



Unemployment in administrative data using survey data as a benchmark

Cristina Lafuente¹

Received: 2 May 2018 / Accepted: 8 July 2019 / Published online: 18 July 2019
© The Author(s) 2019

Abstract

Social security administrative data are increasingly becoming available in many countries. These data have a long panel structure (large N, large T) and allow for the measurement of many different variables with high accuracy. It also captures short-term unemployment spells which are normally unavailable in survey data due to its design. However, the measurement of unemployment differs in both types of datasets. The resulting gap between total unemployment and registered unemployment is not constant across workers characteristics or time. In this paper, I present a simple, systematic method to expand the raw Spanish Social Security administrative data. I identify unemployed workers who are not receiving unemployment benefits, using information from the institutional framework and using the Labour Force Survey as a benchmark. The resulting unemployment rates and labour market flows are comparable across both datasets. Administrative data can also overcome some of the problems of the Labour Force Survey, such as changes in the structure of the survey. This paper aims to provide a comprehensive guide on how to adapt administrative datasets to make them useful for studying unemployment.

Keywords Administrative data · Survey data · Unemployment · Temporary contracts

JEL Classification J21 · J60 · J80

I would like to thank the National Statistics Institute (INE) and the Ministry of Employment and Social Security of Spain for kindly providing the data. I would like to thank my supervisors Maia Güell and Ludo Visschers for all of their support and advice; I have also benefited from the comments and suggestions of Raquel Carrasco, Carlos Carrillo-Tudela, José Ignacio García-Pérez and Rafael Lopes de Melo. Any remaining errors are my own.

✉ Cristina Lafuente
Cristina.Lafuente@ed.ac.uk

¹ University of Edinburgh, Edinburgh, UK

1 Introduction

Administrative datasets are gaining popularity among economists.¹ They offer some advantages over traditional Labour Force Surveys. Most administrative datasets can identify firm-worker pairs and have detailed and extensive working histories (large N, large T). However, using these datasets for the study of unemployment presents some challenges. Firstly, these data were not designed for research, but rather for administrative bookkeeping: calculating contributions to the social security and benefit entitlement of workers. Secondly, the definition of unemployment used in administrative datasets is not the same as the International Labour Office (ILO) standard. Finally, in some countries the administration only keeps track of the unemployed while they are receiving benefits.

These discrepancies are particularly relevant for the case of Spain. Since 2004, the Social Security and the Ministry of Labour of Spain has made one such administrative dataset available to researchers: the *Muestra Continua de Vidas Laborales* (MCVL). This dataset provides complete employment histories of 4% of the Spanish Labour Force in a given year. It can be linked to anonymised tax records, providing comprehensive information on wages and benefits. The MCVL adopts the administrative definition of unemployment: only workers receiving unemployment benefits are considered to be unemployed. This measure systematically excludes individuals who have not accumulated enough contribution periods to be eligible for unemployment insurance. These are mostly young workers in short temporary contracts. Moreover, after the 2008 financial crisis, the number of workers whose unemployment benefits had expired increased considerably. As a result, the unemployment rate as measured by the Labour Force Survey (*Encuesta de Poblacion Activa*) diverged substantially from the MCVL. By the end of 2013, there was a gap of 10 percentage points between the two measures of unemployment.

This paper aims to reconcile this gap by expanding the administrative definition of unemployment in two ways. Firstly, unemployed workers whose benefits run out will appear in the administrative data as if their spell was over. This has been noted by previous authors (García Pérez 2008). The first expansion, which I call the long-term unemployment (LTU) expansion, adds the missing days between the end of a registered unemployment spell and the next employment spell to correct for these artificially short spells. Secondly, workers that are not entitled to any unemployment benefits will not appear as unemployed. Using the institutional framework, three possible situations emerge: quits, workers with too short tenures to qualify for benefits and self-employed workers out of a job. I identify spells corresponding to these cases and add them to the MCVL. I refer to this expansion as the short-term unemployment (STU) expansion. Including these spells is crucial for young people and women, who have a higher incidence of part-time and temporary contracts. This approach refines the most common approach in the literature, which considers all non-employment spells

¹ See for example Moffitt (1985), Katz and Meyer (1990), Sullivan and Von Wachter (2009), Tattara and Valentini (2010), Couch et al. (2011), Krueger and Mueller (2011), Bonhomme and Hospido (2017) among others.

as unemployment.² Together, both expansions can explain most of the gap between the two data sources, supporting the use of the MCVL for the study of unemployment.

The paper then shows how the expanded MCVL can complement the Labour Force Survey for the study of labour market flows. The MCVL can be used to address two of the main challenges faced when trying to quantify labour market transitions: non-responsiveness of the unemployed (attrition) and changes in the survey design. In the first case, up to one in five unemployed individuals fail to respond to two consecutive interviews.³ This overstate flows from unemployment to employment—which are mostly to temporary work. The MCVL does not suffer from attrition problems, as it tracks all of the changes in the status of individuals (except flows into non-participation). The 2005 change in survey design creates breaks in the transition rates of temporary and permanent workers in the LFS. The MCVL does not fundamentally change in this period, so using the 2005 wave it is possible to examine these changes. Finally, the MCVL allows for the observation of high-frequency transition rates. Due to its quarterly frequency, the Labour Force Survey is not adequate to capture very short, frequent employment–unemployment spells. The MCVL can identify these spells with precision, making it a valuable tool for the study of frictional and youth unemployment. This last observation has broader implications beyond Spain, as young workers across Europe are becoming more exposed to unstable employment. From mini-jobs in Germany and zero hour contracts in the UK to the gig economy worldwide.

The paper is structured as follows: Sect. 2 describes the two datasets, their advantages and disadvantages. Section 3 presents the unemployment gap between the LFS and the MCVL and discusses its likely sources. Section 4 expands the MCVL definition of unemployment, comparing the resulting unemployment figures with the Labour Force Survey estimates. Section 5 provides further robustness checks. Section 6 demonstrates the uses of comparable data sources for the study of labour market flows. Section 7 concludes.

2 Data

This section explains the main characteristics of the two data sources I employ throughout the paper: the Spanish Labour Force Survey (LFS hereafter), elaborated by the National Statistics Institute (known as INE), and the *Muestra Continua de Vidas Laborales* (MCVL), provided by the Spanish Social Security Agency. It offers a comparison between the two and outlines how to structure the latter as a quarterly panel.

Official unemployment statistics come from the Spanish LFS, which follows a representative sample comprising over 100,000 people for six consecutive quarters. The INE provides population weights which enable population level estimates to be constructed from the sample. These weights will be used in this paper when reporting stocks, as they also correct some sampling errors. The LFS classifies a worker's

² See for example Rebollo-Sanz (2012), Rebollo-Sanz and García-Pérez (2015), Nagore Garcia and van Soest (2017), Bentolila et al. (2017) and Rebollo-Sanz and Rodríguez-Planas (2018).

³ This is at its worst in the year 2005, but it is over 15% most of the quarters in the 2005–2013 period, as Fig. 19 in Sect. 6.1 shows.

employment status by asking them to report their activities in the week of the interview. In particular, if they were employed or if they searched for a job.⁴ Based on their answers, the LFS can identify when a worker is out of the labour force and why which is its main advantage over administrative data.⁵ The LFS can also account for informal work arrangements, to the extent that workers are honest about their answers.⁶ The unit of observation is the individual response at the quarter of the interview. Linking the different quarters by the individual's identifier gives the survey a rotating panel structure.

However, many participants do not reply in all of the six quarters in which they are supposed to be part of the sample. This leads to problems when calculating stocks and building labour market transitions. The population weights help to correct for non-responsiveness when building labour market stocks, but they do not correct flows. This is because the weights indicate how many other individuals in the population are represented by the interviewee. If one individual fails to respond to the survey, the sample weights can be readjusted to reflect this. However, two consecutive observations are required in order to record labour market transitions. If one respondent misses an interview in one quarter, the weights cannot be used to recalculate the importance of the remaining individuals that have two consecutive observations as they do not account for people conditional on their status in a past survey. As a result, the stocks (counting observations in each labour state) are correct using the weights, but the flows are not. This issue is further discussed in Sect. 6.1. Note that this is a problem that affects all Labour Force Surveys and is not specific to Spain.

Another complication arises from changes in the survey design. Two notable changes, in 2001 and 2005, affected unemployment measurement and produced breaks in the series.⁷ These changes did not affect the stocks of unemployed workers, although they did alter labour market flows. I will revisit this issue in Sect. 5.

The MCVL is comprised of the entire working histories of a 4% sample of the working population extracted each year from 2005 onwards. Similar datasets exist for

⁴ In particular, I use the variables *Type of Contract* to identify employees, *Current working situation* for the self-employed and the variable *AOI* for unemployed and out of the labour force individuals. This last variable combines the answers to other key variables (“Were you working this last week?”, “Are you looking for a job?”/“Are you ready to work in the next 15 days?” and “What type of contract do you hold?”). This variable is the one used for the official unemployment rate series by the government, the EUROSTAT and OECD.

⁵ If the respondent is not employed nor looking for a job it asks her to declare the reason by choosing one of 9 possible answers. These include *studying*, *thinking she is not going to be able to find a job* and *caring for others*, among others.

⁶ A simple test consists of looking at the proportion of individuals that declared to be receiving unemployment benefits but also report themselves as not searching for a job. Being an active searcher is a requirement to receive unemployment benefits. By declaring themselves in this situation, the individuals admit their irregular position. The proportion of claimants in this situation reaches 30% of the total in the 2005–2007 period. This proportion drops to nearly zero after 2008. Therefore, many individuals seem to be honest even when they are revealing some irregularity in their situation. Other than this test, which is more likely to be an underestimate, it is hard to quantify how many workers are in irregular employment.

⁷ In 2001 the requirement for unemployed workers to be available for work in the next two weeks was introduced. This change caused a shift in the stock of unemployed in 2001. In 2005, two meaningful changes were introduced: the sample composition was altered to reflect changes in the population and surveys shifted to an electronic format instead of phone calls.

Portugal, Germany, Italy, Austria and the US, among other countries.⁸ The sample size is large, with over 1 million individuals in each year. In contrast to other administrative datasets, the MCVL contains information on self-employed individuals as well as employees and unemployed workers. There are anonymised identifiers for individual workers and firms. This feature is highly valuable for labour economists, as it allows for the observation of job-to-job transitions and the identification of recalls to the same employer. The LFS lacks information on the employer and therefore cannot provide this information. There, the unit of analysis is the employment or unemployment spell. This feature is useful for the estimation of duration models, although to study labour market flows one has to transform the individual-spell structure into an individual-quarter labour status.

There are two sources of wage information in the MCVL. Firstly, using the worker identifier and an establishment code the working histories file can be linked to the social security contributions file. These contributions specify the gross salary upon which firms have to make contributions to social security. As with similar datasets, these wages are top-coded. Secondly, using the individual worker and firm identifiers the working histories panel can be linked with the “Income Tax complement”. This file contains income tax information on wages, other forms of payment, unemployment or disability benefits, severance payments and any other flow of income between the firm (or the administration) and the worker. The tax file contains declared profits from economic activities. Although that information is highly susceptible to misreporting for tax avoidance purposes, it nevertheless provides some insight into self-employed earnings. Researchers working with these data have used both sources. Bonhomme and Hospido (2017) show how to adapt the contribution file to study wage and earnings dynamics.

Despite all of its useful information, it is important to note that the MCVL is not a matched employer–employee data. The unit of measurement is the worker, not the firm. Therefore, observing two or more workers at the same firm is unlikely. In other words, the firms in the MCVL are not representative of firms in Spain.⁹

The main disadvantage of the MCVL is that in its original format the data are not useful for research. Organising and cleaning the data requires a considerable amount of time and knowledge of Spanish legal terms. This challenge arises because the MCVL is an extraction of administrative records. The LFS, in contrast, is built with researchers in mind. Over the years, different academic articles have been written explaining how to clean and format the data [see García Pérez (2008), Lapuerta (2010), Arranz et al. (2011), López-Roldán (2011) or the online appendix in Roca and Puga (2017) for example]. In particular, García Pérez (2008) provides the most comprehensive guide on the treatment of unemployment in the MCVL. It explains how to deal with overlapping employment spells and censored unemployment spells. However, after cleaning and formatting the data, there is still the question of how to treat unemployed workers who are not registered as unemployed. These periods appear as gaps between observed spells. This is a common feature with other administrative datasets, but in

⁸ See Tattara and Valentini (2010) for a complete summary.

⁹ However, because it is representative of self-employed individuals, the coverage of small firms is better than large firms. This representation issues may be useful, as most firms in Spain are very small.

Spain, this issue is especially relevant because of the prevalence of very short and very long unemployment. These issues and how to deal with them are at the core of this paper.

In principle, it would be possible to build a panel earlier than 2005, using the information on the past working histories of workers. Some spells date back as far back as the 1960s. However, both García Pérez (2008) and Arranz et al. (2011) warn against doing this kind of inference as the sample is representative of the year of extraction. The 2005 file is representative of the working population of Spain in 2005. Every subsequent year new spells are added to adapt to demographic changes. The MCVL does not drop any worker. The cases of workers not appearing in a given year are either migration, transitioning out of the labour force or deaths.¹⁰ The MCVL includes pensioners in its sample, so retirement is not a cause for dropping out. Using the 2005 file to study the labour market in the 2 or 3 years prior would cause minor representativeness problems. However, using the MCVL to look further back would over-represent younger workers. For this reason, it is best to use all of the individual year files. Arranz et al. (2011) provide an algorithm to merge these files while consolidating all individual observations. However, for some applications, using only the latest year can offer some advantages. The later waves have fewer discontinuities and overlaps, more variables and greater accuracy. In particular, it is easier to calculate spell duration. In this paper, I will use each year file from 2005 to 2013, consolidating all of the information into a single unbalanced panel.

