by one of two letters (A or B). These letters have been assigned to show the loss of possible patterns in the sample which occurs if the data are not correctly processed.

Letter A indicates the information on the patterns of registered individuals who had a connection with the Social Security system between 2004 and 2010 that is obtained when only the 2010 file on registered individuals is considered and information on past registration saved (from 2010 back to 2004). This information corresponds to the application of step (1) of the programme described in the previous section. 64 possible patterns are identified, representing 83.3% of the different individuals.

Letter B indicates the additional information obtained when the complete programme described above -steps (1) to (7)- is applied. Unless this is the case, incomplete information on the employment trajectories of the individuals included in the MCVL sample will be obtained: incomplete because it does not include 63 patterns, accounting for 16.7% of the total patterns. Therefore, sample selection bias is produced if only one registration file is considered (2010) rather than the complete linkage of MCVL files (2004 to 2010).

To complement the previous analysis, we estimated a model that would enable us to identify the determinants of the probability of having been included in the MCVL in one of the prior years without having been present in the last one; that is to say, the determinants of the sample loss which occurs when the MCVL is not processed correctly due to not using all the available information. The primary objective of this analysis was to show, as clearly and intuitively as possible, that this sample loss is not random and that some groups (classified according to different personal characteristics and employment situation) are more prone to being part of this sample loss. Thus, researchers conducting studies which focus on these groups should be especially aware of the problem that this loss of information implies for their results.

To carry out this estimation using the two types of patterns described in Table 1, a dichotomic variable was used as the model's dependent variable. This variable takes the value of 1 when an individual is included in the MCVL in one of the years prior to 2010 but not in this last year (individuals with type B patterns) and the value 0 if the individual is present in 2010 independently of whether he or she appeared in previous years (individuals with type A patterns).

This variable, which summarises the patterns of individuals throughout the seven currently available years, enabled us to conduct an intuitive analysis of whether the probability of belonging to category B (sample loss) is random or is predominantly associated with specific groups of individuals. Thus, we analysed whether this probability varied according to gender, age group, region, citizenship or the individual's employment situation throughout the year (defined on the basis of three continuous variables indicating the percentage of spells of unemployment, self-employment and temporary employment). We estimated a probit model.

Table 2 gives the estimation results of the probit model described. The estimated parameters show that those groups whose relationship with the labour market is more

unstable are more prone to being part of the sample loss which occurs when the sample is not processed appropriately. In particular, women, the youngest (less than 25 years old) and oldest (over 55 years old) workers and foreigners presented a greater probability of belonging to the sample lost when the data processing programme described above is not applied correctly.

Table 2.

Estimation results of the probit model on the probability of being present in the MCVL in some of the samples prior to that of 2010 but not in 2010. Source: MCVL (2004-2010).

	Coefficient	*Standard Error*	*Significance*
Intercept	-0.965	0.004	***
Gender (1=Women)	0.080	0.002	***
Age groups			
16-24 years	-	-	-
25-30 years	-0.247	0.003	***
31-45 years	-0.285	0.003	***
46-55 years	-0.243	0.003	***
56-65 years	0.222	0.003	***
Regions			
Andalucía	-	-	-
Aragón	-0.082	0.005	***
Asturias	-0.017	0.006	***
Baleares	-0.042	0.005	***
Canarias	0.030	0.004	***
Cantabria	-0.006	0.007	
Castilla-La Mancha	-0.080	0.004	***
Castilla y León	-0.092	0.004	***
Cataluña	-0.038	0.003	***
C. Valenciana	0.048	0.003	***
Extremadura	-0.146	0.006	***
Galicia	-0.039	0.004	***
Madrid	-0.019	0.003	***
Murcia	-0.025	0.005	***
Navarra	-0.034	0.006	***
País Vasco	-0.110	0.004	***
La Rioja	-0.080	0.009	***
Ceuta	0.056	0.020	***
Melilla	0.148	0.020	***
Missing Values	0.767	0.020	***

Note: *** means statistically significant at 1 percent.

Table 2.

Estimation results of the probit model on the probability of being present in the MCVL in some of the samples prior to that of 2010 but not in 2010. Source: MCVL (2004-2010). (Continued)

	Coefficient	*Standard Error*	*Significance*
Citizenship (1=Spanish)	-0.743	0.002	***
Labour market status			
%Temporary employ. spells	0.660	0.002	***
%Unemployment benefit spells	0.594	0.003	***
%Self-employment spells	0.346	0.002	***
Log-likelihood function		-1,601,777	
Observations		6,079,337	

Note: *** means statistically significant at 1 percent.

