The best of two worlds

The best of two worlds: assessing the use of administrative data for the study of unemployment using the labour force survey as a benchmark

Cristina Lafuente*
University of Edinburgh

March 8, 2018

Social security administrative data are increasingly becoming available in many countries. These are very attractive data as they have a long panel structure (large N, large T) and allow to measure many different variables with higher precision. Because of their nature they can capture short, frictional unemployment which is usually hidden in survey data, due to design or timing of interviews. However, the definition of unemployment is also different in both datasets. As a result, the gap between total unemployment and registered unemployment is not constant neither across workers characteristics nor time. In this paper I augment the Spanish Social Security administrative data by adding missing unemployment spells using the institutional framework and the Labour Force Survey as a benchmark. I compare the resulting unemployment rate to that of the Labour Force Survey, showing that both are comparable and thus the administrative dataset is useful for labour market research. Administrative data can also be used to 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.

JEL classification: J21, J60, J80
Key Words: Administrative data, survey data, unemployment, temporary contracts

*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 received financial support from the ESRC. 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.

1 Introduction

Administrative datasets are being increasingly used for labour market research.[1] They offer many advantages over traditional Labour Force Surveys, from firm-worker identification to detailed and extensive working histories (large N, long T). However there are some challenges when using them for studying unemployment. First, these data were not designed for research, but rather for administrative bookkeeping: who is making contributions to the social security and who is claiming benefits. This is especially true when unemployment benefits are contribution-based, so they are proportional to the social security contributions by the worker in her previous employment. Second, the definition of unemployment is not the same as the International Labour Office standard, which is used as the official statistic for international organizations such as the OECD or the IMF. This definition typically requires the worker to be actively looking for work and ready to accept a suitable offer. Although the reception of benefits is often conditioned on active search on the side of the worker, monitoring may not be perfect. Finally, in some countries the administration only keeps track of the unemployed while they are receiving benefits.

A question that naturally arises is what should economists consider as unemployment. Most of us wouldn't expect for a worker to be constantly searching for employment for all of her time unemployed. Specially in the case of benefit recipients, we know that there is a spike around exit (Card et al. (2007), Rebollo-Sanz (2012)) so search effort may not be constant over the non-employment spell. This means that some of the non-participation in the LFS could be consider unemployment. In particular, if the individual finds a job after a few weeks of non-participation, could we say she was really non-participating? According to the ILO definition, yes, she was not in the labour market. Why did she accepted the job the following quarter then? It could be a misclassification error, something that surveys are more prone to do (see Elsby et al. (2015)). This makes administrative datasets more appealing for studying unemployment for economists, as the only classification error is the administration failing to monitor effectively for job search. However there is always some requirement that the registered unemployed need to fulfil in order to receive benefits, and the administrative dataset can distinguish those who are claiming other kind of benefits that are not related to job search (disability, maternity, pensions). This is the motivation behind the approach I follow with the administrative non-employment gaps when there is no registered unemployment: identify the cases of workers who are not considered unemployed because they can't claim job search benefits but that are *likely* searching for a different job. By excluding recalls I'm excluding those

[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.

who are not likely searching because they know they are coming back to the same firm. In doing so I am applying the ILO requirement of active search in order to be considered unemployed.[2] I argue that this is a good approach for Spain, as temporary contracts that are chained with the same firm will sometimes require that the worker is out of work for a period to be allowed to return.

The aim of this paper is to show how to implement two simple expansions in order to make administrative data ready to be used for the study of unemployment. First, unemployed workers who run out of benefits but keep searching will appear in the administrative data as if their spell was over. I add the missing days in between the end of a registered unemployment spell and the next employment spell to correct for this. Second, those who are not entitled to unemployment benefits because they had too short of a tenure in their previous job will also not appear as unemployed. Using the richness of the data and the institutional setting I identify theses cases and add these spells into the data.

The main methodological check is to compare the resulting unemployment rate from the extended administrative data with that of the Labour Force Survey. To this aim I format the administrative dataset into a quarterly panel structure as in the Labour Force Survey. The results show evidence that the different expansions achieve their aim of adding the missing unemployment to the administrative data. In particular, the second expansion is crucial for youth unemployment and for women, who have a higher incidence of part-time and temporary contracts. As further check I use the information in the Labour Force Survey to reconstruct the administrative data's unemployment rate.

Finally I show how the expanded administrative data can help overcome some of the problems that the Labour Force Survey faces: attrition and changes in survey design. First, unemployed workers sometimes do not reply to two consecutive interviews, resulting in underestimated unemployment-to-unemployment flows. This affects exit rates from unemployment. Second, there was a major change in the survey in 2005 which did not affected stocks, but some flows were severely affected. The MCVL can help to distinguish which jumps correspond to true events and which are artificially created as a result of the redesign of the survey.

The contribution of this paper is twofold: First, it extends the methodology of García Pérez (2008) in adapting the administrative dataset in order make it useful for research. I provide further systematic guidance using the institutional setting and the LFS as a benchmark. The Appendix provides a detailed guide on how to do this. Second, it shows how administrative data can improve the labour force survey on some of

[2]For the case of the US, Fujita and Moscarini (2017) consider this as unemployment too.

its empirical challenges.

The rest of the paper is structured as follows. Section 2 describes the two datasets, their advantages and disadvantages and explains the procedure to build a quarterly panel from the administrative data; Section 3 explains the different expansions to the administrative data, checking their resulting unemployment rates to the Labour Force Survey official unemployment rate; Section 4 provides some further robustness checks; Section 5 shows how the expanded administrative data can help to evaluate and interpret two problems of the LFS; Section 5 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 thereafter), elaborated by the National Statistics Institute (INE in its Spanish acronym), and the Continuous Sample of Woking Lives (MCVL), provided by the Spanish Social Security. It offers a comparison between the two and briefly explains how to structure the latter as a quarterly panel.