Table 1 summarises the advantages and strengths of the dataset as described above.

3 The unemployment gap

In order to compare unemployment in both datasets, it is necessary to format the MCVL into a quarterly panel format. The formatting and panelling procedure is detailed in “Appendix”. The main strategy to transform the MCVL into a panel uses the first two weeks of every quarter as a reference period. It then classifies workers in their current employment status. If the worker has more than one status in these two weeks, the last observed status is chosen. Sections 5.2 and 5.3 provide robustness checks on these assumptions.

The main challenge when classifying worker status in the MCVL is that unemployment and non-participation cannot be clearly distinguished. In this paper, the definition of unemployment used is the same as in the standard ILO definition (which coincides with the LFS). It follows that the differences in between the MCVL and the LFS unemployment series will represent unemployed workers that are missing in the administrative data or the LFS. For example, frictional unemployment is unemployment by the ILO definition but is often not captured by the LFS. The timing and structure of the survey mean that this kind of unemployment is unlikely to be recorded. The marginally attached (those who are not employed nor actively searching) would not be unemployed by that definition. These cases should not be included the MCVL. However, we cannot exclude these cases from the MCVL without some imputation.

¹⁰ The MCVL records the date of passing.

Table 1 Data comparison table

	Labour Force Survey	Administrative data (MCVL)
Description	Rotating panel of quarterly interviews with a sample size of over 100,000. It is available from 1987	A sample of about 1,000,000 job records of people who have any sort of affiliation with the Social Security. It constitutes a panel since it follows most of the same people from 2004
Advantages	<p>Detailed and accurate information for personal characteristics (such as education)^a</p> <p>It has the potential to track labour status changes that are made out of the scope of administrative records (first job seekers, informal market jobs, inactive workers)</p>	<p>Firm and worker identifiers allow for the study of job-to-job transitions (not available in the LFS)</p> <p>Very accurate information on employment spells, with precise dates of entry and exit into and out of jobs/unemployment</p> <p>Can be matched to a fiscal dataset for wage/benefit information</p> <p>Consistent through time</p>
Disadvantages	<p>Fails to capture short-term jobs and some very short unemployment spells due to its <i>quarterly</i> frequency</p> <p>There are important series breaks in 1992, 1999, 2001 and 2005</p>	<p>It can't track anyone who has no formal relationship with the Social Security Agency. As such, it does not track people outside of the labour force well</p> <p>For the same reason, it is also unable to track informal labour market activities</p>

^aAlthough there have been some improvements on the measure of this variable in the MCVL as noted by Roca and Puga (2017), this variable is still unreliable, according to the 2018 MCVL user guide manual

That imputation would require constructing a “propensity to be marginally attached” measure with observables in both the LFS and the MCVL. This approach is not followed for three reasons. First, the set of personal characteristics variables common to both datasets is small.¹¹ Building a propensity score in the LFS and applying it in the MCVL would introduce noise which will be difficult to measure. The LFS measure of attachment is already a noisy estimate.¹² Although I choose to follow the ILO definition, most labour economists consider the marginally attached as being unemployed. For all of these reasons, I find that trimming the marginally attached from the administrative data is not worthwhile for the purpose of this paper.

As a final note, the LFS and the MCVL have a different number of observations: an average of 108,136 in the LFS¹³ and 678,183 in the MCVL. In order compare the

¹¹ Some key variables, like education, occupation and sector are defined differently in both datasets or they are highly aggregated.

¹² This is true even after cleaning up the frequent UNU transitions (unemployment to non-participant to unemployment) the LFS as in Elsbey et al. (2015).

¹³ The weighted Labour Force Survey has a mean of 31,360,266 observations per period—which amounts to the total population of working age in Spain.

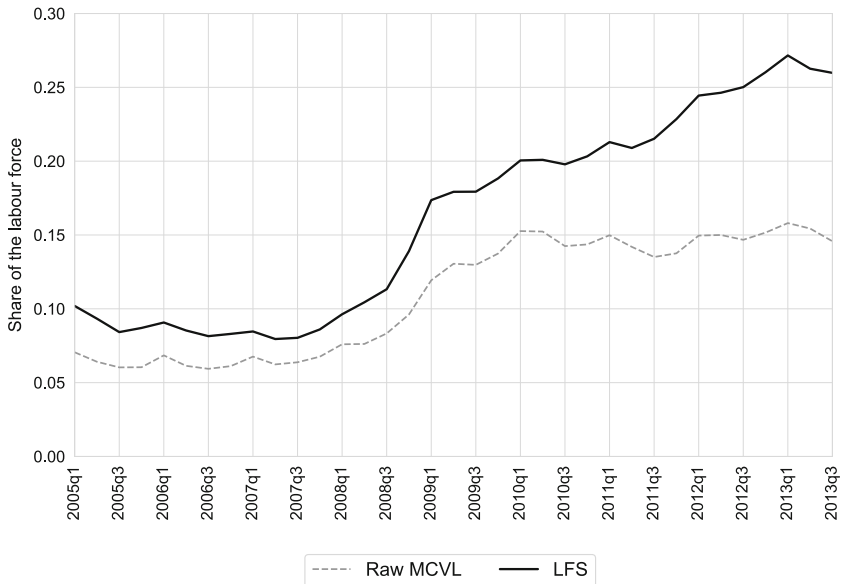


Fig. 1 Unemployment rate in Spain

two datasets, I express the labour market stocks as shares of the labour force for the remaining of this paper.

After building the MCVL into a panel, we can compare both datasets. Figure 1 shows the unemployment rates from the raw MCVL data and the LFS. There is a growing disparity between the two, that reaches ten percentage points by the second quarter of 2013. These differences persist across age groups and gender: Fig. 2 shows the gap was wider for women before 2008 and Fig. 3 shows that it is very large for young workers. By the end of the sample, their unemployment rate in the MCVL is half that the LFS. Moreover, in the MCVL the unemployment rate of young workers appears to trend down from 2010, while in the LFS it is rapidly increasing.

The main source of this discrepancy is the different definitions of unemployment:

- The LFS considers a person unemployed if: (1) they are not currently employed, (2) they are actively looking for a job and (3) they are ready to start working within the next 15 days.
- The MCVL considers a person unemployed if: (1) she has been in the social security system before (had a previous job) and (2) is receiving unemployment benefits.

In other words, the MCVL excludes all unemployed people whose benefits have expired. The Spanish Social Security Agency does not record any other benefit for those who exhaust their unemployment compensation. All of those unemployed for more than 2 years (4 years for some groups) are missing from the registry.¹⁴ As long-term unemployment reached 60% of total unemployment by the end of 2013, this

¹⁴ Some workers are entitled to an absolute maximum of 4 years if they have family obligations or other special circumstances.

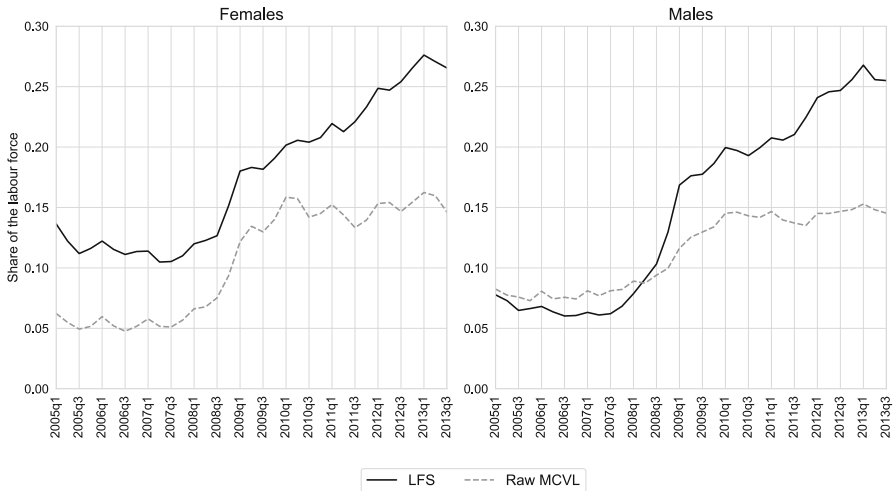


Fig. 2 Unemployment rate by sex

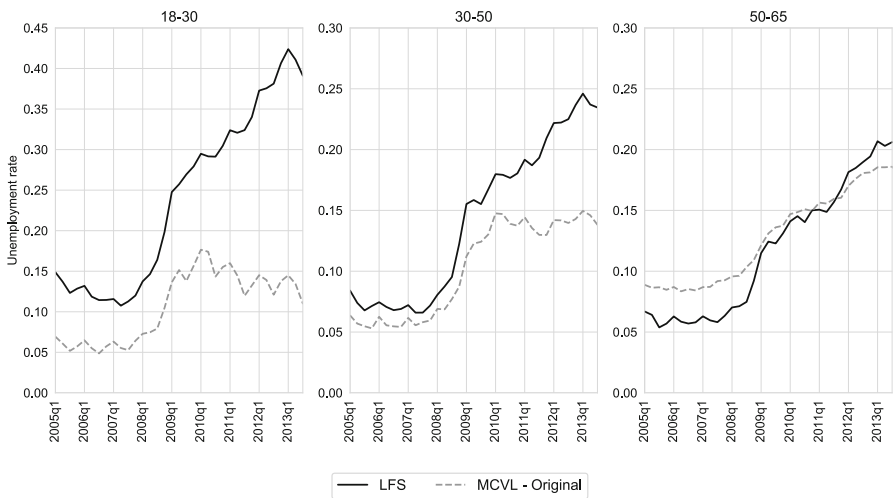


Fig. 3 Unemployment rate by age. *Source:* MCVL and the LFS

is the principal potential source of disagreement. The first expansion deals with this issue by extending observed spells until the start of the next job or until the end of the sample.

The Social Security agency also excludes all individuals without the right to claim unemployment compensation. That is the case for those who have less than a year of employment in the last 4 years.¹⁵ The second expansion aims to recover these spells by

¹⁵ Before 1992, the threshold was six months of contributions. The labour reform of that year increases the minimum to twelve. No other major reform, including the most recent 2012 one, has changed this threshold.

adding gaps between employment spells of those without the right to claim benefits, under certain conditions.

Finally, it is worth noting that there is another potential source of the disparity is that the MCVL may be counting some inactive workers (by the definition of the LFS) as unemployed. That is individuals who are not actively searching for a job or are not ready to work in the next 15 days. Notice however that if this was the primary source of disparity, then the MCVL unemployment rate would be above the LFS. That is not what we observe in the data, except for older workers.¹⁶

4 Unemployment expansions

The disparity between both datasets suggests that there may be some unemployment missing from the MCVL.

This section shows how to implement two simple unemployment expansions to capture the missing spells. The long-term unemployment (LTU) expansion includes unemployment spells beyond the expiration of unemployment benefits. This expansion is routinely applied in the empirical literature (see García Pérez 2008; Rebollo-Sanz 2012; Bentolila et al. 2017; Fernández-Navia 2019 for example). The short-term unemployment (STU) expansion aims to capture unregistered unemployment spells which do not count as unemployment by the administrative definition. These are mostly very short, frictional unemployment spells. It offers an alternative to counting all gaps between spells as “non-employment”, which is the other approach followed in the literature (see for example Rebollo-Sanz 2012; Rebollo-Sanz and García-Pérez 2015; Nagore García and van Soest 2017; Bentolila et al. 2017; Rebollo-Sanz and Rodríguez-Planas 2018).

This section contrasts the resulting unemployment rates after each expansion against the LFS, which provides some insights into the different treatments of unemployment.

4.1 LTU expansion

Given the importance of long-term unemployment in Spain, it seems natural to extend those spells that become right-censored due to benefit expiry until the start of the next job. This expansion is already noted by García Pérez (2008) as a necessary treatment to work with unemployment in the MCVL. However, many of the long-term unemployed have not found a job by the end of the sample. Figure 4 shows that merely adding the days before another job spell does not help to reconcile the post-2009 trend in both datasets. Comparing this series with the raw MCVL series, it makes little difference.

The *LTU expansion* adds all of the unfinished unemployment spells by the end of 2013, as well as extending the duration of registered unemployment spells between jobs as before. Unfinished spells were very prevalent at the end of 2013. These spells could be a lesser issue for researchers using later years as the end of their sample. My approach is to extend all unfinished spells after benefit expiration. This assumption

¹⁶ I consider workers who transition from unemployment to retirement as out of the labour force. They are unlikely to engage in active search, thus not counting as unemployed for the LFS.