Furthermore, in terms of employment status, the results also clearly indicate that the composition of the sample used can significantly be altered too. In particular, those individuals who had experienced more frequent spells of unemployment benefit or temporary employment exhibited a lower probability of being present in the sample when only the last available year is considered.

4.3 Second application: counting the number of spells starting each year and comparison with the Labour Statistics

This application also illustrates the sample selection bias which occurs if the MCVL is not processed appropriately. Here we focus on a particular case: the total number of spells of employment and unemployment in receipt of benefits that began in each of the years in the period 1997-2010. The aim is to determine whether this number coincide or not when using the two procedures considered: namely, the procedure which follows all the steps in the programme described in section 4.1, generating the complete cumulative spells file (containing present and past information from all the registration files from 2004 to 2010), and the other procedure, which only considers step (1) of the programme, generating the 2010 data file (which only uses the present and past information contained in the registration file for the year 2010). Note that we have selected years belonging to the sample reference period (2004-2010) but also years prior to that (1997-2003). Our objective is to show how the amount of spells changes (declines) as we move back further in time from the reference period.

The top panel of Table 3 shows the information for the total number of spells of employment and unemployment with benefit that began not only during the period of reference of the sample but also some years before, obtained using the two files described above. There are four columns in this Table. The first column contains the complete cumulative spells file information referring to spells of employment or unemployment with benefits for each year. The second column contains the same information obtained from the 2010 data file. The third column highlights the differences between both files. Finally, the last column gives the

ratio between the information obtained from the 2010 data file and that captured by the complete cumulative spells file (shown as a percentage).

Table 3

Total number of spells (either of employment or of covered unemployment) beginning each year. Source: MCVL.

Panel (a): Information obtained from the 2004-2010 complete cumulative spells file (CCSF0410) and the 2010 data file (F2010).

	CCSF0410	F2010	F2010-CCSF0410	Ratio (F2010/CCSF0410)*100
1997	14,382,000	13,405,250	-976,750	93.2
1998	16,522,025	15,346,975	-1,175,050	92.9
1999	18,572,275	17,164,775	-1,407,500	92.4
2000	18,722,525	17,180,300	-1,542,225	91.8
2001	19,046,350	17,311,025	-1,735,325	90.9
2002	22,058,600	19,937,750	-2,120,850	90.4
2003	22,423,450	19,997,800	-2,425,650	89.2
2004	23,766,525	20,963,900	-2,802,625	88.2
2005	25,674,075	22,606,350	-3,067,725	88.1
2006	26,494,225	23,340,775	-3,153,450	88.1
2007	27,496,400	24,325,575	-3,170,825	88.5
2008	27,147,725	24,678,975	-2,468,750	90.9
2009	28,146,050	27,124,825	-1,021,225	96.4
2010	28,938,750	28,938,750	0	100.0

Panel (b): Information obtained from the 2004-2007 complete cumulative spells file (CCSF0407) and the 2007 data file (F2007).

	CCSF0407	F2007	F2007-CCSF0407	Ratio (F2007/CCSF0407)*100
1997	14,274,875	13,710,650	-564,225	96.0
1998	16,381,425	15,703,850	-677,575	95.9
1999	18,418,275	17,602,975	-815,300	95.6
2000	18,555,200	17,653,850	-901,350	95.1
2001	18,844,175	17,858,425	-985,750	94.8
2002	21,803,275	20,611,250	-1,192,025	94.5
2003	22,167,650	20,755,275	-1,412,375	93.6
2004	23,501,525	21,909,250	-1,592,275	93.2
2005	25,303,525	23,883,725	-1,419,800	94.4
2006	26,060,800	25,174,550	-886,250	96.6
2007	26,924,600	26,924,600	0	100.0

Note: authors' own calculations

As can be observed, the number of spells that are "missing" in the 2010 data file when compared to the complete cumulative spells file is enormous: about 3 million each year of the period 2004-2007, 2.5 million in 2008 and 1 million in 2009. With the exception of 2010, which coincides 100% by definition, the 2010 data file captures around 9% and 4% less information than the CCSF for the 2008 and 2009, respectively, and approximately 12% less for the years 2004 to 2007. As we move far to the past, this proportion remains in the range 7%-10%.