2.1 Description

Official unemployment statistics come from the Spanish Labour Force Survey (LFS thereafter) which follows a representative sample over 100,000 people for six consecutive quarters. The sample size is about 100,000 observations per quarter from 1987 to 2013, weighted to account for true population numbers. I will use these weights when reporting stocks, as it corrects for the sampling errors - some groups are over-represented and some under-represented. As other labour force surveys, it classifies workers by asking them to report their activities in the week of the interview - if they were employed or if they searched for a job, for example.[3] This allows the LFS to observe whether a worker is out of the labour force and why.[4] The LFS is thus structured as a panel by design.

However, many participants do not reply all of the six quarters that they are interviewed, which leads to problems when calculating stocks and more so when building

[3]To classify labour market status of the population I use the variable *Type of Contract* for employees, *Current working situation* for the self-employed and the variable *AOI* for unemployed and out of the labour force individuals. This last variable encompasses the answer 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?"). The main advantage of using this variable is that it is the one used for official unemployment rate series, which are then used in the EUROSTAT and OECD

[4]In particular, 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 they are not going to be able to find a job", "caring for others" etc.

labour market transitions. These problems are partly corrected by adjusting the weights, which results in consistent stocks over time. Another complication arises from the changes the survey itself has undertaken over the years. In particular, two major changes in 2001 and 2005 affected how unemployment was recorded and produced breaks in some series.[5] These changes do not affect, by design, the stocks, but do alter labour market flows. In Section 4 I will explore these issues in more detail.

The Spanish Continuous Working Life Sample (MCVL thereafter) comprises the working histories of a 4% sample of the working population for the years 2004-2013. Similar datasets exist for Germany, Italy, Austria and the US, among other countries (see Tattara and Valentini (2010) for a table summary). The MCVL stands out as being very accessible and big - there are more than 20 million observations in total as of 2013. It uniquely identifies workers and firms, allowing to observe job-to-job transitions and distinguish quits and lay-offs. It also keeps track of self-employment, something that is excluded in other datasets. The firm and worker identifiers allow to link the working histories panel to a yearly Income Tax complement, containing fiscal information about wages, other retributions in kind, unemployment or disability benefits, severance payments, and any flow of income between the firm and the worker (or the Social Security and the worker). For self-employed workers and firm CEOs, it contains declared profits, and although that information is highly susceptible to misreporting for tax avoidance purposes, it nevertheless provides an insight into self-employed earnings. These characteristics could in principle allow the dataset to be treated as a matched employer-employee data. However, as the unit of measurement is the worker and not the firm, it is very unlikely that we observe all of the workforce from the firms in the data.[6] Alternatively, it also contains a file that details the taxable income received by the worker in all of their previous spells. This is not the same as gross wages from the tax data, but as shown in Bonhomme and Hospido (2017) it can be adapted to study wage and earnings dynamics.

The raw MCVL is not ready to use because it is not designed for the purpose of research, but for administrative bookkeeping. There are some academic articles explaining how to clean and format the data (see Lapuerta (2010) or García Pérez (2008), for example). In particular García Pérez (2008) provides a comprehensive identification for the main problems when dealing with overlapping employment spells and censored unemployment spells. After implementing most of the cleaning and formatting procedures, there is

[5]The 2001 reform added the requirement for unemployed workers to be available for work in the next two weeks. This change caused a shift in the stock of unemployed in 2001. But the major change came in 2005, when the sample was altered to reflect the growing impact of migrant workers and an electronic way of carrying the survey from quarter was introduced.

[6]It could be argued that the sampling of firms is representative of the universe of firms in Spain, as the sample is representative of the worker side, it thus should be representative of firms too. As self employment is represented too, the coverage of small firms is good but few large firms are represented.

still the question of how to handle unemployed workers who are not registered within the Social Security. These periods appear as blanks, gaps between observed spells. This is a common feature to other administrative datasets, but in Spain this issue is eespecially 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 use the retrospective information of the MCVL to build a panel earlier than 2004, as we have information on the complete working histories of workers, dating as far back as the 1960's. However, García Pérez (2008) warns against doing this kind of inference as the sample is representative of the year that it is collected from. That is, the 2005 file is representative of the working population of Spain in 2005. In the next years the sample adds new spells to adapt to demographic changes, but it does not "drop" any worker. The cases of workers dropping are either migration, transitioning out of the labour force or deceased.[7] Using the 2005 to get any inference on the labour market in 2000 would cause some relatively minor representativeness problems, as there was a substantial influx of migrants in the 2000-2005 period. But using the MCVL to look at the 1992 recessions would over-represent younger workers as the average age falls considerably. The sample size of the MCVL increased considerably in 2005 to account for better representation of different groups, so the MCVL of 2004 is not very well suited for study. in the For this reason, I follow García Pérez (2008) and only use the year 2005 onwards when building stocks, to make it comparable to the LFS. I will use the 2005 file to account for flows in 2004 as the sampling error of a year is not too significant. Finally, the potential accuracy gains that can derive from using only the final wave (2013) can outweigh the problems with representativeness for some applications. One of these exceptions is unemployment duration, as using the last year only provides with higher accuracy. Using instead all of the waves can result in more overlapping spells and discontinuities.

The main points of the description of the two datasets can be summarised in table 1 below.

2.2 Prepare the MCVL for use

In order to work with the MCVL to analyse job market variables it is necessary to at least stablish a reference variable for labour market status and treat some simultaneous spells. If we also want to make meaningful comparisons with the LFS, we need to build a panel with one observation per quarter. Laborda (2013) for example uses this approach.