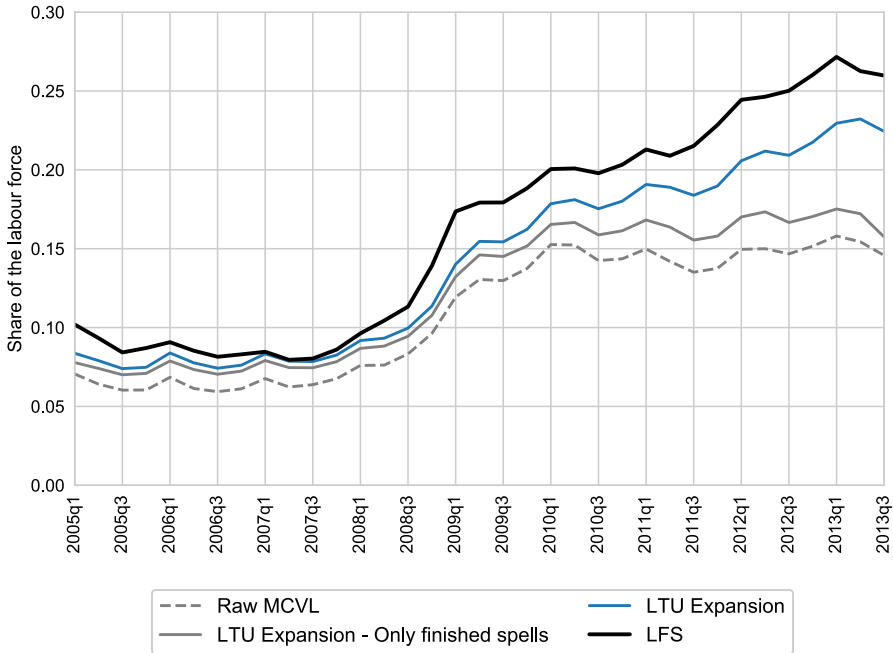


Fig. 4 Unemployment rates—LFS and expanded social security. *Source:* MCVL and the LFS. The *LTU expansion—only finished spells* series is calculated by applying the LTU expansion but excluding right-censored spells as the end of 2013

fits the trend of the LFS better than other specifications. Section 5.1 provides a robustness check on this assumption, showing that imposing restrictions on the duration of unfinished spells does not seem to affect the main results.

After this expansion, both trend and level are closer to the LFS, as shown in Fig. 4. However, the expanded series is still below the LFS by 3.7–2.5 percentage points by the end of 2013.

4.2 STU expansion

In order to close the gap between unemployment rates, we need to add the unemployed individuals without the right to claim. That is the case of:

- Quits into unemployment. Voluntary terminations of employment do not entitle workers to unemployment compensation.
- New entrants to the labour market (with no previous employment record).
- Workers with employment spells below the minimum requirement—less than a year of employment in the last 4 years. Young and temporary workers are particularly susceptible to this, lacking the right to claim.
- Self-employed workers are not entitled to unemployment insurance.¹⁷

¹⁷ This rule changed in the 2012 reform, but the change was not effective until 2013. For most of the sample, self-employed workers do not have the right to claim.

These spells tend to be shorter than the rest, as they represent frictional unemployment: short unemployment spells between jobs. This is therefore called the *short-term unemployment* (STU) expansion. To identify these spells, I chose to include all gaps between employment spells that lasted at least 15 days¹⁸ where at least one of the following conditions was also fulfilled:

- The worker quit her last job.
- The worker was self-employed in her last spell.
- The worker had not contributed enough to be eligible.^{19,20}

In addition to these restrictions, I added the requirement that the worker is not to be recalled to work on the same firm. Fujita and Moscarini (2017) noted that the dynamics of unemployment for those workers whose unemployment spell ends in a recall are very different from the rest. María Arranz and García-Serrano (2014) documented this for the case of Spain. Recalled workers may have little incentives to search and may answer “no” to the question “did you look for employment this week?” in the LFS. This case is particularly relevant for Spain because employers have incentives to use this tactic to extend the maximum duration of temporary contracts. Instead of renewing the contract beyond the 2-year limit, the firm can ask the worker to take a short leave and then rehire her.²¹

Finally, it is likely that some unemployment spells are censored at the end of 2013, as was the case with the LTU expansion. However, if we observe the worker over a long period of registered unemployment, we have some assurance that the worker was indeed looking for a job in this period. Restricting STU spells to those which end in a job strongly suggests that the worker was looking (and eventually obtained) employment. However, with right-censored gaps, it is harder to distinguish a transition into non-participation from a true unemployment spell. For this reason, I have restricted unfinished non-employment gaps that qualify for a STU expansion to be of at most 3 years in duration.²² Section 5.1 checks for alternative specifications, showing that there is not much difference between 3, 4 or 5 years, but there are significant changes

¹⁸ This restriction aims to exclude likely job-to-job transitions and small clerical errors. Other authors (see Rebollo-Sanz and García-Pérez 2015; Bentolila et al. 2017 for example) also take this approach.

¹⁹ The threshold, as stated before, is less than 360 days of employment, according to Spanish legislation. This threshold does not change for the period of study.

²⁰ The law in Spain does not allow to claim benefits for the days leftover in the last unemployment spell. For example, consider a worker has three months left of benefits and finds a six-month job. After that job ends, she cannot claim the three missing months from her last unemployment spell. However, she can choose not to claim benefits when she first becomes unemployed. In that case, after her six-month contract, she can claim her previous unclaimed benefits plus two more months for her six-month job. In this way, workers can save unemployment benefits for later. This mechanism may explain why the number of claimants rose significantly during the recession.

²¹ It is hard to quantify the extent of this practice, since it is illegal. However, using data from the LFS, in the 1999 to the 2013 period an average of 24.9% of temporary workers report to have been working for their current employer for more than 2 years—9.6% report 5 years or more. This is only possible if the employee was laid off and rehired by the firm at least once.

²² In comparison, Rebollo-Sanz (2012) chooses to censored all spells being at most 2 years. For her period of study (the 2005–2008 period), it makes sense to limit the length of unemployment spells given the low long-term unemployment rate. For studies that also looks at the recession period, the 3 years limit is preferred—see for example Rebollo-Sanz and García-Pérez (2015) and Bentolila et al. (2017).

Table 2 STU expansion spells, by type. *Source:* MCVL

	Quit	Self-employment	No right to UI
Total	173,186	100,907	622,164
Percentage	19.32	11.26	69.42

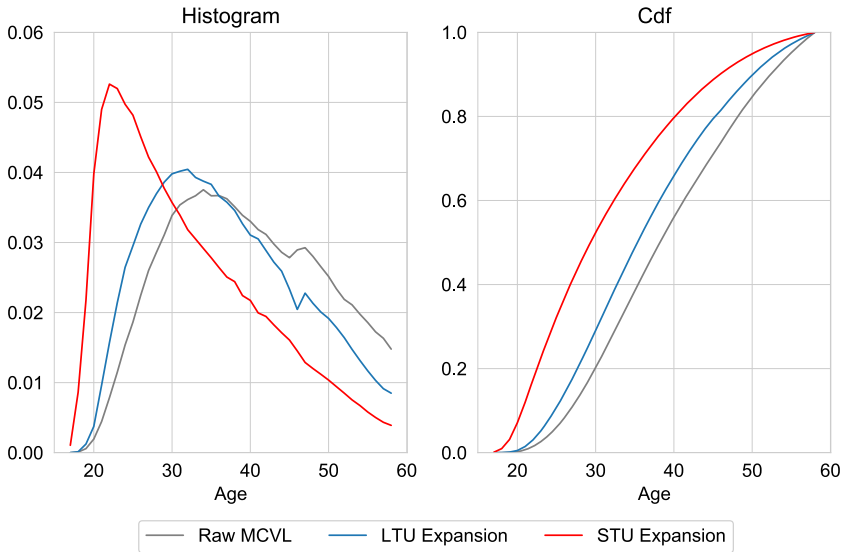


Fig. 5 Age distribution by expansion. All spells in the 2005–2013 period, by expansion. *Source:* MCVL

in trend unemployment if we use all censored unemployment spells or if we use only those lasting for 2 years or less.

These conditions ensure that inactive individuals are not counted in the STU expansion. Table 2 shows the breakdown of spells by which of the three conditions above are met. The majority of spells belong to individuals who did not have the right to claim unemployment benefits. Self-employment and quits also account for a significant fraction of these spells. Figure 5 shows the distribution for the ages of the unemployed at the start of their unemployment spell. Unemployed individuals from the STU expansion are overwhelmingly younger, with 80% of them under 40.

Figure 6 shows that STU spells are more likely to originate from temporary contracts than in the LTU expansion.²³ Future spells are also more likely to be another temporary job. If we exclude self-employment, 86% of all previous spells in the STU addition are temporary jobs, while the LTU and the raw data only have 70% and 74%, respectively. The vast majority of unemployed workers coming from self-employment are only counted in the STU expansion. The spells added by the STU expansion are also shorter than those of the LTU expansion, as Fig. 7 shows. I do not restrict any of the very short registered spells in the raw data. Tables 4 and 5 show the detailed

²³ This plot excludes spells beginning or ending in self-employment. These are more frequent in the STU. As explained before, self-employed workers had no right to claim unemployment benefits, so only the STU expansion manages to capture them.

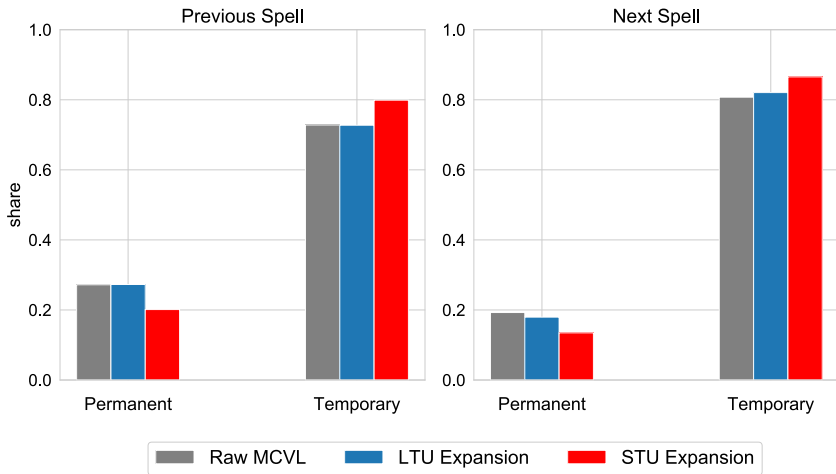


Fig. 6 Contract type before and after. All spells in the 2005–2013 period, by expansion. *Source:* MCVL

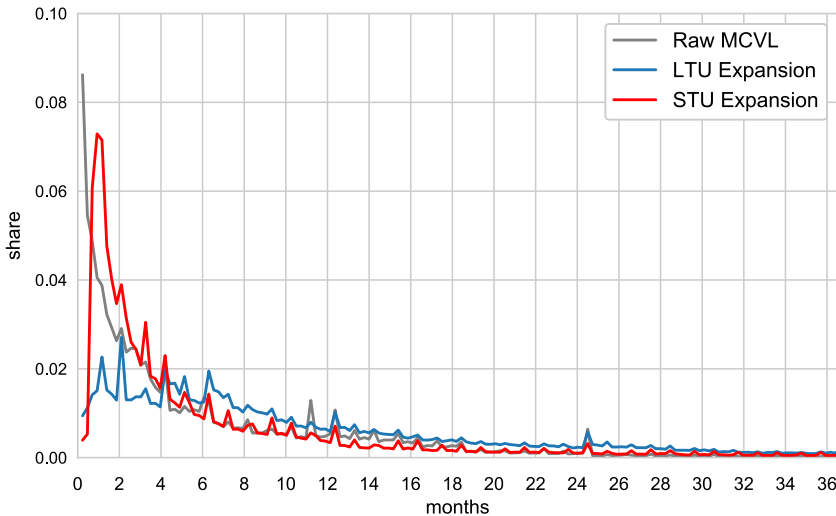


Fig. 7 Histogram of spell duration, by expansion. *Source:* MCVL. Raw MCVL refers to spells not affected by either expansion, while LTU and STU expansion refer to spells affected by each expansion, respectively

results. These very short registered spells in the raw data are unlikely to appear in the unemployment calculations. Table 3 in “Appendix” provides more detailed statistics of the unemployment spells by expansion.

Figure 8 shows that after the STU expansion, the MCVL unemployment rate gets closer to the LFS, although there is still an overestimation of unemployment before 2009. The differences become smaller towards the end of the sample. It is not surprising that the STU expansion results in more unemployment than the LTU expansion in the 2005–2008 period. These years coincide with the construction boom during

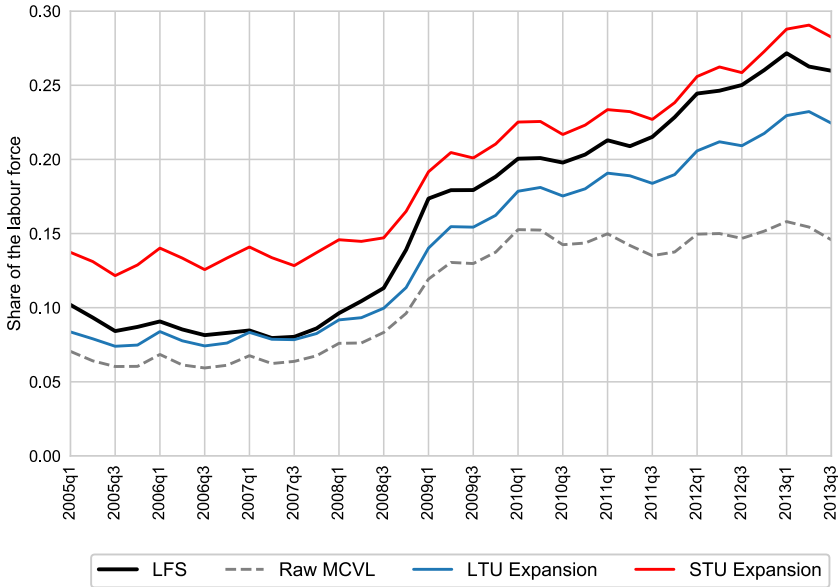


Fig. 8 Unemployment rates—short gaps expansion. Source: MCVL and the LFS

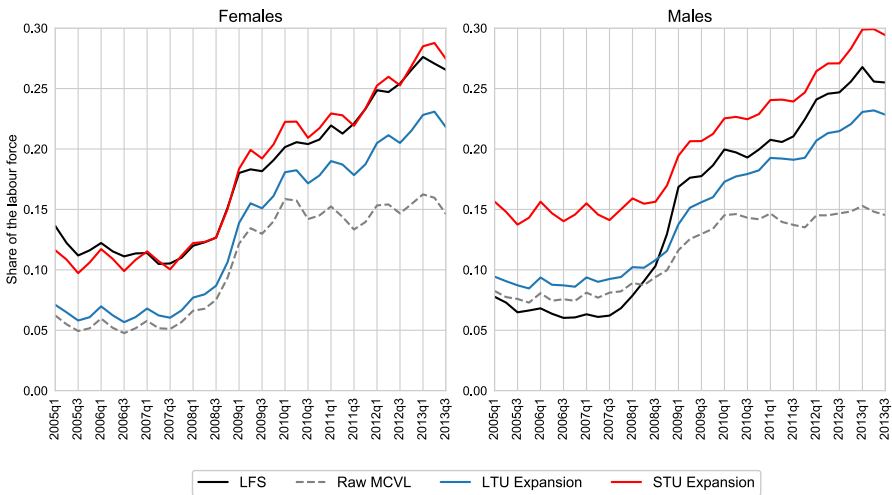


Fig. 9 Unemployment rates by gender. Source: MCVL and the LFS

which temporary contracts represented over 30% of total employment. In the following years, the gap reduces as the importance of long-term unemployment grows. The STU expansion has a similar trend to the LTU expansion and the LFS.