The bottom panel of Table 3 provides similar information to that provided in the top panel but taking 2007 (instead of 2010) as the reference year. This means that the CCSF now contains present and past information from the registration files from 2004 to 2007, while the 2007 data file only uses the present and past information contained in the registration file for 2007. The aim of selecting 2007 (an expansion year) is to compare the results with those obtained with 2010 (a recession year) in order to examine whether the business cycle (and, thus, the higher or lower probability of having workers not connected with the Social Security) affects the number of spells captured.

In this case, the number of spells not recorded by the 2007 data file amounts to less than 1 million in 2006 and about 1.5 million in 2004 and 2005, so this file captures between 4% and 7% less information than the CCSF not only for the period of reference (2004-2007) but also for the years before that. Therefore, the difference is smaller than the one of the previous comparison. The reason does not rest on the fact that the CSSF captures not so many spells (in fact, the difference between the first column of both panels is less than 0.5 million spells during the period of reference) but hinges upon the number of spells captured retrospectively by the 2007 and 2010 data files (the difference between the second column of both panels). As can be seen, the amount of spells obtained for each year of the period declines substantially if we use 2010 instead 2007, i.e. the longer the distance between the data file used and the year for which the information is extracted. This result allows us to conclude that the difference in the number of spells captured when we compare a CCSF and the data file of a particular year does not depend crucially on whether this file corresponds to an expansion or a recession year.

We now focus on the number of spells of receipt of unemployment benefit. Table 4 provides the information on the total number of spells of benefits that began during 1997-2010[5]. There are three grand columns in this table. The first column contains the information obtained from the complete cumulative spells file referring to spells of receipt of unemployment benefits beginning in each year, distinguishing between entitlements that start immediately after the loss of a job or due to other reasons (mainly after the exhaustion of other unemployment benefit). The latter is the difference between the total and those resulting from employment. The second grand column contains the same information obtained from the 2010 data file. Lastly, the third column highlights the differences between

[5] This and the following table do not include the agricultural subsidy because the Annual Report on Labour Statistics (*Anuario de Estadísticas Laborales*) does not contain this information for the spells of unemployment benefit starting a given year and, as will be seen subsequently, one of our objectives is to compare the figures from the Annual Report with those for entries into unemployment benefit receipt obtained from the MCVL.

both files, giving a ratio between the information captured by the 2010 data file and that obtained from the complete cumulative spells file (shown as a percentage) for the total information, broken down according to the origin of the benefit entitlement[6].

Table 4.

Total number of spells of unemployment in receipt of benefit beginning each year (entries into de unemployment compensation system). Information obtained from the 2004-2010 complete cumulative spells file (CCSF0410) and the 2010 data file (F2010). Source: MCVL.

	CCSF0410			*F2010*		
	Total	*From employment*	*Others*	*Total*	*From employment*	*Others*
1997	2,406,400	1,780,450	625,950	2,259,775	1,669,900	589,875
1998	2,085,375	1,684,350	401,025	1,947,675	1,573,250	374,425
1999	2,291,500	1,795,325	496,175	2,127,050	1,668,650	458,400
2000	2,370,725	1,865,550	505,175	2,189,200	1,725,850	463,350
2001	2,688,850	2,070,325	618,525	2,469,000	1,903,925	565,075
2002	2,959,850	2,267,725	692,125	2,692,600	2,069,975	622,625
2003	3,111,275	2,430,800	680,475	2,786,775	2,193,575	593,200
2004	3,291,400	2,527,325	764,075	2,936,625	2,274,375	662,250
2005	3,382,300	2,571,575	810,725	3,026,025	2,314,650	711,375
2006	3,470,225	2,667,925	802,300	3,114,350	2,413,175	701,175
2007	4,044,100	3,066,100	978,000	3,644,825	2,776,300	868,525
2008	5,560,625	4,284,175	1,276,450	5,157,400	3,983,100	1,174,300
2009	9,126,325	4,954,375	4,171,950	8,958,975	4,843,375	4,115,600
2010	9,557,000	4,725,425	4,831,575	9,556,675	4,725,325	4,831,350

(Continued)

	F2010/CCSF0410 (%)		
	Total	*From employment*	*Others*
1997	93.9	93.8	94.2
1998	93.4	93.4	93.4
1999	92.8	92.9	92.4
2000	92.3	92.5	91.7
2001	91.8	92.0	91.4
2002	91.0	91.3	90.0
2003	89.6	90.2	87.2
2004	89.2	90.0	86.7
2005	89.5	90.0	87.7
2006	89.7	90.5	87.4
2007	90.1	90.5	88.8
2008	92.7	93.0	92.0
2009	98.2	97.8	98.6
2010	100.0	100.0	100.0

Note: authors' own calculations.