[7] In this last case, the date is recorded in the MCVL.

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 since 1987.	A sample of about 1,000,000 job records of people with any sort of affiliation with the Social Security. It can constitute a panel since it follows most of the same people over the period 2004-2012.
Advantages	* Detailed and accurate information for personal characteristics (such as education). * 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)	*Firm and worker identifiers allow for the study of job-to-job transitions (rarely available in the LFS). * Very accurate information on employment spells, with precise dates of entry and exit into and out of jobs/unemployment * Can be matched to a fiscal dataset for wage/benefit information. * Consistent through time.
Disadvantages	* Fails to capture short term jobs and some very short unemployment spells due to it being a *quarterly* dataset. * There are important series breaks in 1992, 1999, 2001 and 2005.	* It can't track anyone who has no formal relationship with the Social Security. As such, it serves poorly for tracking people out of the labour force. * For the same reason, it is also unable to track informal market activities.

The Appendix provided with a detailed guide that allows to interpret the accompanying Stata files to this paper. Here I explain the main steps of the procedure to make the MCVL ready for use in research:

1. Classify the labour status of the worker in each spell
2. Modify overlapping spells and extend censored spells
3. Build a panel by selecting one spells per period of time

Labour Status of the worker

The aim of the first step is to create a variable that classifies workers in four categories: *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 90% of all job creation[8] and 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 someone is participating or not. In order to keep their benefits, unemployed workers are formally required to: prove they are actively searching for a job, attend job interviews and to not reject job offers. The Employment Centre monitors workers at least each month upon receiving the payments. A priori, I treat this an sufficient prove of unemployment.[9] Retired workers are not part of the main sample, and their information is in another linkable dataset. They constitute, according to García Pérez, 27% of the total number of observations in 2007, which is far from the 40% of inactive workers that the labour force survey reports for that year. To this is necessary to add population that is too young or have never participated in the labour force. As these groups are excluded from the sample, the remaining option is to count periods in which the Social Security has no information as out-of the labour force. This is the initial treatment I give them, as in previous studies. Later on I will relax this assumption.

Three variables contain all the information needed to classify workers in the same working status as in the LFS:

[8]Temporary contracts who expire and are not renewed do not incur into dismissal transfers, which is the case of termination of open-ended contracts. These contracts always end in dismissal, quit or retirement.

[9]This may not be the case if monitoring fails, for example if an unemployed worker that lies about searching for a job. This is acknowledged by most authors (see for example García Pérez (2008), Lapuerta (2010) and Arranz et al. (2011)). García-Pérez points out that most of these dropouts from the sample correspond to women and young people.

1. *Type of Labour Relationship* (TRL) codes the different links each worker has with the social security – working, receiving unemployment benefits. This way I separate unemployed workers.

2. *Contract Type* contains the code for each type labour contract.[10] There are 557 kinds of different contracts in the registry, but most of them are "legacy contracts" that do not exist in the present. Most temporary contracts are grouped under the 400s numerical codes while regular contracts are coded in the 100s. This way I distinguish between temporary (which have an specific termination date) or permanent (open-ended) contracts.[11]

3. *Contribution Class* allows for the identification of self-employed workers, as they have a different arrangement with the Social Security. These correspond to variable values 500-600.[12]

Using those three variables suffices to classify most observations, but there are special cases: the unemployed close to retirement that choose to pay their contributions to the social security as if they were employed to boost their pension, discontinuous and seasonal workers who get a compensation in between working seasons or students that receive benefits under apprenticeship contracts. The treatment for these specific cases is left for the appendix.

Clean overlaps and extend unemployment spells

After classifying the state of the worker, it can still be the case that a worker is classified in different status at the same time. García Pérez (2008) recommends to drop overlapping spells where the worker is simultaneously employed in more than one firm[13] or when the beginning and the end of the spell overlap for a few days. It can be the cases where employed or unemployed individuals are receiving some form of complementary benefit during their current spell. For example, because of an incapacitating illness or a widowhood pension. In these cases the best approach is to keep the employment spell only, and

[10]There are some kinds of contract that don't exist any more - usually contracts with some kind of temporary subsidy created in the 1990's. These are not relevant for the present study as I focus in the 2005-2013 period.

[11]In the particular case of discontinuous workers (those who work only on specific periods of time every year) I treat them as permanent, as they are subject to firing costs and have no pre established termination date.

[12]There are some especial categories for domestic workers, agriculture workers, farmers and sailors. I select those who are self-employed in these special regimes.

[13]These cases are rare in Spain and mostly refer to part-time jobs. However some employees of the Catholic Church are recorded to have two full-time jobs: one for their religious duties and one for their other dealings - teaching for example.

to merge some of these overlapping spells in unemployment.[14]

The second step can be omitted if the researcher wants to use the MCVL definition of unemployment. García Pérez (2008) recommends to add to unemployment between two employment spells the missing days before and after the unemployment spell if these are not recorded, especially in the case of days between the end of the job and the beginning of the unemployment spell. These are likely due to administrative delays and should be added. In Section 3 I analyse the different expansions that can be done to unemployment spells in more detail.

Building the panel

Once each worker has a unique observation for each quarter, we can proceed to build a panel by only keeping the observations that relate to each quarter of interest. The easiest approach would be to take the 2013 file and use its retrospective information. However, as each year file is representative for the population in that particular year, using the 2013 retrospectively will lead to a bias in favour of younger workers in previous years, as discussed before. The second main reason is that is more practical for handling the data: to mimic the LFS we would need to sample a spell that lasts beyond the year as separate, different spells. That is, year frequencies allow to build an observation per year per worker, with details of the states in the different quarters. In this way, each worker ends with an entry for each year she is in the sample. Using a unique file would require the use of many more auxiliary variables.