We can gain some insights into the difference between the two expansions by looking at the unemployment rates broken down by gender (Fig. 9) and age (Fig. 10). By gender, the STU expansion brings the MCVL closer to the LFS for women. The incidence of temporary contracts and part-time employment is higher among women,

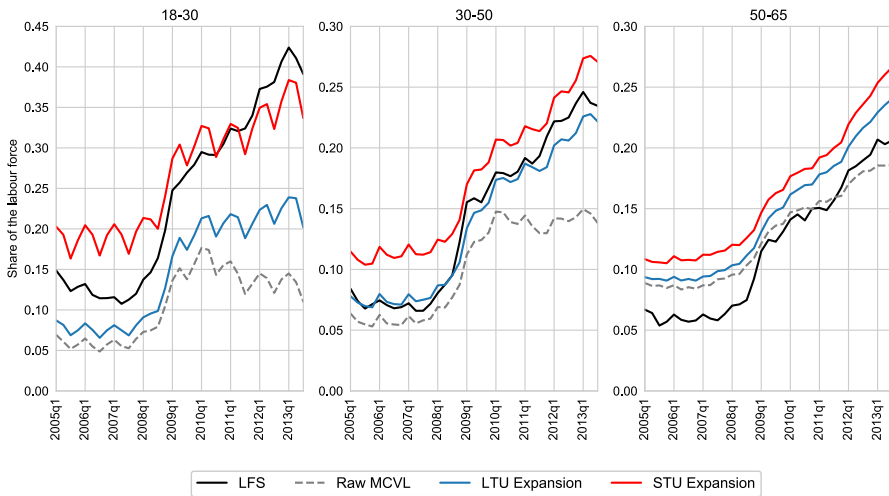


Fig. 10 Unemployment rates by age group. *Source:* MCVL and the LFS

which may explain why non-claimants of unemployment benefits seem to matter a lot for their unemployment rate. The STU expansion explains less of the gap for men, particularly before 2008. Notice that even the raw MCVL overestimates unemployment in this period. The construction sector was booming in the 2005–2008 period. This sector employed mostly young men in temporary contracts, which could explain the overestimation of the STU expansion in this period. After the recession, the gap becomes much smaller.

Looking at age: the STU expansion helps to reconcile the unemployment rates of younger workers, in a way that the LTU expansion is not able to match. There is a positive gap in the 2006–2008 period, likely due to young men on temporary contracts. For middle-aged workers, the differences mirror those in Fig. 8: unemployment is overstated by the STU expansion, more so at the beginning. Here the LTU expansion arguably performs better. For older workers, there is little difference between expansions, perhaps because of the smaller incidence of frictional unemployment among older workers. Each of the expansions offer a more coherent picture of the overall labour market than the raw MCVL series.

The importance of the LTU and STU expansions changes before and after the recession. Figure 11 shows the histogram of all unemployment spells (not just the ones in the panel) before and after 2008 by expansion and contract type. Both LTU and non-expanded spells fall during the recession for both contract types, reflecting the longer durations of regular unemployment. On the other hand, STU expanded spells fall during this period, this is mostly due to those coming from temporary contracts. Figure 12 shows this fall is mostly driven by quits and also by a small drop in short-term employment (those who are ineligible to claim benefits). The fall in the number of quits reflects that during the expansion it was easier to find jobs for those who quit their previous employment. This points towards lower mobility during the recession: workers that would separate during an expansion prefer to keep their jobs during a

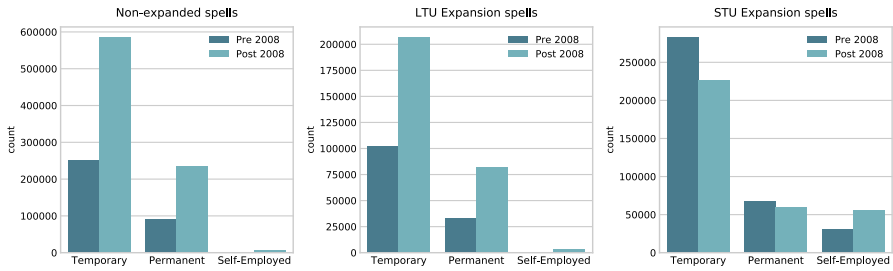


Fig. 11 Unemployment spells by expansion, before and after the 2008 Recession. *Source:* MCVL. Raw number of unemployment spells, by unemployment expansion and contract type. Pre 2008: 2004–2008. Post 2008: 2009–2013



Fig. 12 STU expanded spells by type, before and after the 2008 Recession. *Source:* MCVL. Raw number unemployment spells affected by the STU expansion, by category. Pre 2008: 2004–2008. Post 2008: 2009–2013

recession. Figures 24 and 25 in “Appendix” provide a more detailed breakdown by year.

These results confirm the idea that the LTU expansion is the main driver of the widening gap between the LFS and the MCVL after the recession. In addition, accounting for unregistered unemployment is important (particularly for young workers and women) but it is a less relevant source of unemployment in the recession.

5 Robustness checks

This section checks the consistency of the results from the previous section against different ways of counting unemployment in the MCVL. In particular, I look at the maximum length of extension for unfinished spells, the two reference weeks (whether they are on the first, second or third month of the quarter) and the rules for choosing overlapping spells in the reference period. Additionally, using information from the

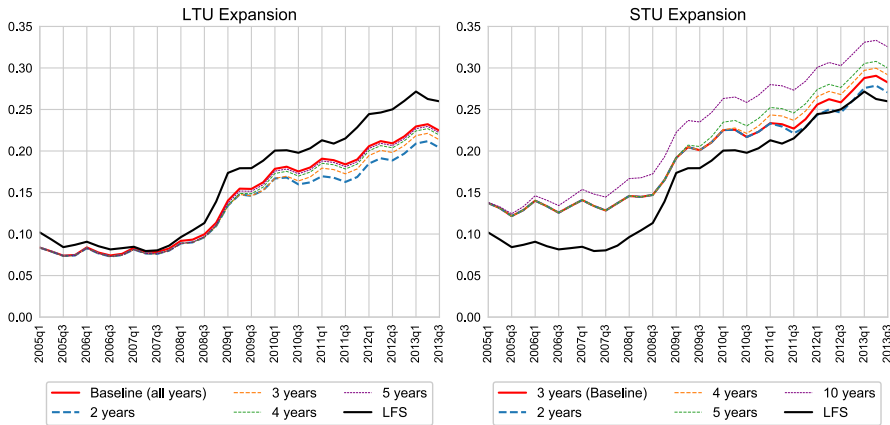


Fig. 13 Maximum duration of expanded unfinished spells, by expansion. *Source:* MCVL and the LFS

LFS, I provide more evidence of the kind of unemployment added in each expansion. Because the LFS allows unemployment to be distinguished from inactivity, and workers report whether they are receiving benefits, we can say something about the possible biases upwards and downwards of the MCVL.

5.1 Expanding unfinished spells

Recall that when implementing the LTU expansion it was crucial to extend unfinished unemployment spells until the end of the sample. However, a possible concern is that some of these extensions correspond to individuals who are dropping out of the labour force.

Notice that one of the advantages of the MCVL is that it allows for the identification of pension beneficiaries. Therefore, individuals transitioning to retirement or who have become permanently incapacitated will not be included in this expansion. The cases of individuals dropping out of the labour force that are likely to be captured by this expansion include emigrants, full-time (unsupported) carers and students.

A simple test of the decision to extend unfinished unemployment spells is to restrict the extension to those spells ending within the 2, 3, 4 or 5 years before the end of the sample. In this way, we can restrict some of the long-term non-participants from the LTU expansion. The left panel of Fig. 13 shows the result of imposing these restrictions. There is very little difference overall, with the most substantial difference between the 2-year restriction and the original being two percentage points. There is also a noticeable gap between the 2- and the 3-year restrictions. The likely reason for this distance is the increase in job destruction in 2008–2009. Workers losing their jobs in these years with a maximum extension of 2 years should lose their benefits around 2010–2011. Overall, it seems that not imposing any restriction does not lead to a persistent, noticeable increase in measured unemployment.

The right panel of Fig. 13 shows the results of a similar exercise for the STU expansion. Recall that in the STU expansion unfinished, non-registered unemployment spells

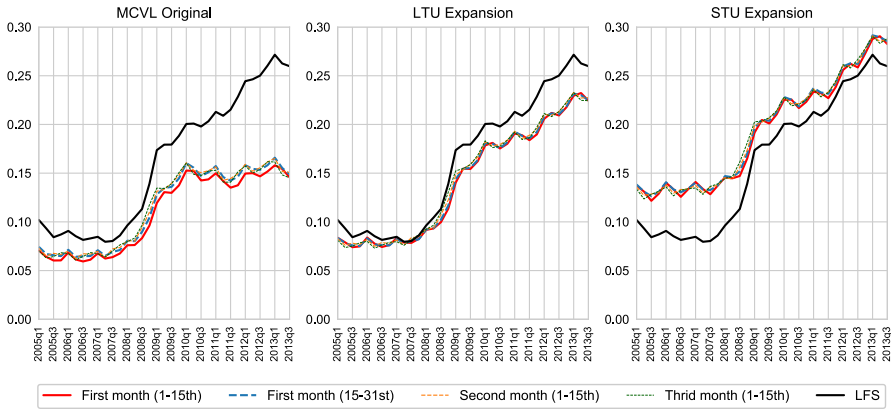


Fig. 14 Changing the reference weeks, by expansion. *Source:* MCVL and the LFS

within 3 years of the end of the sample are extended. This extension can potentially capture more inactive workers than the LTU expansion, as the latter required workers to be actively searching for employment before losing their benefits. By not registering with the employment office, individuals added in the STU expansion could be signalling that their intention is to become inactive.

The results in Fig. 13 suggest that allowing all unfinished spells in the sample to be extended increases measured unemployment noticeably: the gap between the baseline (3 years) and extending all spells reaches 4.6 percentage points at the end of 2011. By the end of the sample, the differences among the rest of the series are in the range of 1 percentage point. However, note that the 2-year threshold series misses the trend of the LFS. The 3-year and 4-year thresholds manage to capture the upwards trend of the 2011–2013 years better. This better fit is likely due to their ability to capture workers that were dismissed in this period, which saw an increase in job destruction. This exercise provides some empirical backing for other studies that also choose a 3-year limit for non-employment spells, as in the case of Rebollo-Sanz and García-Pérez (2015) and Bentolila et al. (2017).

Overall the differences are not large, except for when expanding all unfinished STU spells without restrictions. This result supports imposing some restriction on these extensions in the STU expansion. The LTU expansion yields similar results to the baseline.

5.2 Different reference periods

When constructing the unemployment series, I chose to focus on the first two weeks of each quarter. Recall that the LFS carries out interviews throughout the three months of each quarter. What would happen if we choose a different period for selecting worker status in a given quarter?

Figure 14 shows the results of using different reference fortnights for the raw MCVL series and both expansions. The lines are very close to each other. This result supports the idea that the choice reference period does not influence the results. In Figs. 26, 27

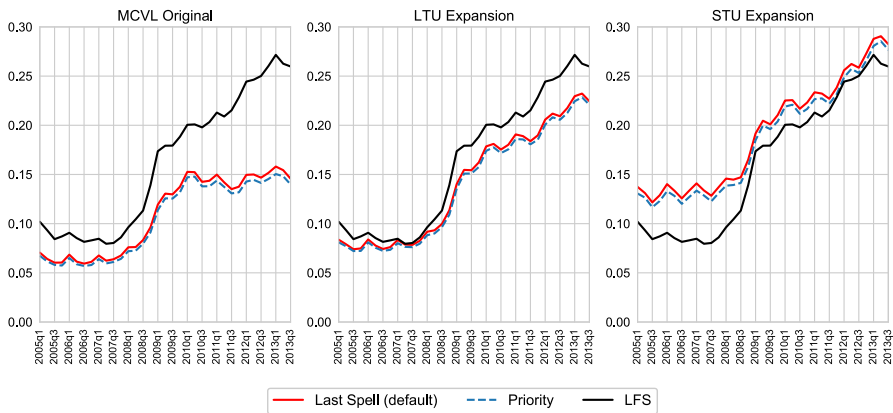


Fig. 15 Different choosing criteria, by expansion. *Source:* MCVL and the LFS

and 28 in “Appendix”, I take the first difference of all series and compare the seasonal patterns to those of the LFS. Choosing the first month of each quarter seems to deliver the closest fit to the seasonal patterns of the LFS. This additional result gives a weak preference for the first fortnight, but this choice is ultimately irrelevant to the results.