[6] Table 4 includes the administrative errors that appear in the sample when the cause for the start of benefit receipt is not known or is a lost value. The results with and without these administrative errors are almost identical.

The first feature to mention is that the number of spells of unemployment benefit that began in 2010 is identical in both files (the ratio is 100%). This result is logical because the spells in 2010 were almost the same[7]. However, for 2009 the 2010 data file contained nearly 2% less unemployment benefit entries than the complete cumulative spells file. This percentage of information lost rises as we go further back in the period of reference (exceeds 10% in 2004-2007); and it ranges between 6% and 10% in the years before that period. If we focus on the origin of the start of unemployment benefit receipt, this missing information continues to exist whether it is due to loss of a job or to exhaustion of a different benefit. The reduced percentage of information obtained from the 2010 data file compared to the complete cumulative spells file for the period 2004-2007 is around 10% when receipt originated in the loss of employment and 11%-13% when originating in entitlement to another benefit.

Lastly, Table 5 provides a comparison of the information obtained from the complete cumulative spells file and the 2010 data file with the information on entry into the unemployment compensation system obtained from the Annual Report on Labour Statistics (Anuario de Estadísticas Laborales) published by the Ministry of Employment based on data from the Public Employment Service. The first column shows the data from the Annual Report[8]. The second and third columns provide the ratio between the total number of entries in the complete cumulative spells file and the Annual Report, given as a percentage, and between the total number of entries in the 2010 data file and the Annual Report, given as a percentage, respectively. If these ratios are equal to 100, the information in the Annual Report and complete cumulative spells file (2010 file) coincides; if they are greater than 100, the complete cumulative spells file (2010 file) contains more information than the Annual Report; and if they are lower, the opposite is true. In this Table, we also distinguish between the entitlements that start immediately after the loss of a job and the remainder[9].

[7] There are very small differences which may be due to the elimination of repeated spells that are observed when the years in the CCSF are combined rather than to processing the *2010 data file* separately.

[8] The Annual Report provides the total number of entries after excluding the subsidy for agricultural temporary workers ("subsidio de trabajadores eventuales agrarios"). Furthermore, to make figures comparable, since the MCVL does not include some types of benefits such as the "Renta Activa de Inserción" (RAI) and the temporary programmes designed to help the unemployed after the exhaustion of UI and UA (the so-called "PRODI", passed after the Real Decreto-ley 10/2009), we should subtract the entries into these programmes from the total. However, it is impossible to do it for the RAI for the whole period since after 2005 they are merged into "Other reasons" (until then, the yearly number of entries was below 100,000). Therefore, we have subtracted only the entries into the PRODI (255,501 in 2009 and 822,137 in 2010).

[9] To take account of this difference in the Annual Report, we consider that the entries after loss of employment are due to the following reasons: individual firings, collective layoffs and ending of temporary contracts. The rest of entries are caused by exhaustion of a previous benefit, short-time and reduction of working hours, voluntary quits, specific causes for the restart of the entitlement, and other reasons. However, a change of codification in 2010 makes it difficult to maintain the homogeneity of the series.

Table 5.

Number of entries into the unemployment compensation system obtained from the Annual Report on Labour Statistics (ARLS) and comparison with the information obtained from the 2004-2010 complete cumulative spells file (CCSF0410) and the 2010 data file (F2010)