To build the panel I first establish a point in time at which I will evaluate people's working states: the two weeks starting 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. The labour status variable will reflect the relationship each individual have on those weeks: self-employed, open-ended employee, temporary employee or unemployed.[15] Laborda (2013) uses the fifth week or each quarter, as the LFS takes place around that date. Most job contracts tend to start on the first day of the month, so the first weeks of the quarter seem a natural choice. For workers that have more than one spell in the same two-week period, I give priority to the longest spell. That is, if a worker starts the two-week observation window unemployed but end with a job that lasts for one

[14]The Social Security records pensions separately, but the pension file can be easily merged with the main working records file using the individual identifier.

[15]This corrects for sort 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 Christmas holidays in Spain ending on the 6th.

more quarter, I count her as employed on that quarter. If there is a tie (0.03% of the total number of observations) I chose employment over unemployment, and self-employment over employment. This is because some jobs (especially part-time) can be complemented with unemployment benefits, but that doesn't mean the worker is unemployed.

3 Methodological Check

The challenge is thus how to treat unemployment spells in the MCVL. In the second step of the procedure outline in Section 2.2, the researcher needs to take a stance on what she's going to consider as unemployment, or alternatively as in Alvarez et al. (2015) consider all the gaps between employment spells as non-employment. In this section I argue that some extensions allows us to consider most of these gaps as unemployment spells, providing a series of methodological checks against the LFS unemployment series. The motivation of these expansions is to match the unemployment rates from both datasets, which differ considerably after 2008.

3.1 The unemployment gap

The LFS and the MCVL have a different number of observations (an average of 108,136 in the LFS[16] and 678,183 in the MCVL) so in order compare the stocks I express them as shares of the labour force thereafter.

Figure 2 shows the main discrepancy between the MCVL and the LFS: the unemployment rate. This disparity reaches 10 percentage points by the second quarter of 2013. The differences persist when unemployment rates are compared by age and sex. For young workers between 20 and 30 years of age, the MCVL unemployment is half of the LFS.

The main source of the differences comes from the different definitions of unemployment they use:

- The LFS considers a person unemployed if: (1) they are not currently employed (2) 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.

[16]The weighted labour force survey has a mean of 31,360,266 observations per period – which amounts to the total population of Spain.