5.3 Different overlapping spells criteria

In some cases, workers have more than one spell in the two-week reference period. The last spell in the quarter is used in order to decide the state of the worker. For example, if a worker starts a period unemployed but ends with a temporary job the temporary job is used. The assumption is that the individual has a good idea of her situation by the end of the two weeks. However, this may not be the case for many workers.

An alternative would be to give preference to employment. The LFS asks individuals to report if they have worked in the last week. If they respond affirmatively, the LFS classifies them as employed—even when workers know they are going to be dismissed soon or are already non-employed.

Figure 15 shows the result of both approaches for the MCVL original series and all expansions. As expected, the unemployment rates when preferring employment are marginally lower. However, the differences are not substantial. The most significant discrepancy, in the STU expanded series, is of 0.7%. These small differences are not relevant for comparing unemployment, but they may matter for other applications.

5.4 Alternative LFS measures

The motivation of the LTU expansion was that the MCVL was excluding individuals whose unemployment benefits have expired. The LFS includes a variable that codes the self-reported relationship to the Public Employment Office (INEM in Spanish). This question is answered by all respondents, as some workers out of the labour force

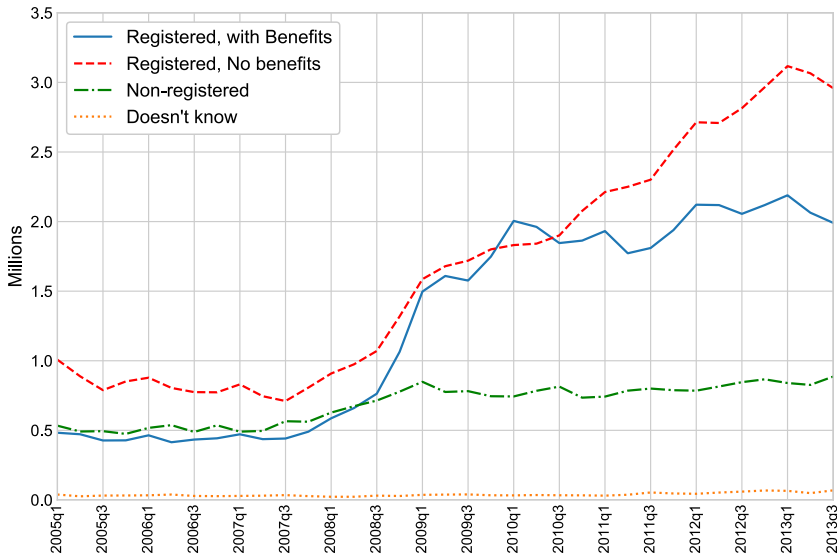


Fig. 16 Relationship with the Employment Office. Source: LFS

may be receiving benefits, such as pensions, or temporary or permanent incapacity. It is possible to quantify the number of unemployed workers who report receiving benefits using this variable.

In particular, the possible answers are ‘Registered, with Benefits’, ‘Registered, No Benefits’, ‘Non-registered’ and ‘Doesn’t know’. The ‘Registered, with Benefits’ answer will be recorded as unemployment in the MCVL, while ‘Registered, No Benefits’ should not. Unemployed people without benefits are the group that the LTU correction is targeting. While individuals reporting not being registered can either not have registered because they are not eligible for benefits or because they are first-time job seekers. The MCVL is unable to capture first-time job seekers, but the STU expansion should capture those ineligible to claim benefits.

Figure 16 shows the evolution of the responses (in millions) in the 2005–2013 period. The ‘Registered, with Benefits’ line looks very similar to the raw MCVL series, as expected. The stock of ‘Registered, No Benefits’ on the other hand keeps increasing beyond 2009, reflecting the workers whose benefits expired after the recession. The amount of non-registered unemployed increased only slightly during the recession. Using these stocks, we can construct alternative unemployment rates and compare them to the MCVL and its expansions.

Figure 17 shows the results of this comparison. In all panels, the *Registered with UI* line represents the unemployment rate that considers only unemployed people who report receiving unemployment benefits. The *All registered* line adds the unemployed registered but without benefits, which should correspond to the LTU expansion. Finally, the solid black line represents all of the unemployed in the LFS while the red line corresponds to the MCVL and its expansions.

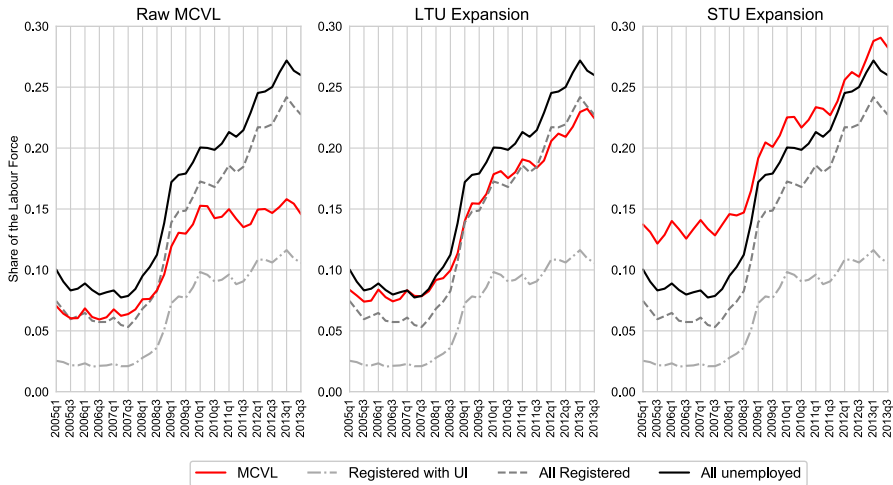


Fig. 17 Alternative unemployment rate series—registered unemployed. *Source:* MCVL and the LFS

The first panel shows raw MCVL unemployment is consistently over *Registered with UI*. However, both series have very similar trends. The difference can be coming from two sources: those who report receiving unemployment benefits but are nevertheless registered and those who are claiming unemployment benefits but who report as not actively searching for employment. There are reasons why a person may still be registered with the employment office even if she is not receiving unemployment insurance. For example, in order to claim discounts and other benefits. The MCVL will record these cases.

The second panel shows how the expanded LTU follows the *All registered* line pretty closely, even after the recession. This result provides a strong argument in favour of always using the LTU expansion when working with the MCVL. The final panel shows the difference between the STU expansion and the rest of the lines. As discussed in Sect. 4.1, the STU overestimates unemployment when compared with the LFS, particularly before 2008.

Does this disparity come from people who are out of the labour force but still claim unemployment benefits? The raw MCVL series will capture those individuals. Figure 18 shows three different measures. *All with benefits* includes all individuals that are registered and claiming benefits—both active searchers and non-participants. *All registered* adds the former active searchers and non-participants that are not receiving unemployment benefits but that are registered with the employment office. Finally, *All registered + non-registered unemployed* adds non-registered unemployed workers.

Consider the raw MCVL data (first panel of Fig. 18). If this measure included all active and inactive claiming benefits, it should match the *All with benefits* line. This is not the case, as the raw MCVL data lies above this alternative measure. As discussed above, the MCVL may be picking up some workers who are registered but not receiving benefits. The fact that the raw MCVL is higher than the sum of active and non-active searchers means that it is capturing some extra-registered unemployment.

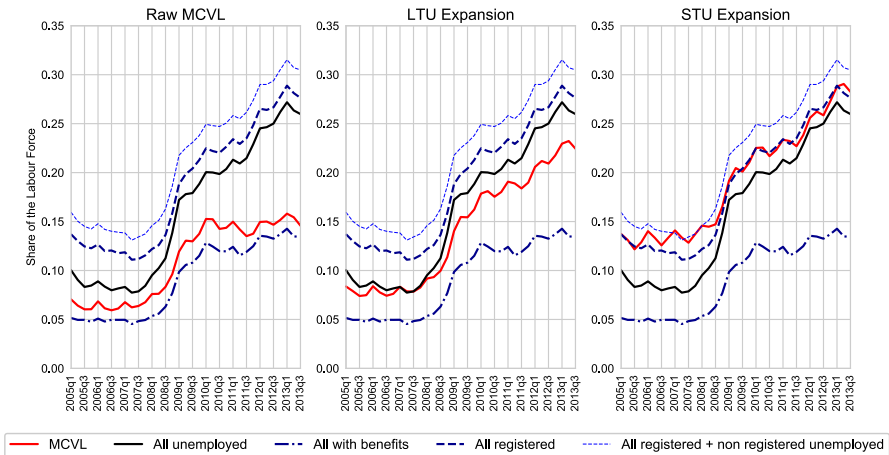


Fig. 18 Alternative unemployment rate series—all unemployed. *Source:* MCVL and the LFS

The second panel of Fig. 18 is crucial. If we include all registered individuals (with benefits or not, active or not) then the LTU line should align with this measure. But this is not the case, as the LTU line is always below this measure. This result shows that the LTU expansion (and by extension, the raw MCVL) is not simply picking up registered non-employment. The individuals in the MCVL are mostly highly attached individuals.

This idea is further reinforced by the third panel: if the STU expansion was including all unemployed workers and all registered non-participants, it should align with the *All registered + non-registered unemployed* line. However, it is consistently below except a brief period in 2007. Based on the duration of the spells and the demographics, the STU must be capturing some unregistered but highly attached individuals. We know that most of the spells added by the STU expansion are mostly short, in between job periods. These unemployment spells are hard to capture in the LFS and thus should not appear as either unemployed or out of the labour force.

6 An application to labour market flows

This section presents an application that combines both datasets, using the MCVL to address two issues of the LFS: attrition (non-responsiveness) of unemployed workers the effects of changes in survey design in 2005 in labour market flows. These flows are very relevant, as they help us to understand unemployment dynamics.

6.1 Attrition and labour market flows

The LFS is a rotating, panel, such that each household is interviewed in 6 consecutive quarters. I define *attrition* as a respondent failing to respond to two interview, one after the other. The size of the attrition bias has not been constant over time. Figure 19 shows the share of respondents who report being unemployed any given quarter but do not

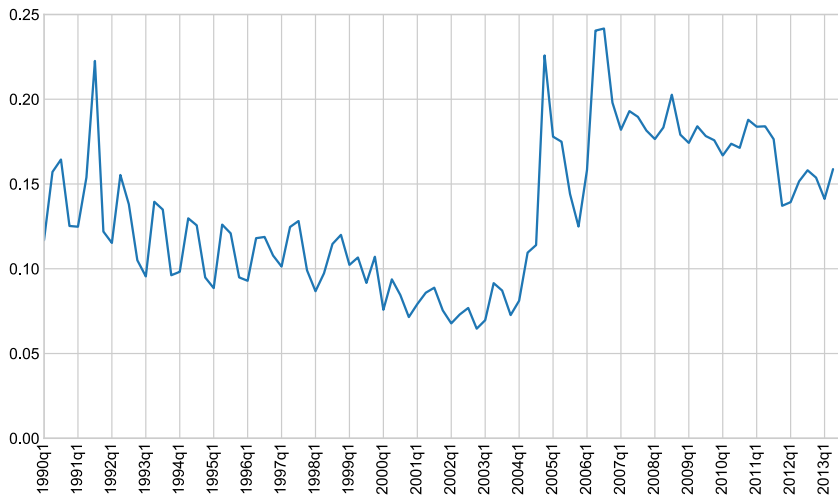


Fig. 19 Attrition of unemployment stocks in the LFS. *Source:* LFS

respond to the survey in the next quarter. These workers are not in their last interview, so they should have answered. For example, about 8% of all individuals reporting being unemployed in the 2000–2005 period do not respond in the next quarter. After 2005, that number shoots up to over 15%, reaching 20% in some quarters.

The LFS tries to correct for this problem by changing the weights attached to the observations and introducing more people into the sample. These modifications make stocks consistent over time. However, if we want to calculate labour market transitions, the weights alone do not solve the problem. This issue is not unique to the Spanish LFS. Labour market flows researchers follow different approaches to correct for attrition. For example, Silva and Vázquez-Grenno (2013) and Elsby et al. (2015) take the stocks as given, and adjust some transition rates, so that the flows are consistent with the evolution of the stocks.

I use a much simpler approach. Consider the transition from state X to Y as the number of observed individual transitions between X and Y , divided by the sum of all individual transitions starting from X , as Eq. 1 shows:²⁴

$$\lambda_{t,flows}^{XY} = \frac{Y_{t+1}|X_t}{\sum_Y Y_{t+1}|X_t} \quad (1)$$

Assume that there is attrition in this data, but rather than having an effect on the flows from X to Y it affects the number that remain in their original state. Then the denominator would be *lower* than it should be, as the non-respondents are not in the sample. Consider instead the transition rate defined in Eq. 2: number of observed individual transitions between X and Y , divided by the number of observed individuals in state X .