	ARLS		*CCSF0410/ARLS (%)*		*F2010/ARLS (%)*	
Años	*Total*	*From employment*	*Total*	*From employment*	*Total*	*From employment*
1997	2,775,165	2,092,898	86.7	85.1	81.4	79.8
1998	2,453,235	1,985,035	85.0	84.9	79.4	79.3
1999	2,561,449	1,996,235	89.5	89.9	83.0	83.6
2000	2,592,407	2,033,562	91.4	91.7	84.4	84.9
2001	2,774,045	2,173,265	96.9	95.3	89.0	87.6
2002	3,146,449	2,447,936	94.1	92.6	85.6	84.6
2003	3,149,006	2,512,033	98.8	96.8	88.5	87.3
2004	3,342,041	2,636,650	98.5	95.9	87.9	86.3
2005	3,440,086	2,718,191	98.3	94.6	88.0	85.2
2006	3,553,329	2,827,155	97.7	94.4	87.6	85.4
2007	3,992,332	3,042,054	101.3	100.8	91.3	91.3
2008	5,506,529	4,323,486	101.0	99.1	93.7	92.1
2009	8,954,132	5,143,164	101.9	96.3	100.1	94.2
2010	9,340,403	4,345,776	102.3	108.7	102.3	108.7

Note: authors' own calculations. The heading "Total" of the Annual Report gives the total number of entries after excluding the subsidy for agricultural temporary workers ("subsidio de trabajadores eventuales agrarios") and the temporary programme for helping the unemployed after the exhaustion of UI and UA (the so-called "PRODI"). The heading "From employment" of the Annual Report gives the number of entries due to the following reasons: individual firings, collective layoffs and ending of temporary contracts. A change of codification in 2010 makes it difficult to maintain the homogeneity of the series.

According to data from the Public Employment Service, there were some 3.3-3.5 million entries (spells) into the unemployment compensation system yearly in the period 2004-2006. These numbers increased to nearly 9 million in 2009 and 9.3 million in 2010. These figures generally coincide with those obtained from the complete cumulative spells file, so the ratio between both sources of information is around 98%-102% during the period of reference of the sample[10]. When the analysis is repeated for the number of entries due to loss of employment, the ratio ranges between 94% and 100% during the years 2004-2009[11]. Obviously, the ratio declines substantially as we go further back. This result implies that the procedure presented in section 4.1 and applied to the MCVL 2004-2010 obtains information regarding the process of registering unemployment benefit recipients that is roughly identical to that given in the Annual Report for the period of reference.

[10] It should be borne in mind that the Public Employment Service may conduct some type of additional processing (for example, processing administrative errors in variables such as gender, cause of termination, year of birth, etc.) which is not conducted in our case. Moreover, the increase in the number of entries into the RAI programme after 2008 may explain why the CCSF is capturing more spells than the Annual Report (see footnote 8).

[11] The exception is the year 2010. The same occurs when using the *2010 data file* (see below). As said in footnote 9, this is due to a change in the way the Annual Report provides the entries by reason. This fact makes comparisons extremely difficult.

As for the 2010 data file, it contains a smaller quantity of spells than the complete cumulative spells file, as indicated previously, and so the ratios calculated and displayed in the third column always offer values lower than those of the second column. Evidently, if the 2010 data file were used to obtain information on spells of benefit receipt beginning in years prior to and thus more distant from the representative period 2004-2010, these ratios would be even lower, as can be seen in the Table. This is an additional reason to that given in section 3 for not using the MCVL retrospectively, in other words, using the information on the past spells of those individuals who are representative of a given year.

4.4 Third application: counting the number of transitions and durations

This application illustrates the bias which can arise in any analysis of labour mobility or duration of covered unemployment and employment if the MCVL is not used correctly. Firstly, Table 6 provides information on the total number of transitions towards employment or unemployment with benefit that occurred in a given year (in this case, in 2005). Each column in the Table shows the situation to which each person has moved, which might be employment (distinguishing between self-employment, wage and salary employment with open-ended or temporary contract, agricultural employment or domestic work) or unemployment in receipt of benefits. Each row provides information on each transition in accordance with the procedures of the complete cumulative spells file and of the 2010 data file described in the previous applications. The third row highlights the differences in the number of transitions between both files. The analysis was conducted for the entire sample, for people aged 16-30 and for women.

Table 6

Distribution of the total number of transitions into employment or covered unemployment that occurred in 2005, by state of destination: entire sample, individuals aged 16-30 and women. Information obtained from the 2004-2010 complete cumulative spells file (CCSF0410) and the 2010 data file (F2010). Source: MCVL.