Figure 1: Unemployment rate in Spain

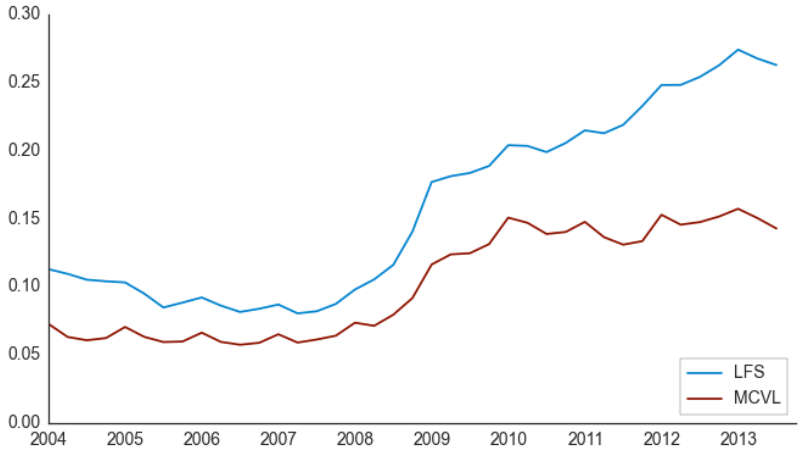


Figure 2: Unemployment rate by sex

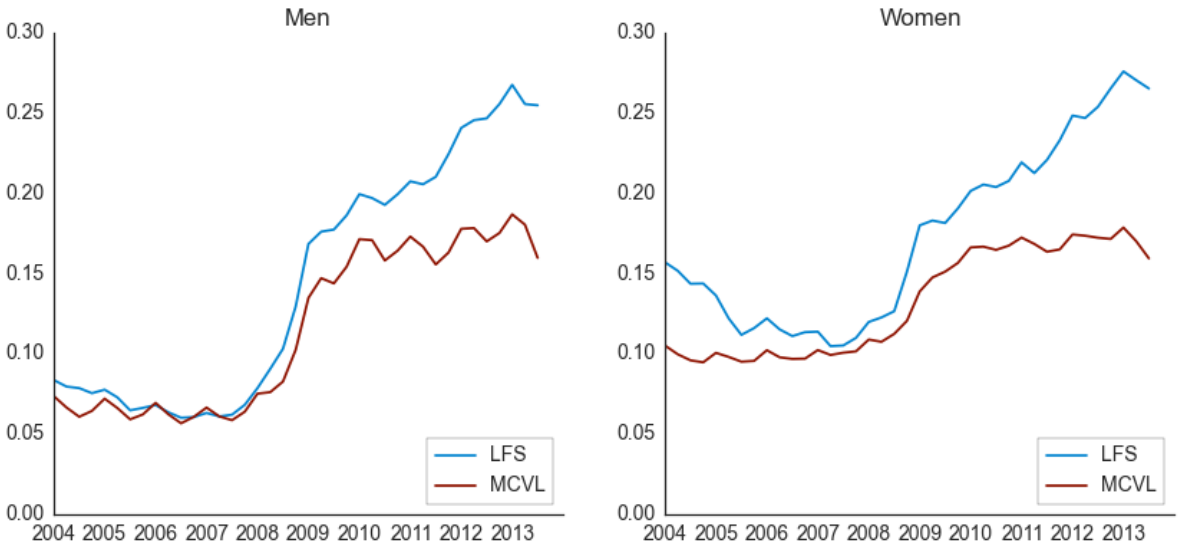
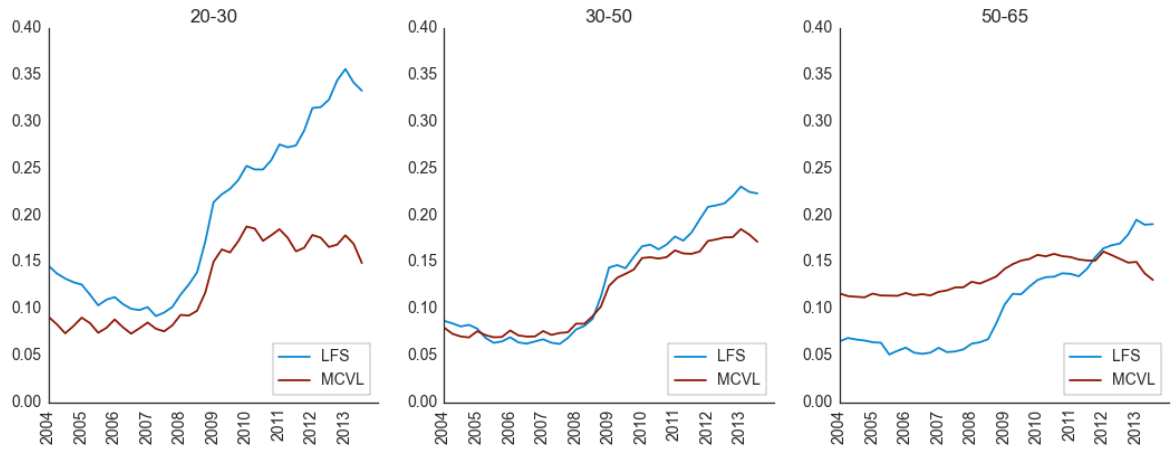


Figure 3: Unemployment rate by age



This means the MCVL excludes all unemployed whose benefits have expired. The Spanish Social Security does not provide any other benefit for those who exhaust their unemployment compensation, so all unemployed beyond 2 years[17] *disappear* from the registry. As long term unemployment reached 50% of total unemployment by the end of 2013, this 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 the end of the sample.

The Social Security also excludes all without the right to claim unemployment compensation (who have been employed for less than a year in the last 4 years) and unemployed that have not held a job yet. The second expansion aims to recover these spells (usually related to young workers in short lived temporary contracts) by adding gaps between employment spells of those without the right to claim.

Finally the Social Security may be counting as unemployed some inactive workers by the definition of the LFS (not actively searching for a job and/or not ready to work in the next 15 days). This would imply the MCVL has a bias upwards with respect to the LFS. But this is not what we observe in the data, except for older workers.[18]

In all that follows my consideration of unemployment is the same as in the LFS (standard ILO definition). This allows me to compare the unemployment rates from both datasets as both measuring the same phenomena. Therefore the differences in between the two series will be interpreted as the unemployment *within the ILO definition* that is accounted for in administrative data. For example, frictional unemployment that is not captured by the LFS would be unemployment by the ILO definition, but because of the timing and structure of the survey is not recorded. The marginally attached (those who are not employed nor actively searching at a point in time yet end their non-employment spell in a job) would not be unemployed by that definition, and so ideally be excluded from the MCVL. However we cannot exclude these cases from the MCVL without some imputation. This would require to build a measure of "propensity to be marginally attached" with observables in both the LFS and the MCVL. I do not follow this approach for three reasons: first, the set of observables in common between the two datasets is small[19] so the propensity score will be noisy; second, there is noise in the LFS measure-

[17]This limit is expanded to an absolute maximum of 4 years for workers with family obligations or other special circumstances

[18]It is possible to observe when an individual is receiving unemployment benefits immediately before retirement in the MCVL. I consider these workers unlikely to engage in active search. In order to keep consistency with the LFS I do not count these observations as unemployment.

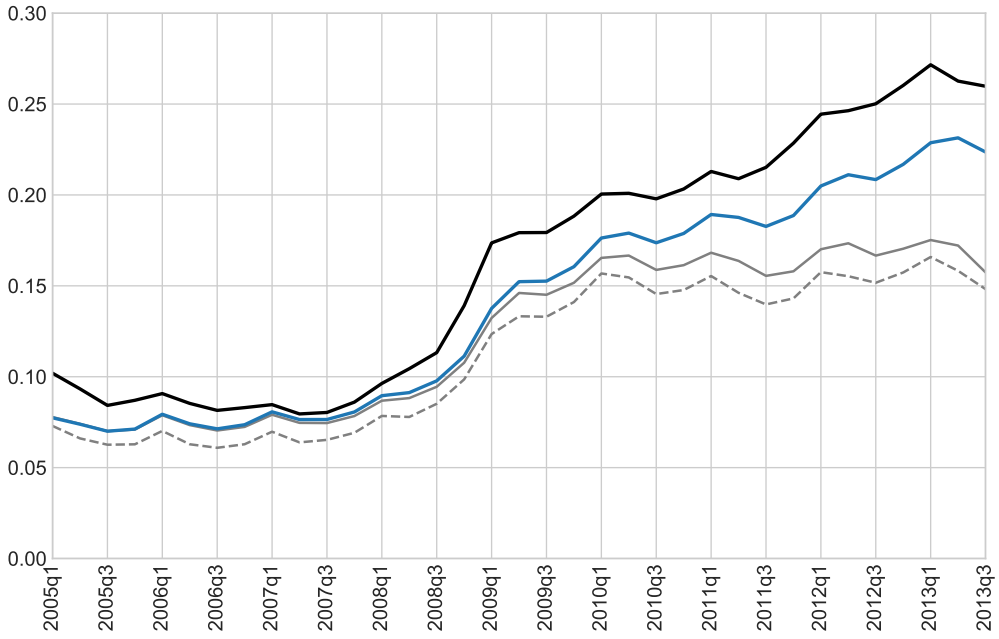
[19]Some key variables, like education, occupation and sector are defined differently in both datasets or at a large aggregation level (industry is coded at the top level in the LFS for example).