²⁴ I define transitions rates *forward*—from one quarter to the next. The literature tends to use the *backwards* approach—transitions from the previous quarter to the present. This distinction does not matter for results.

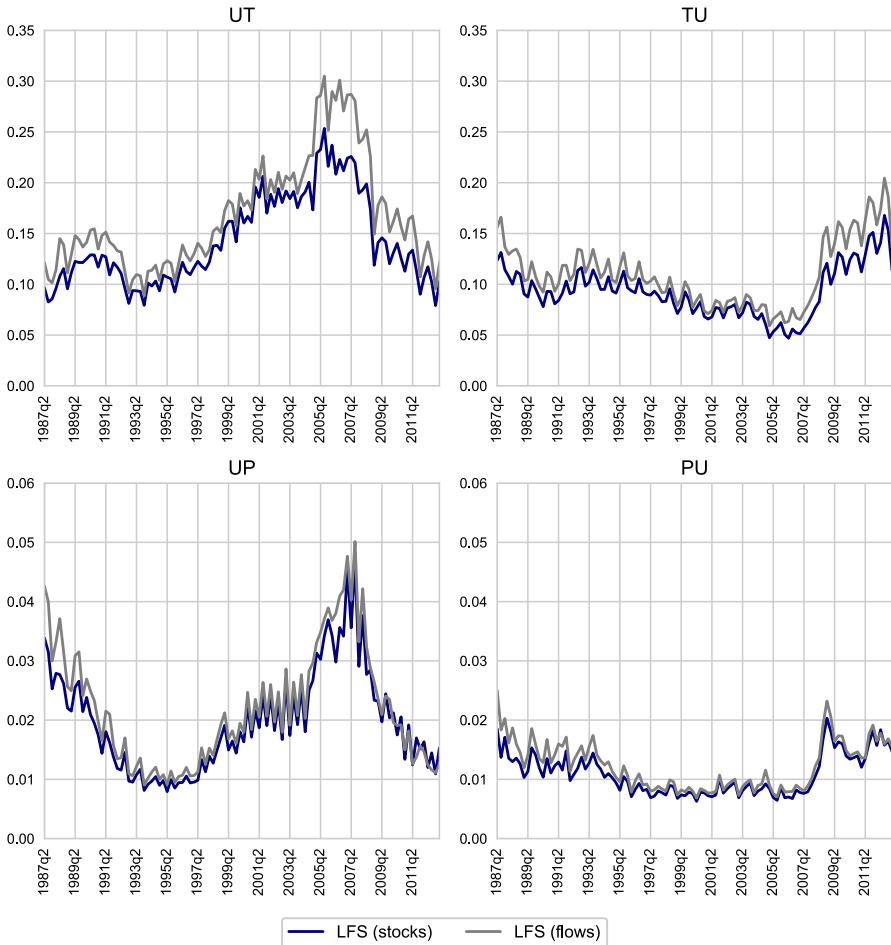


Fig. 20 Labour market flows in LFS. *Source:* LFS. Shows quarterly observed transition rates, using as the denominator: the observed stock of workers in a given state (stocks) and the sum of the flows originating for a given state (flows)

$$\lambda_{t,stocks}^{XY} = \frac{Y_{t+1}|X_t}{X_t} \tag{2}$$

This way, the transition rate would be consistent with the data. In practice, attrition can affect all of the flows out of state X , so the resulting bias of $\lambda_{t,flows}^{XY}$ is ambiguous. We can consider the case of $\lambda_{t,stocks}^{XY}$ as the extreme case when all of the attrition comes from those who remain in their original state. Figure 20 shows the evolution of $\lambda_{t,flows}^{XY}$ and $\lambda_{t,stocks}^{XY}$ from 1987 to 2013. There is not much difference between the two except for the flows between unemployment and temporary contracts. Here the gap is very noticeable in the 2005–2008 period, which coincides with the attrition spike in Fig. 19. The gap is also noticeable for the temporary to unemployment (TU) flow after 2008.

The MCVL does not suffer from this bias, as we can observe more precisely the changes in labour status of workers. Once workers are added to the data, they remain in it. The definitions of unemployment are different in both, as discussed, but given that the expansions align them more closely we can compare the resulting transition rates to the LFS. As the MCVL does not suffer from attrition issues, comparing the LFS and the MCVL can give us some insight into the source of the discrepancies in the LFS flows that are due to attrition.

Of course, the MCVL does not identify non-participants. Being able to capture non-participation is the main advantage of the LFS. We know that there are significant flows between unemployment and inactivity. These flows have been extensively discussed in the literature.²⁵ The approach in this paper is to focus on flows among labour force participants as it is hard to separate non-participation and unemployment in the MCVL. As discussed previously, trying to identify non-participation in the MCVL presents an empirical challenge beyond the scope of this paper. Ignoring transitions into inactivity may bias the flows from the MCVL downwards, as the denominator (stock of those remaining in unemployment) may be overstated. Another potential source of bias is that some non-participants may be mistakenly included in the denominator as well. However, as shown in Sect. 5.4, there is evidence that not all non-participants claiming unemployment benefits are in the raw or expanded MCVL. While it is not possible to quantify this bias, it is not too large.

Another source of downward bias in the MCVL is the frictional unemployment it captures. Some frictional or very short-term unemployment is captured in the raw MCVL (see Fig. 7), and the STU expansion increases this further. If the LFS fails to capture these short-term workers, as discussed previously, then the flows out of unemployment will be higher in the MCVL.

Figures 21 and 22 compare the flows resulting from the LFS to the MCVL. The *LFS (flows)* line shows the transition rates from the LFS calculated as in Eq. 1 (the denominator being the sum of transitions) while the *LFS (stocks)* line shows it as in Eq. 2 (the denominator being the stock).²⁶ The blue lines correspond to the LTU expansion and the red line to the STU expansion. Given the increase in the attrition of unemployed workers in 2005, I present the series from 2003.²⁷

In general, the level and trend of the flows are close between the two datasets. The MCVL series has both higher seasonal variation, which is due to the higher frequency of the data. The LFS struggles to capture these seasonal increases, leading to a smoother series. Notice as well that the *LFS (flows)* line is always higher than *LFS (stocks)*. This suggests that the unemployed individuals who do not respond to the next interview remain unemployed.

Figure 21 shows the flows out of unemployment in the LFS and the MCVL in its two expansions. The first thing to note is that in both the LFS and the MCVL the flows to temporary contracts are an order of magnitude higher than those to permanent jobs. Regarding flows to temporary contracts, before 2005 both MCVL expansions

²⁵ See for example Elsby et al. (2015).

²⁶ When calculating the stock, I naturally exclude those who are in their last interview, as they would not reply in the next quarter because they are out of the sample.

²⁷ The observations from before 2005 are from the 2005 file, so there might be some small representative issues. These are unlikely to be substantial, as we only go back for 2 years earlier in the sample.

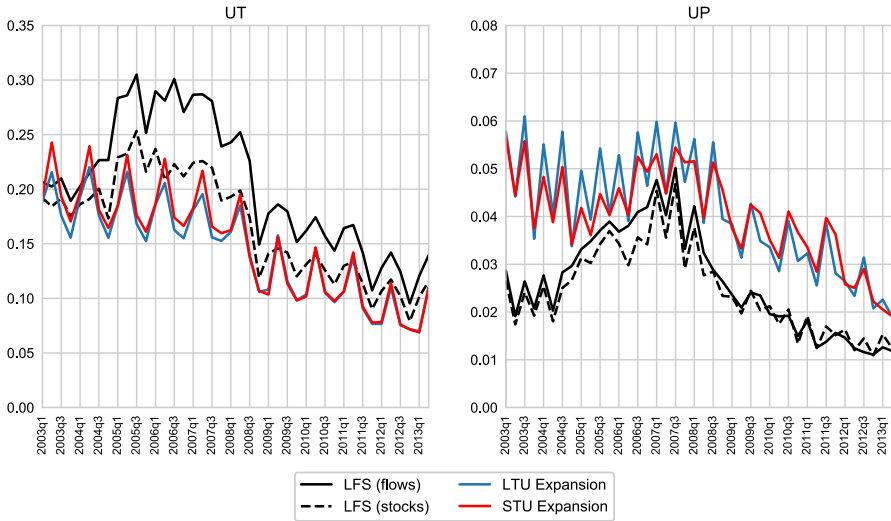


Fig. 21 Flows out of unemployment

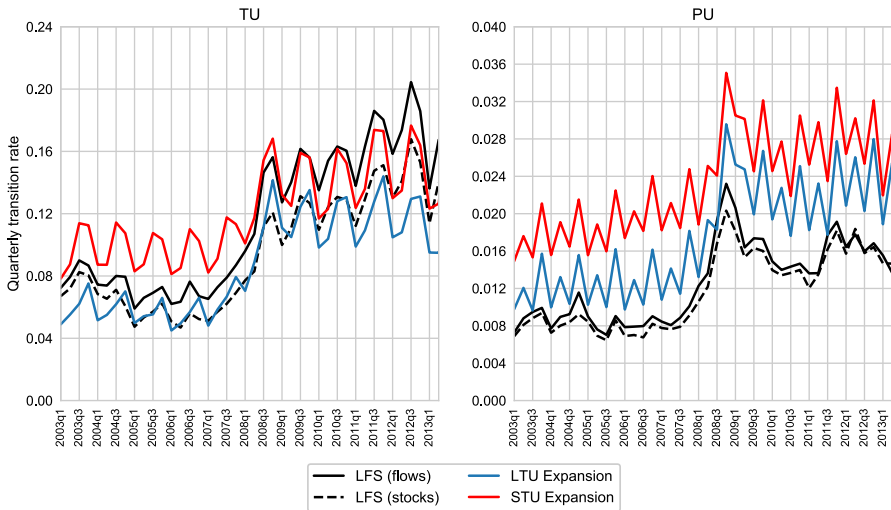


Fig. 22 Flows into unemployment. Source: MCVL and the LFS

are very close to the stock version of the LFS. The MCVL exhibits a more pronounced seasonality, as noted before. Then, in 2005, the transition rates from the LFS increase sharply. This increase is more pronounced in the *LFS (flows)* series, where it stays well over 0.25 until 2008. There is no evidence of a similar break in the MCVL series. This result shows that the attrition problem may be behind the large flows to temporary contracts observed in the LFS using the flows accounting approach. But this does not explain why using the stocks approach also results in a visible increase in 2005—an

increase that is not backed by the MCVL flows. The break in the survey design of 2005 may also play a role.

After the recession, the distance between all series reduces. The STU and LTU flows are almost identical after 2008, while before that the STU was above the LTU series. This disparity is likely due to the frictional unemployment captured by the STU during the boom. Notice that if we look at the MCVL or the stocks LFS, the fall in the job finding rate was not as dramatic as implied by the LFS. This finding is very relevant for papers that try to decompose the variance of unemployment, such as Silva and Vázquez-Grenno (2013).

The right panel of Fig. 21 shows that the unemployment to permanent flow is higher in the MCVL, but the differences are small—notice that the scale is only from 0 to 8%. The missing contract modification variable before 2005 could explain the divergence of the MCVL prior to that year. This result is a reminder that the data cannot be used retrospectively without problems, and of the importance of the contract modification cleaning step that is outlined in “Appendix”. The fall in the unemployment to permanent flow is not as sharp in the recession as the unemployment to temporary flow.

Figure 22 shows the results for flows into unemployment. As before, the STU expansion seems to be adding short spells from temporary contracts that otherwise would count as job-to-job transitions. The LFS does not capture these quick changes well and has a tendency to smooth them out, so both LFS series are below the STU expansion before the crisis. Recall that the denominator in this case is the stock of temporary contracts. There should, therefore, be less bias in the MCVL than in the case of flows from unemployment. This changes after the recession, as job destruction increases considerably. Towards the end the LTU series falls below the LFS, while the STU series aligns with the LFS. This result underlines the importance of capturing unemployed individuals without the right to claim benefits that have not found a job by the end of the sample. As a final note, the unemployment to permanent rates are higher in the STU than the LTU series, and both are above the LFS. The absolute differences are minimal, as this rate is again an order of magnitude less than the temporary to unemployment flow.

In conclusion, the flows from the expanded MCVL are very close to the LFS. This result provides strong evidence that the MCVL is capturing actual unemployment. If the workers added by the expansions were non-participants, they would behave differently from regular unemployed workers in the LFS. That does not seem to be the case. Moreover, the MCVL suggests that the volatility of the unemployment to temporary transition rate is lower than what the LFS suggests after 2005. The combination of both datasets brings new insights into how labour market flows behave in Spain.