	State of destination			
	Unemployment benefits	*Self-employment*	*Open-ended contracts*	*Temporary contracts*
All				
(1) CCSF0410	3,610,800	535,975	2,938,375	16,274,875
(2) F2010	3,248,775	467,150	2,716,775	14,176,850
(2)-(1)	-362,025	-68,825	-221,600	-2,098,025
<30 years				
(1) CCSF0410	1,084,175	163,125	1,200,150	8,553,525
(2) F2010	962,525	144,625	1,097,750	7,418,375
(2)-(1)	-121,650	-18,500	-102,400	-1,135,150
Women				
(1) CCSF0410	1,845,950	202,600	1,396,825	7,847,475
(2) F2010	1,647,475	174,075	1,287,150	6,874,150
(2)-(1)	-198,475	-28,525	-109,675	-973,325

Table 6

Distribution of the total number of transitions into employment or covered unemployment that occurred in 2005, by state of destination: entire sample, individuals aged 16-30 and women. Information obtained from the 2004-2010 complete cumulative spells file (CCSF0410) and the 2010 data file (F2010). Source: MCVL.

(Continued)

	State of destination			
	Agricultural employment	*Domestic work*	*Unknown*	*Total*
All				
(1) CCSF0410	1,869,600	338,950	105,500	25,674,075
(2) F2010	1,661,000	246,225	89,575	22,606,350
(2)-(1)	-208,600	-92,725	-15,925	-3,067,725
<30 years				
(1) CCSF0410	659,300	125,075	43,350	11,828,700
(2) F2010	567,575	90,500	35,525	10,316,875
(2)-(1)	-91,725	-34,575	-7,825	-1,511,825
Women				
(1) CCSF0410	806,275	302,625	31,725	12,433,475
(2) F2010	722,850	219,875	27,825	10,953,400
(2)-(1)	-83,425	-82,750	-3,900	-1,480,075

Note: the column "Unknown" captures the information of unknown destination (administrative errors).

The number of transitions not recorded in the 2010 data file compared to the complete cumulative spells file is enormous: some 3.1 million spells. Looking at the destination of these transitions, it can be seen that the largest sample loss occurring in the 2010 data file corresponded to transitions towards temporary employment (some 2.1 million spells) and towards open-ended contracts (some 222,000), in addition to transitions towards unemployment in receipt of benefit (some 362,000 spells). In the case of the groups consisting of people aged 16-30, or women, the 2010 data file lost a considerable amount of information compared to the complete cumulative spells file, with approximately 1.5 million spells less in the total for each of these groups.

The fact that the 2010 data file does not collect such a large number of past spells can have consequences for the calculation of certain labour variables, such as the duration of the spells of employment or unemployment in receipt of benefits. For the sake of illustration, Table 7 provides the total number of days of employment and of unemployment in receipt of benefit of the spells that began in 2005. As can be seen, the 2010 data file collected 343 million days of employment (especially in the case of temporary contracts) and 55 million days of covered unemployment less than the CCSF. For the groups consisting of people aged 16-30 and women, the 2010 data file lost a considerable amount of information compared to the CCSF, with approximately 145 (16) million days less in employment (unemployment benefits) for individuals aged between 16 and 30 years old and 170 (32) million for women.

Table 7

Total number of days spent in employment (by kind) and covered unemployment of individuals starting spells in 2005. Information obtained from the 2004-2010 complete cumulative spells file (CCSF0410) and the 2010 data file (F2010). Source: MCVL

	All		
	CCSF0410 (1)	*F2010 (2)*	*(2)-(1)*
Unemployment benefits	493,811,203	438,875,390	-54,935,812
Employment	4,948,806,077	4,605,699,713	-343,106,364
Self-employment	546,926,577	505,848,239	-41,078,338
Open-ended contracts	2,432,495,839	2,353,247,956	-79,247,883
Temporary contracts	1,542,279,131	1,378,992,535	-163,286,597
Agricultural employment	247,805,945	227,090,425	-20,715,520
Domestic work	179,298,584	140,520,558	-38,778,026
Unknown situations	4,772,324	-510,900	4,772,324
Observations	25,674,075	22,606,350	-3,067,725

(Continued)

	<30 years		
	CCSF0410 (1)	*F2010 (2)*	*(2)-(1)*
Unemployment benefits	121,338,806	104,822,823	-16,515,983
Employment	1,973,653,768	1,828,881,660	-144,772,108
Self-employment	162,487,508	152,238,205	-10,249,303
Open-ended contracts	950,987,819	917,274,521	-33,713,298
Temporary contracts	711,648,205	633,451,354	-78,196,851
Agricultural employment	94,419,540	85,696,276	-8,723,263
Domestic work	54,110,697	40,221,304	-13,889,393
Unknown situations	-510,900	4,772,324	-510,900
Observations	11,828,700	10,316,875	-1,511,825