ment of the marginally attached;[20] third, although for comparison I choose to follow the ILO definition, most labour economists do consider the marginally attached as unemployment. Therefore I find that trimming the marginally attached from the administrative data not worthwhile for the purpose of this paper.

3.2 Closing the unemployment gap: LTU expansion

Given the importance of long term unemployment in the last years in Spain, it is natural to start by adding the days elapsed between the recorded end of the unemployment spell and the start of the next job. This is already suggested by García Pérez (2008) as a ncesary treatment to work with the MCVL. This extension is easy to implement but has a shortcoming: many of the long term unemployed have not found a job by the end of the sample. This is reflected on the small difference between *Original* and *Only finished spells* unemployment rates in figure 4: the difference in trend and level of unemployment widens from 2009 onwards. In fact, comparing it with the original MCVL series, it barely makes a difference.

Figure 4: Unemployment rates - LFS and expanded Social Security



The *LTU Expansion* adds all the unfinished unemployment spells by the end of 2013, as well as extending the duration of registered unemployment spells between jobs as be-

[20] This is true even after de-UNU-fying the LFS as in Elsby et al. (2015).

fore. After this expansion both trend an level are very close to the LFS, as shown in figure 4. The expanded series is still below the LFS by 3.7-2.5 percentage points by 2013. This shows us that if we want to match the trend of the LFS unemployment rate, we need to add unfinished spells. How many depends on the sample size. 2013 was the worst year of the recession in Spain, so many unfinished spells are to be expected. But if a researcher has access to further years, most of the difference may be captured by adding the finished unemployment spells only. I also condition the extension of unfinished unemployment spells on being at most of 2 years/[21]

3.3 Closing the unemployment gap: STU expansion

Other kind of unemployment the MCVL is missing is people without the right to claim. This will be the case of:

- Quits to unemployment. Quits do not have the right to unemployment compensation.
- New entrants to the labour market (with no previous employment record)
- Temporary workers with employment spells below the minimum requirement - less than a year of tenure in the last 4 years.
- Self-employed workers who have no right to unemployment compensation.

Notice that all of these cases are not the main source of discrepancy between the LFS and the Social Security, as the the first expansion is already very close to the LFS. Underestimating long term unemployment a bigger issue, but capturing short term spells is important for matching youth unemployment rates. Therefore I refer to the resulting expanded unemployment series the *Short Term Unemployment* (STU) expansion.

To identify these spells, I chose to include all gaps between employment spells that last at least 15 days and at least one of the following conditions:

- The worker quitted her last job.
- The worker was self-employed in her last spell.
- The worker does not have enough contribution periods to be eligible.[22] [23])

[21]Robustness checks can be used to asses if the 2 year maximum duration is a good threshold. In the LFS X% of all unemployment spells are beyond 4 years and are not receiving UI.

[22]The threshold is less than 360 days of employment, according to Spanish legislation.

[23]It is worth noting that the law in Spain does not allow to claim benefits that were not consume in

Figure 5: Age Distribution by Expansion

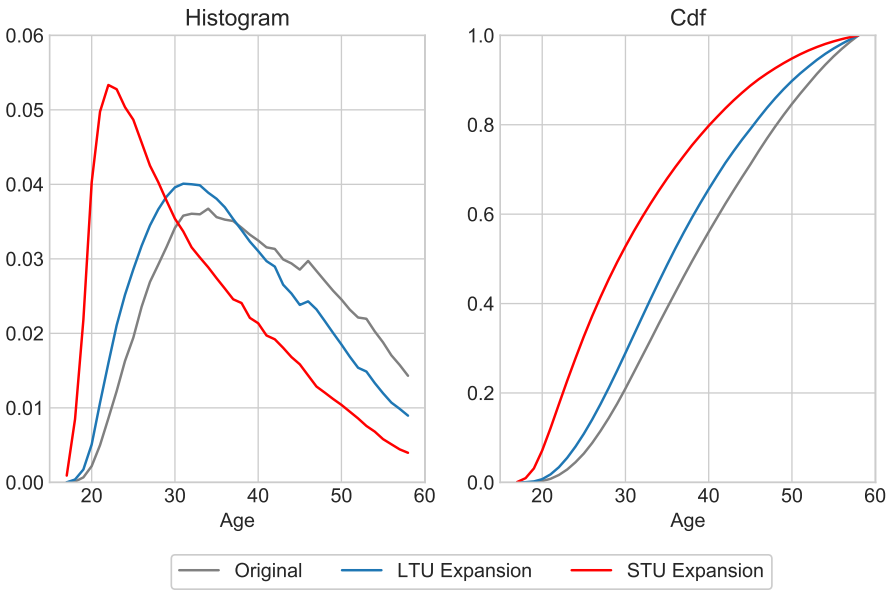
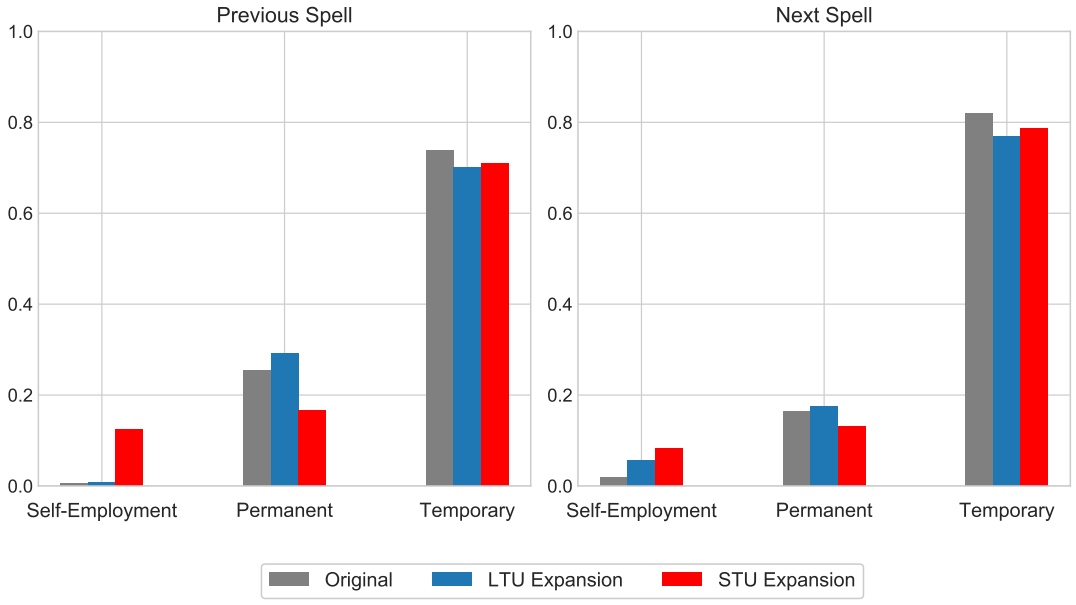