6.2 Changes in survey design: temporary to permanent flows

Attrition is not the only challenge when computing labour market flows with the LFS, changes in the structure of the interview can cause severe discontinuities. These breaks are not present in stocks, because the National Institute of Statistics (INE) ensures that

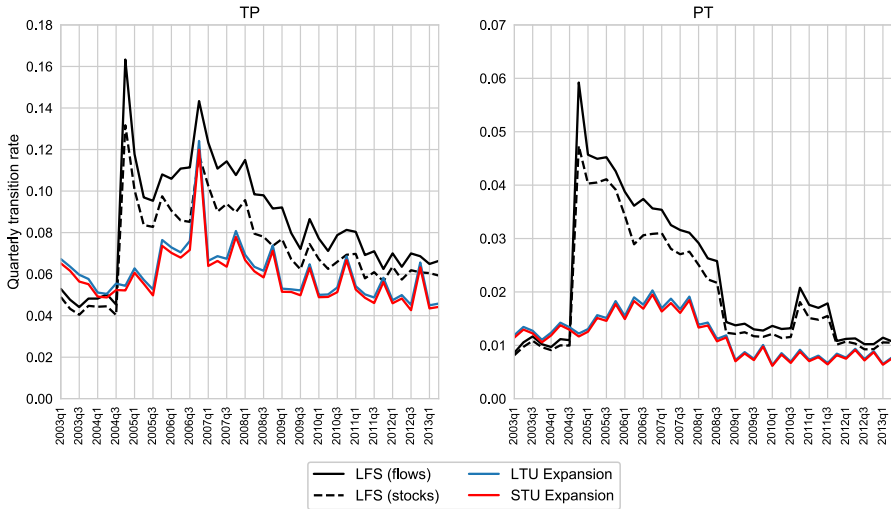


Fig. 23 Quarterly flows: between contract types. *Source:* MCVL and the LFS

stocks are consistent over time. Figure 23 shows one of the main breaks in the flows between different types of contracts (TP and PT).²⁸

The transition rate between temporary to permanent was between 4 and 5% before 2005, which was consistent with the literature on contract upgrading (see Güell and Petrongolo (2007) for an example). Immediately after 2005, the transition rate shoots up to 12% (or 16% following the flows calculation). There is another spike in 2006, which coincides with a labour market reform that encouraged conversion of temporary to permanent contracts.²⁹ The MCVL series, on the other hand, does not display an abrupt increase in 2005, although it shows a 12% spike in 2006. A natural interpretation of this discrepancy is that firms already told workers they would make them permanent employees before the contracts changed—and hence the survey responses pre-empt the administrative data. Whether this is in fact the case is a question for further research.

The right panel of Fig. 23 shows a similar pattern, where the LFS permanent to temporary flow (PT) increases from 1 to 6% before slowly falling back to previous levels. In contrast, the MCVL only increases to 1.7% before falling after the recession.

7 Conclusion

Administrative datasets are an important source of information for economists, but they also present some challenges. In this paper, I have analysed the case of the

²⁸ Other flow rates that suffer breaks relate to non-participation. However, since the MCVL cannot speak to non-participation flows, then there is nothing administrative data can add to that question.

²⁹ In particular, all temporary contracts converted to permanent before 2007 benefited from a tax exemption scheme. Firms reacted very strongly by upgrading many temporary contracts in the last quarter of 2006. This reaction suggests that firms widespread use temporary contracts is due to its lower cost. It seems that a simple tax rebate is enough to overcome all of the screening problems that the firm may have and would induce it to upgrade them to permanent positions.

Spanish *Muestra Continua de Vidas Laborales* (MCVL), a rich administrative dataset containing working histories of a representative sample of the Spanish workforce. However, it has important shortcomings in its original format. In particular, the way it records unemployment spells. In this paper, I presented an approach to expand the data by including two kinds of missing unemployment: long-term and short-term unemployment. While the first expansion is widely applied in the literature, the second expansion offers an alternative to considering all gaps as non-employment spells. I then check the results of these expansions using the LFS unemployment rates as a benchmark.

Most of the large gap between the LFS and the MCVL unemployment is explained by the workers affected by the long-term expansion. These are mostly long-term unemployed workers whose benefit entitlement expires before they find a new job. Their importance increases in the years of the crisis. However, this expansions alone underestimates unemployment, particularly for women and young workers.

The second expansion adds short-term unemployment spells of workers that are not entitled to receive unemployment benefits. After adding these workers, the gap closes down in the recession, but it overestimates unemployment compared to the LFS. This is likely due to the frictional nature of these spells, mostly linked to short temporary contracts. The changes in composition over the business cycle indicate that these spells were less common in the recession. A possible interpretation is that mobility, through quits or short temporary contracts, slowed down in the recession.

I provided further robustness checks to both expansions and the general methodology to build the panel. The results support the assumptions made, such as restricting the expansion of unfinished spells to those starting within the last years of the sample. These checks also provide empirical support to common assumptions in the literature.

Finally, I analyse the main implications for the study of labour market flows, which traditionally use Labour Force Surveys. The MCVL provides some insight into two main challenges faced by these datasets: attrition bias from unemployed individuals failing to respond for two consecutive quarters and changes in the survey. Overall, the flows from the MCVL match those from the LFS, which supports the idea that the datasets are comparable. However, there is considerable evidence that the temporary job finding rate may have been overestimated before 2008 because of attrition bias.

The MCVL and the LFS together provide a more comprehensive picture of the evolution of the labour market. There are also some general lessons that can be applied to similar administrative datasets in other countries. In particular, it is necessary to make sure that unemployed individuals without benefits count as being unemployed. Frictional unemployment, which the LFS cannot capture, is becoming increasingly important which calls for a more widespread use of microdatasets that can effectively capture high-frequency movements.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix

A.1 Formatting the MCVL

This appendix aims to give an overview of how to format and structure the MCVL in four main steps: joining files, identifying workers status, cleaning overlapping spells and building a panel. This last step allows for the comparison of the results of different unemployment expansions to the LFS. If the researcher wants to use the MCVL as a series of spells per worker, there is no need to build a panel.

Please note that this appendix gives a general overview of the methodology developed in this paper. The full online appendix gives a more comprehensive, step-by-step guide.³⁰

Joining the files

The files containing the working histories are the “*ficheros de afiliación*” files. These have different names depending on the wave of the MCVL but are easily identifiable. After appending these files, so all observations are in one folder, two other files need to be merged: the “*personal*” file and the “*pensiones*” file. The first file provides essential individual information, such as the date and province of birth, gender and nationality. The “*pensiones*” file provides retirement information, which allows identifying transitions to retirement. Because the purpose of this paper is working population labour stocks and flows, I only keep the spell where the worker becomes a full-time pensioner.

The next step is to repeat this process for each year, and then join all the years in a single file. This roughly follows Arranz et al. (2011) approach. Using the start and end dates and the firm codes, repeated spells can be safely discarded. Bear in mind that later files add some information to some individuals that was missing in the previous year files.

Identifying worker status

After joining the files, the first step is creating a variable that classifies workers in four labour market status: *self-employment*, *working with a permanent (open-ended) contract*, *working with a temporary contract* and *unemployed*. It is important to separate both kinds of contract because their dynamics are very different, with temporary contracts accounting for most of the flows in and out of unemployment.

The only category missing is *out of the labour force*. The administration does not provide reliable information to judge whether an individual is participating in the labour market or not. This lack of information on participation is the main drawback of administrative datasets. There is a separate question of whether the distinction between unemployment and out of the labour force is essential for labour market flows. Looking at the LFS, for workers in the 20–60 age range, the transition rates to employment from unemployment are very similar to transitions to employment from

³⁰ This is available, together with the STATA code, at <http://github.com/crisla/MCVL>.

out of the labour force, 12% versus 10%. This similarity highlights the shortcomings of the official definition of unemployment. This discussion is however beyond the scope of this paper. For classification purposes, following other authors (García Pérez 2008; Lapuerta 2010; Arranz et al. 2011) I assume registered unemployed workers are genuinely unemployed, and that retired workers and periods where the MCVL has no information are out of the labour force spells.³¹

In the MCVL, three variables contain all the information needed to classify workers:

1. *Type of labour relationship* codes the different links each worker has with the social security—working, receiving unemployment benefits. Here it is possible to identify unemployed workers.
2. *Contract type* contains the code for each type of labour contract. There are 557 different contract codes in the registry, but most of them are “legacy contracts” that do not exist in the present.³² Most temporary contracts are grouped under the 400s numerical codes while regular contracts are in the 100s. This way I distinguish between temporary or permanent (open-ended) contracts.³³ Temporary contracts cannot be renewed beyond 2 years with the same firm and are subject to smaller severance payments than regular contracts. As discussed above, their dynamics are very different.
3. *Contribution class* allows for the identification of self-employed workers, as they follow a different contribution system. These correspond to variable values 500–600.³⁴

These three variables have the necessary information to classify most observations, but there are exceptional cases. Most notably, some unemployed workers close to retirement pay their social security contributions as if they were employed. By doing this, they can boost their pension. Other examples include discontinuous and seasonal workers and students that receive benefits under apprenticeship contracts. The online appendix provides a comprehensive list of these exceptions.

Cleaning overlapping spells

Many workers have different labour status at the same time. Having simultaneous spells may not be an issue of many applications, but as García Pérez (2008) argues, many overlapping spells are administrative errors—such as recording the end of one spell before the start of the next one. Another example corresponds to workers who are working part-time and receive compensatory unemployment benefits while working. These workers will have an unemployment and employment spell simultaneously. For

³¹ In particular, in order to keep their benefits, unemployed workers are formally required to: prove they are actively searching for a job, attend job interviews and not to reject job offers. The Employment Centre monitors workers at least each month upon receiving the payments.

³² There are some kinds of contract that are extinct—usually subsidised contracts created in the 1990s. These are not relevant to the present study, which focuses on the 2005–2013 period.

³³ For the special case of discontinuous workers (those who work only on specific periods every year), I treat them as permanent, as they are subject to firing costs and are open-ended.

³⁴ There are some special categories for domestic workers, agriculture workers, farmers and sailors. I select those who are self-employed in these special regimes and treat the rest as employees.

these reasons, García Pérez (2008) recommends trimming the start and end dates of spells to avoid overlaps and dropping unemployment spells that overlap with employment. See García Pérez (2008) for more details.

After these easy adjustments, a small number of spells remain where the worker has two or more jobs simultaneously. Depending on the study, these simultaneous jobs can be kept or dropped. The objective of this paper is to build a panel of labour statuses to quantify labour market stocks and flows. At this stage, I do not take a stance on whether to drop overlapping spells or whether to keep them and leave that decision to the building the panel phase.

One final note of caution relates to conversions of temporary contracts into permanent contracts and vice versa. These changes are not recorded in separate spells. Instead, the variable “modificación de tipo de contrato” keeps track of the change. Splitting these spells into two is convenient. More detailed instructions can be found in the online appendix.

Building the panel

In order to build a panel, we need to select one observation per individual per unit of time. I choose the quarter as the time unit to keep consistency with the LFS. In each quarter, I select a date at which I will evaluate people’s working states: the 1st of January, 1st of April, 1st of July and 1st of October, which coincide with the start of the year’s quarters. Because some jobs may start after that date, I also consider all the spells in the following two weeks, until the 15th of each month.³⁵ In the LFS, the interviews take place during a long period in each quarter. This way, some individuals are interviewed at the beginning and some at the end of each quarter. For workers that have two or more labour status in the same 2-week period (2% of all spells)³⁶ I give priority to the longest spell: if a worker starts the 2-week observation window unemployed but finds a job that lasts for one more quarter, I count her as employed in that quarter (Tables 3, 4, 5; Figs. 24, 25, 26, 27, 28).

³⁵ This corrects for short periods in which the worker may not be in either state. For example, it is likely that many jobs will start on the 7th of January instead of the 1st, due to the Christmas holidays in Spain ending on the 6th.

³⁶ Of these 2%, roughly a quarter belong to two or more simultaneous temporary contracts. For the purpose of building unemployment rate series, the choice between them is irrelevant.

Table 3 Summary statistics of unemployment spells

	No modification	LTU expansion	STU expansion
Observations	1,151,839	431,562	723,830
Share	0.499	0.187	0.314
Age	39.63 (10.50)	36.86 (10.32)	31.41 (10.29)
Female	0.448 (0.497)	0.482 (0.499)	0.464 (0.499)
Duration (days)	131 (193)	430 (494)	205 (286)
Experience PC	6.228 (7.668)	4.039 (5.911)	1.891 (3.842)
Experience TC	3.945 (3.081)	3.227 (2.613)	1.796 (2.086)
Experience unemp	2.416 (2.488)	2.082 (2.05)	0.749 (1.401)

Sample is all unemployment spells ending in the 2005–2013 period. Averages with standard errors in parenthesis. Experience is measured in years, duration of the spell in days. Experience of Unemp refers to previous years of experience in *registered unemployment* only

Table 4 Previous status of unemployed workers

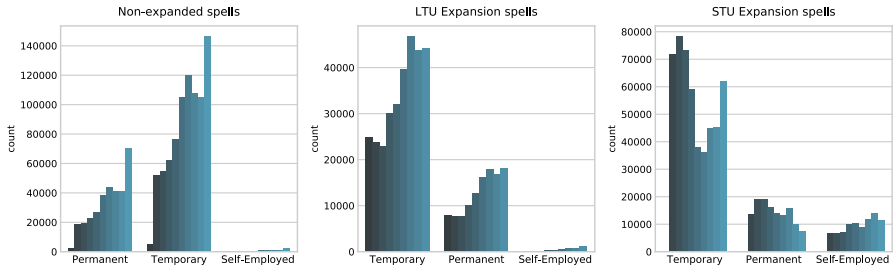
	No modification	LTU expansion	STU expansion
Self-employment	7,789 (0.007)	4,152 (0.01)	86,529 (0.120)
Permanent contract	308,146 (0.270)	115,323 0.27	128,479 (0.177)
Temporary contract	825,010 (0.723)	308,148 (0.72)	508,822 (0.703)
Total	1,140,945	427,623	723,830

Sample is all unemployment spells ending in the 2005–2013 period. Shares of the total in parenthesis

Table 5 Next status of unemployed workers

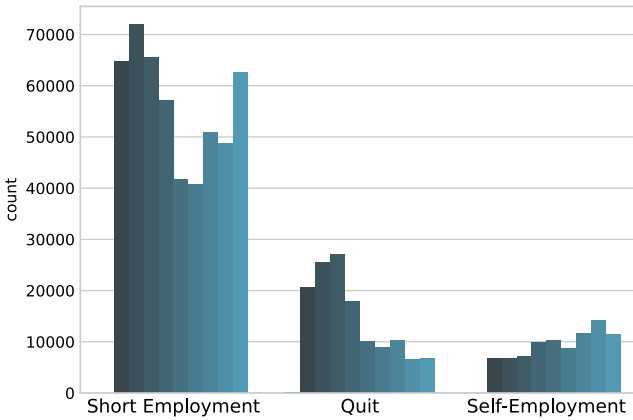
	No modification	LTU expansion	STU expansion
Self-employment	18,763 (0.018)	25,810 (0.075)	50,979 (0.079)
Permanent contract	196,503 (0.189)	57,556 (0.166)	79,477 (0.124)
Temporary contract	821,702 (0.792)	262,653 (0.759)	512,284 (0.797)
Total	1,140,981	427,652	723,830

Sample is all unemployment spells ending in the 2005–2013 period. Shares of the total in parenthesis



Notes: Each bar represents a year, from 2004 to 2013 (both inclusive). Raw number of unemployment spells, by unemployment expansion and contract type.