(Continued)

	Women		
	CCSF0410 (1)	*F2010 (2)*	*(2)-(1)*
Unemployment benefits	259,171,380	226,869,499	-32,301,881
Employment	2,202,423,317	2,032,148,491	-170,274,827
Self-employment	205,791,355	188,928,298	-16,863,058
Open-ended contracts	1,094,473,149	1,056,246,746	-38,226,404
Temporary contracts	637,996,472	566,061,610	-71,934,862
Agricultural employment	99,260,677	91,183,696	-8,076,980
Domestic work	164,901,664	129,728,141	-35,173,523
Unknown situations	4,772,324	-510,900	4,772,324
Observations	12,433,475	10,316,875	-1,480,075

Note: See note to Table 6

Therefore, an analysis of duration of employment (or unemployment in receipt of benefit) for spells that began in one year (such as 2005) using the information from another, subsequent year (such as the 2010 data file) would lead to an underestimation of the total time spent in those states on the part of the population. This point is relevant since, for instance, any study concerning the public expenditure on unemployment compensation using that information on unemployment benefit duration will underestimate the real expenditure.

The fundamental reason for this bias in durations is that most of the spells not reflected in the 2010 data file are short term, indicating the existence of a group of unstable workers who enter and leave employment very quickly, and whose spells are, in contrast, reflected in the complete cumulative spells file. Meanwhile, information on another group of workers who are more stable, both in terms of employment and of unemployment in receipt of benefit, is obtained by both procedures.

5. Conclusions

The birth of a new database in the world of statistics is usually seen as an opportunity to expand the field of research, enabling the study of areas which were not previously covered by statistical sources, or the analysis of already known aspects from a different or more complete perspective. The same is the case of the Spanish MCVL, which has elicited great expectations among economists (above all, among labour economists), given its enormous sample size and the availability of information on the work histories of the individuals included in the sample. However, as Hamermesh (2000) has indicated, the fact that the data are available does not imply that they will answer the research question analysed.

As we have tried to argue in this article, despite the potential the MCVL offers for studying longitudinal aspects of the labour market, a clear limitation exists. The sample can be used to analyse issues affecting individuals from the moment that the information has been extracted (2004 onwards), but not the ones that affected them in the past, since the samples are not representative of past cohorts. This implies that, provided that they are adequately organised, the data can be used to study the work history of individuals or to conduct microeconomic assessments of public policies during the years to which the samples refer (from 2004 onwards), but they cannot be used to study labour transitions or the impact of policies in years prior to the beginning of the samples (2004), at least not without the risk of biased results.

Furthermore, in order to improve the (correct) use of information from the MCVL, we have described a process through which the original data can be organised in such a way as to permit the accurate study of work histories. This process collects information both on individuals who are always part of the sample because their connection with the Social Security system is stable, and on those whose connection presents interruptions and who thus do not reappear in the representative observation period. A comparison of the information obtained using this procedure (complete cumulative spells file 2004-2009) and that obtained using only one sample (2010 data file) in three empirical

applications demonstrates the sample loss and bias which occurs when data on spells and individuals are not processed correctly. These applications refer to the analysis of patterns on the presence of individuals in the sample, to total number of spells of employment and receipt of unemployment benefit beginning in a specific period, and to the number of transitions and the days spent in employment and unemployment in receipt of benefits. In particular, we find that the use of only one sample retrospectively would lead to an underestimation of the total time spent in employment and covered unemployment and of the total number of transitions.

In conclusion, retrospective use of the MCVL, that is to say, using the information on the past spells of individuals who are representative of a given year (for instance, 2010) generates biases, since the information on those people who belong to the sample for that year only represents a portion of the total past information, due to the existence of individuals who, although being part of the past population, no longer are present in the year in question, either because they have died, are no longer connected with the Social Security system, or for other reasons. It also provides a lower number of spells than the complete cumulative spells file or the official statistics. These reasons invalidate the retrospective use of the MCVL both for the closest "representative" period (2004-2010) and, above all, for a more distant, prior period in time, even where computerised records go back many years.