Table 2: STU Expansion spells, by type

	Quit	Self-employment	No right to UI
Total	151,461	91,972	551,272
Percentage	19.06	11.57	69.37

Figure 6: Spells before and After Unemployment



In addition to these restrictions I add the requirement that the worker is not to be recalled to work on the same firm. The reason for this is that there is a good chance that the worker knew that she was going to be called back and thus had no incentives to search. This is particularly important because employers could use these tactic to extend the maximum duration of temporary contracts. That is, instead of renewing the worker beyond the two month period, the firm asks the worker to take a leave and then return.

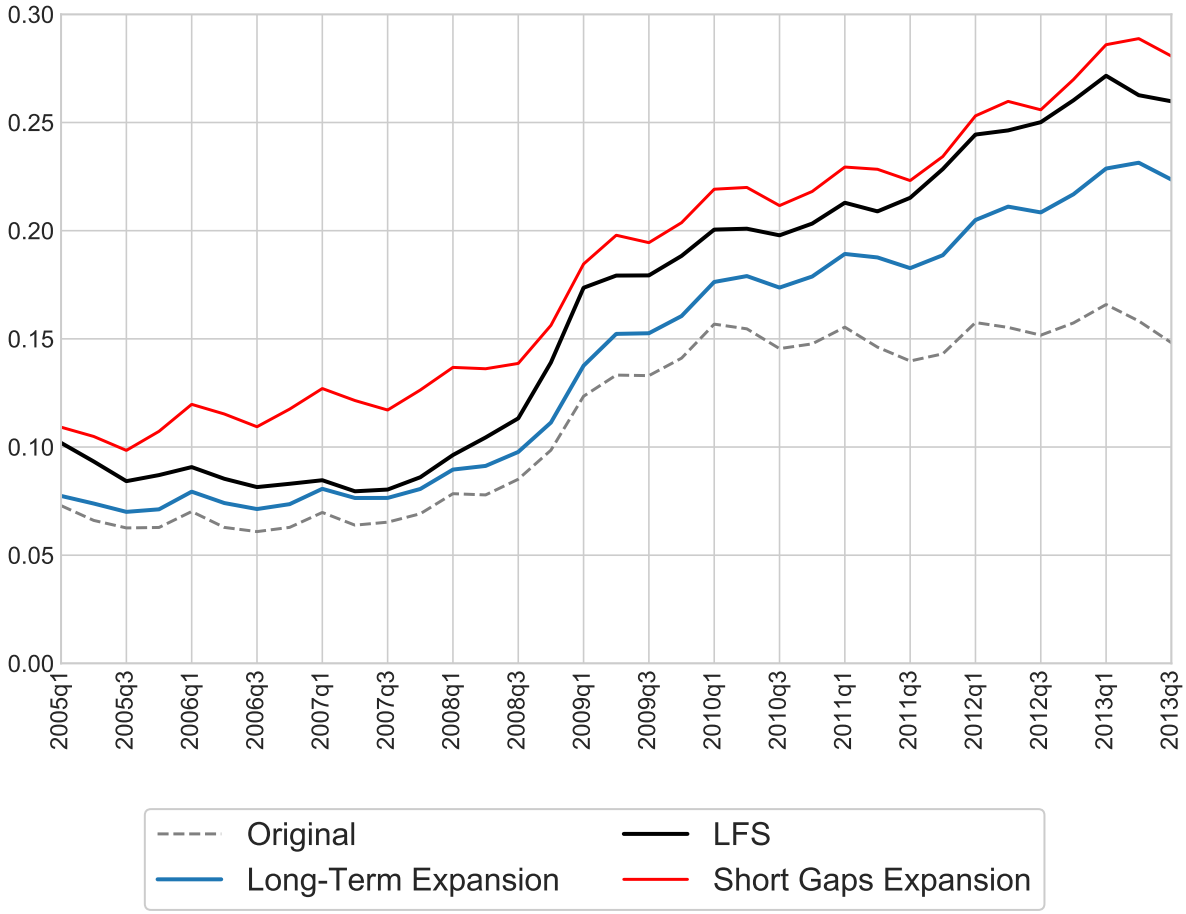
The conditions imposed make it very unlikely that a person detached from the labour market would qualified for an unemployment extension. Table 2 shows the different spells that are added in the STU expansion by each case. Most of them are coming from not having the right to claim unemployment benefits, but a non-trivial amount also come from self-employment and quits. Figure 5 shows the histogram and empirical CDF of the ages of the unemployed at the time of the start of their unemployment spell, broken down by expansion. Unemployed individuals from the STU expansion are overwhelmingly younger, with 80% of them under 40. In terms of what was their previous spell, figure 6 shows that they are more likely to come from temporary contracts than in the LTU expansion, and their next spell is also more likely to be another temporary job. This gap increases if we consider that virtually all unemployed workers coming from self-employment are in the STU expansion. If we exclude self-employment, 86% of all previous spells in the STU addition are temporary jobs, while the LTU and the original only have 70% and 74% respectively. Appendix A.2 analyses these spells in more detail.