Fig. 24 Unemployment spells by expansion by year. *Notes* Each bar represents a year, from 2004 to 2013 (both inclusive). Raw number of unemployment spells, by unemployment expansion and contract type



Notes: Each bar represents a year, from 2004 to 2013 (both inclusive). Raw number of unemployment spells, STU expanded, by category.

Fig. 25 STU expanded spells by type by year. *Notes* Each bar represents a year, from 2004 to 2013 (both inclusive). Raw number of unemployment spells, STU expanded, by category

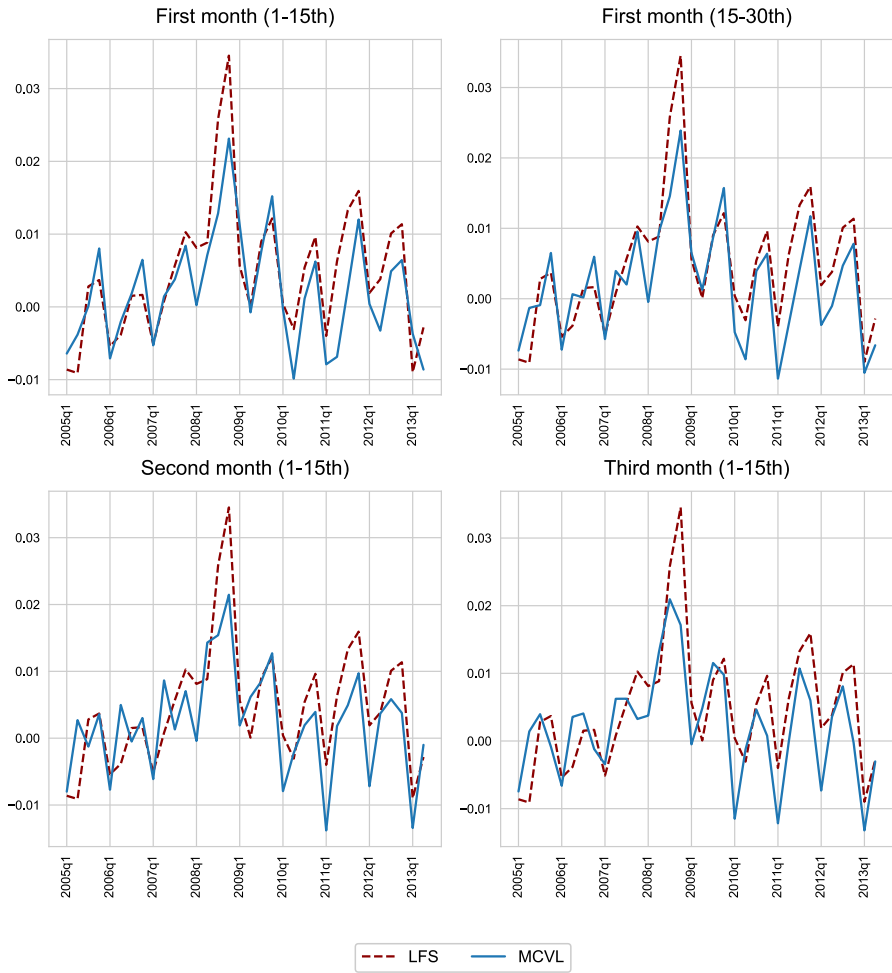


Fig. 26 Seasonality of unemployment by reference week, raw MCVL

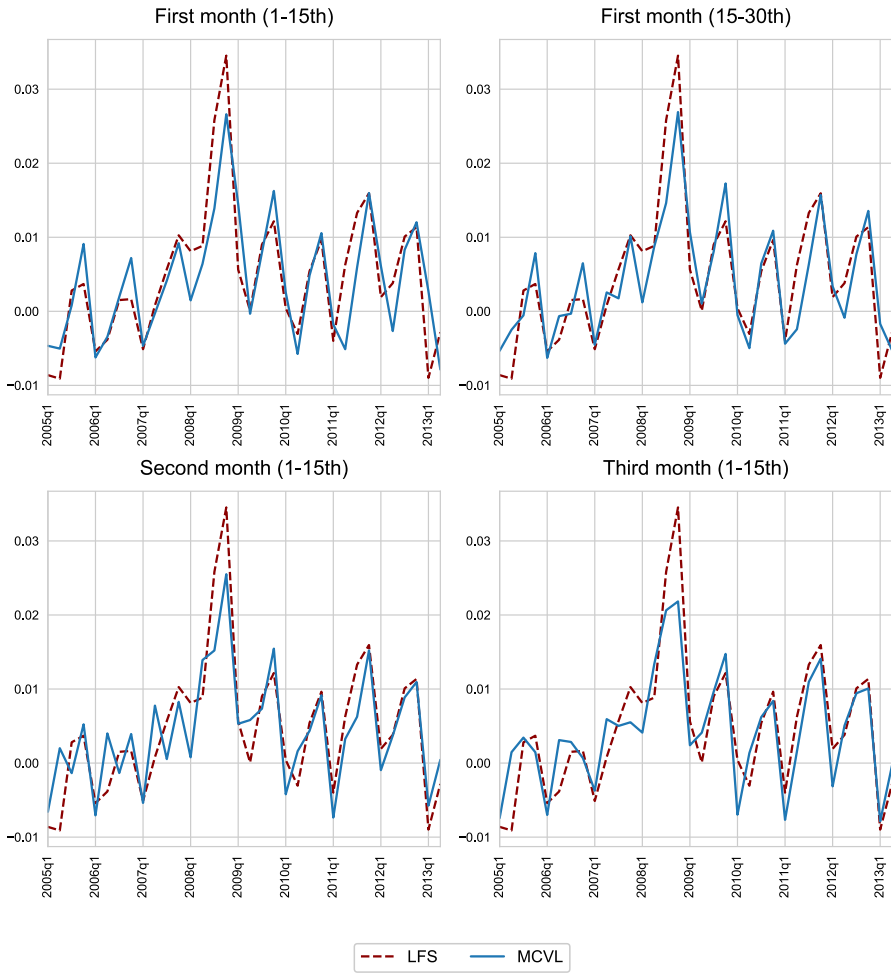


Fig. 27 Seasonality of unemployment by reference week, LTU expansion

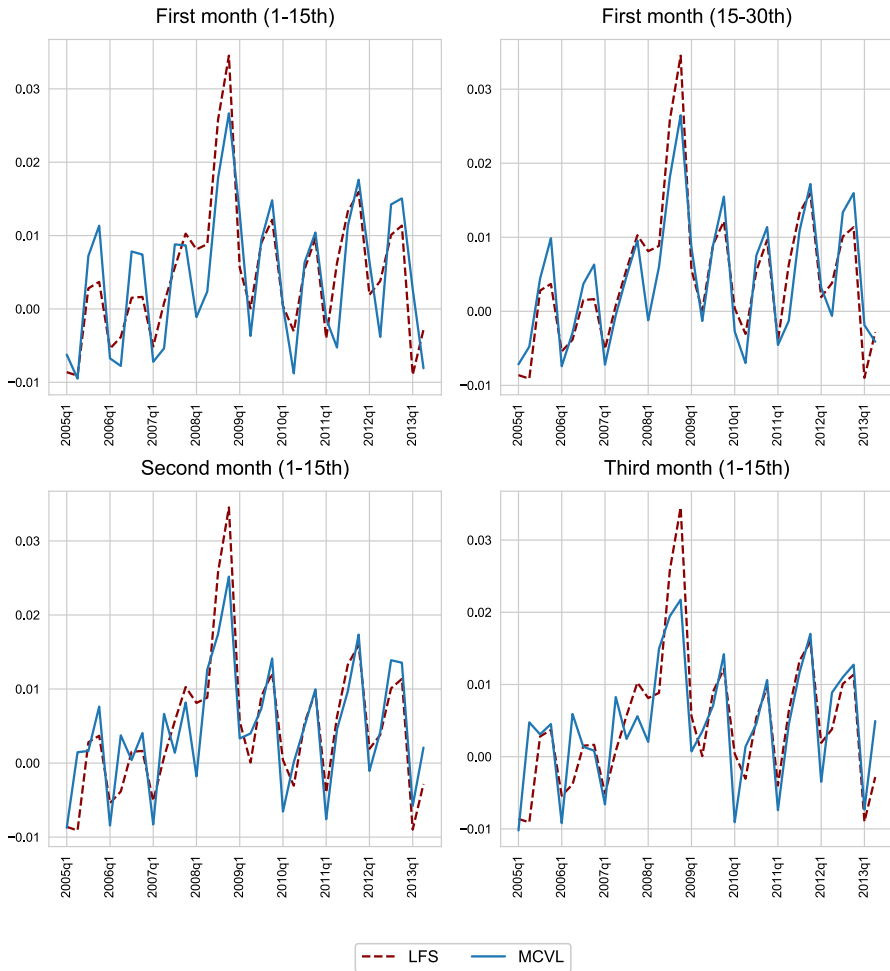


Fig. 28 Seasonality of unemployment by reference week, STU expansion

References

- Arranz JM, Serrano CG et al (2011) Are the MCVL tax data useful? Ideas for mining. *Hacienda Pública Española* 199(4):151–186
- Bentolila S, García-Pérez JL, Jansen M (2017) Are the Spanish long-term unemployed unemployable? *SERIEs* 8(1):1–41
- Bonhomme S, Hospido L (2017) The cycle of earnings inequality: evidence from Spanish Social Security data. *Econ J* 127:1244–1278
- Couch KA, Jolly NA, Placzek DW (2011) Earnings losses of displaced workers and the business cycle: an analysis with administrative data. *Econ Lett* 111(1):16–19
- Elsby MW, Hobijn B, Şahin A (2015) On the importance of the participation margin for labor market fluctuations. *J Monet Econ* 72:64–82
- Fernández-Navia T (2019) Unemployment insurance and geographical mobility: evidence from a quasi-natural experiment. *UB working papers* (forthcoming)
- Fujita S, Moscarini G (2017) Recall and unemployment. *Am Econ Rev* 107(12):3875–3916

- García Pérez JI (2008) La muestra continua de vidas laborales: una guía de uso para el análisis de transiciones. *Revista de Economía Aplicada* 16(1):5–28
- Güell M, Petrongolo B (2007) How binding are legal limits? Transitions from temporary to permanent work in Spain. *Labour Econ* 14(2):153–183
- Katz LF, Meyer BD (1990) The impact of the potential duration of unemployment benefits on the duration of unemployment. *J Public Econ* 41(1):45–72
- Krueger AB, Mueller A (2011) Job search, emotional well-being, and job finding in a period of mass unemployment: evidence from high-frequency longitudinal data. *Brook Pap Econ Act* 2011(1):1–57
- Lapueta I (2010) Claves para el trabajo con la muestra continua de vidas laborales
- López-Roldán P (2011) La muestra continua de vidas laborales: posibilidades y limitaciones. aplicación al estudio de la ocupación de la población inmigrante. *Metodología de Encuestas* 13(1):7–32
- María Arranz J, García-Serrano C (2014) The interplay of the unemployment compensation system, fixed-term contracts and rehiring: the case of Spain. *Int J Manpow* 35(8):1236–1259
- Moffitt R (1985) Unemployment insurance and the distribution of unemployment spells. *J Econom* 28(1):85–101
- Nagore García A, van Soest A (2017) Unemployment exits before and during the crisis. *Labour* 31(4):337–368
- Rebollo-Sanz Y (2012) Unemployment insurance and job turnover in Spain. *Labour Econ* 19(3):403–426
- Rebollo-Sanz YF, García-Pérez JI (2015) Are unemployment benefits harmful to the stability of working careers? The case of Spain. *SERIEs* 6(1):1–41
- Rebollo-Sanz YF, Rodríguez-Planas N (2018) When the going gets tough... financial incentives, duration of unemployment and job-match quality. *J Hum Resour* 1015-7420R2
- Roca JDL, Puga D (2017) Learning by working in big cities. *Rev Econ Stud* 84(1):106–142
- Silva JI, Vázquez-Grenno J (2013) The ins and outs of unemployment in a two-tier labor market. *Labour Econ* 24:161–169
- Sullivan D, Von Wachter T (2009) Job displacement and mortality: an analysis using administrative data. *Q J Econ* 124(3):1265–1306
- Tattara G, Valentini M (2010) Turnover and excess worker reallocation. The Veneto labour market between 1982 and 1996. *Labour* 24(4):474–500

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.