References

ARGIMÓN, I. AND GONZÁLEZ, C.I. (2006). «La Muestra Continua de Vidas Laborales de la Seguridad Social», *Boletín Económico del Banco de España*, May, 40-53.

ARRANZ, J.M. AND GARCÍA-SERRANO, C. (2011). «Are the MCVL tax data useful? Ideas for mining», *Hacienda Pública Española*, 199(4), 151-186.

BLOSSFELD, H.P., GOLSCH, K. AND ROHWER, G. (2007). *Event history analysis with Stata. Lawrence Erlbaum Associates, Taylor & Francis Group.*

CEBRIÁN, I. AND TOHARIA, L. (2008). «La entrada en el mercado de trabajo. Un análisis basado en la MCVL», *Revista de Economía Aplicada*, 16 (E-1), 137-172.

CEBRIÁN, I., MORENO, G. AND TOHARIA, L. (2009). «¿Por qué no reducen las bonificaciones la temporalidad?», *Paper presented at the VIII Jornadas de Economía Laboral, Zaragoza.*

CLEMENTE, J., GARCÍA-MAINAR, I. AND SANSO, M. (2008). «Análisis de las diferencias salariales entre trabajadores indefinidos», *Revista de Economía Aplicada*, 16 (E-1), 93-136.

CONDE-RUIZ, J.I., FELGUEROSO, F. AND GARCÍA-PÉREZ, J.I. (2010). «Las reformas laborales en España: un modelo agotado», Papeles de Economía Española, 124, 128-147.

DURÁN, A. AND SEVILLA, M.A. (2006). «Una Muestra Continua de Vidas Laborales, en C. Marcos (dir.), El papel de los registros administrativos en el análisis social y económico y el desarrollo del sistema estadístico», Madrid, *Instituto de Estudios Fiscales*, 241-252.

DURÁN, A. (2007). «La Muestra Continua de Vidas Laborales de la Seguridad Social», *Revista del Ministerio de Trabajo y Asuntos Sociales*, 1, 231-240.

GARCÍA-PÉREZ, J.I. (2008). «La Muestra Continua de Vidas Laborales (MCVL): una guía de uso para el análisis de transiciones», *Revista de Economía Aplicada*, 16 (E-1), 5-28.

GARCÍA-PÉREZ, J.I. AND REBOLLO, Y. (2009a). «The use of permanent contracts across Spanish regions: do regional wage subsidies work?», *Investigaciones Económicas*, 33(1), 97-130.

GARCÍA-PÉREZ, J.I. AND REBOLLO, Y. (2009b). «Do wage subsidies the subsequent employment stability of permanent workers: the case of Spain», *Moneda y Crédito*, 28, 65-102.

GARCÍA-SEGOVIA, F. AND DURÁN, A. (2008). «Nuevos avances en la información laboral: la Muestra Continua de Vidas Laborales», *Economistas*, 116, 228-231.

HAMERMESH, D. (2000). «The craft of Labormetrics», *Industrial and Labor Relations Review*, 53(3), 363-380.

IZQUIERDO, M., LACUESTA, A. AND VEGAS, R. (2009). «Assimilation of immigrants in Spain: A longitudinal analysis», *Labour Economics*, 16(6), 669-678.

LAPUERTA, I. (2010), «Claves para el trabajo con la Muestra Continua de Vidas Laborales», *DemoSoc Working Paper*, nº 2010-37, *Universitat Pompeu Fabra*.

REBOLLO, Y. (2011). «Landing a permanent contract in Spain: Do job interruptions and employer diversification matter?», *The Manchester School*, 79(6), 1197-1236.

REBOLLO, Y. (2012). «Unemployment insurance and job turnover in Spain», *Labour Economics*, 19(3), 403-426.

TOHARIA, L., ARRANZ, J.M., GARCÍA-SERRANO, C. AND HERNANZ, V. (2009). «El sistema español de protección por desempleo: eficiencia, equidad y perspectivas», *Dirección General de Ordenación de la Seguridad Social, Ministerio de Trabajo e Inmigración, Madrid.*

TOHARIA, L., ARRANZ, J.M., GARCÍA-SERRANO, C. AND HERNANZ, V. (2010). «El sistema español de protección por desempleo y la salida del paro», *Papeles de Economía Española*, 124, 230-246.