After adding these spells, the MCVL unemployment rate gets closer to the LFS after 2009, as figure 7 shows. There is still an overestimation of unemployment before 2009, but as for the end of 2013 the differences are small. It is not surprising that the STU expansion adds more unemployment relative to the Long-Term expansion in the 2005-2008 period. These years coincide with the construction boom and the highest rates of temporary contracts over total employment. In the following years the gap is reduced as long term unemployment increases its incidence. The trend is similar to the Long-Term 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 (figure 8) and age (figure 9). By gender, the Short-Gaps expansion brings the MCVL closer to the LFS. This is expected as women are more commonly employed in the services sector, where temporary contracts are very common. It is less successful for men, in particular before 2008. This again can

the last unemployment spell if the worker wants the past employment spell to count for future benefits. For example, say a worker has 3 months left of UB, and finds a 6 month job; after the job ends, she can chose to claim the 3 missing months from before or the 2 months she has accumulated with the last job. She can't have both.

Figure 7: Unemployment rates - Short Gaps Expansion



relate to the use of temporary contracts among construction workers, but notice as well that even the original MCVL overestimates unemployment in this period. This suggests that men are less likely to report to be searching a job when claiming benefits, something we will see corresponds to what the LFS records in the robustness checks. The overestimation of unemployment persists even after onset of the recession.

By age things are also clear: the STU expansion helps reconcile the unemployment rates of younger workers, in a way the Long-Term expansion is not able to match. There is a small positive gap in the 2006-2008 years, again likely driven by males on temporary contracts. For middle-aged workers the differences mirror those in figure 7: slightly overstatement at the beginning and at the end. Here the Long Term expansion performs arguably better. For older workers the STU expansion barely makes a difference over the Long-Term one. Still both offer a more coherent picture than the original MCVL series. Notice the overestimation of unemployment even for the original MCVL in the 2005-2008 period. This strongly suggests monitoring problems for this segment of the workforce. However, given the small share of older workers on total unemployment for those years they are not likely going to drive a big difference in the overall unemployment rate. Both

Figure 8: Unemployment rates by gender

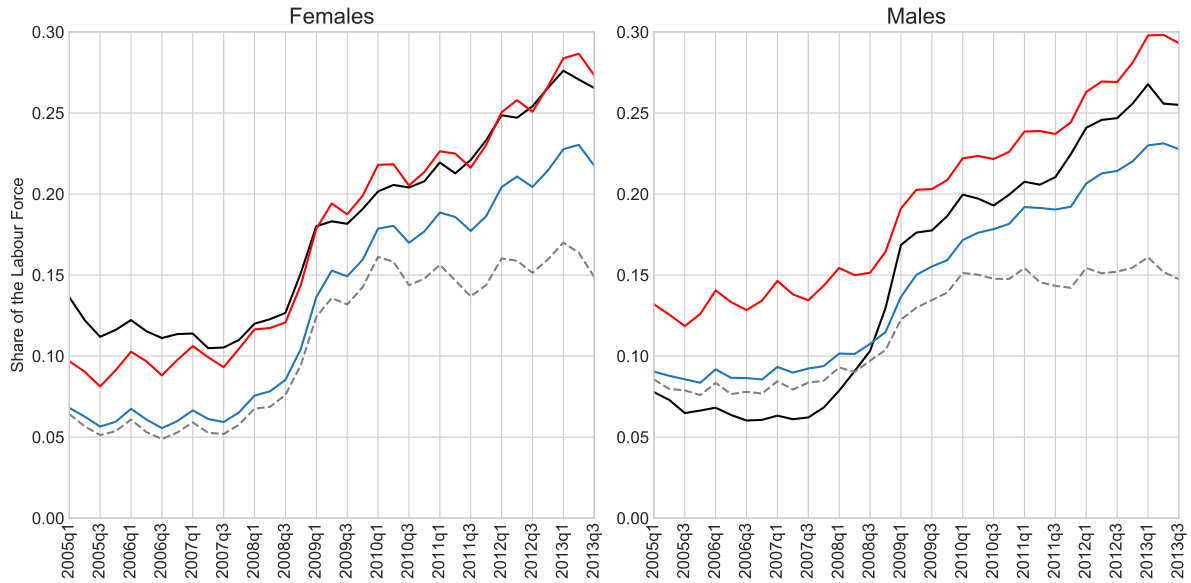
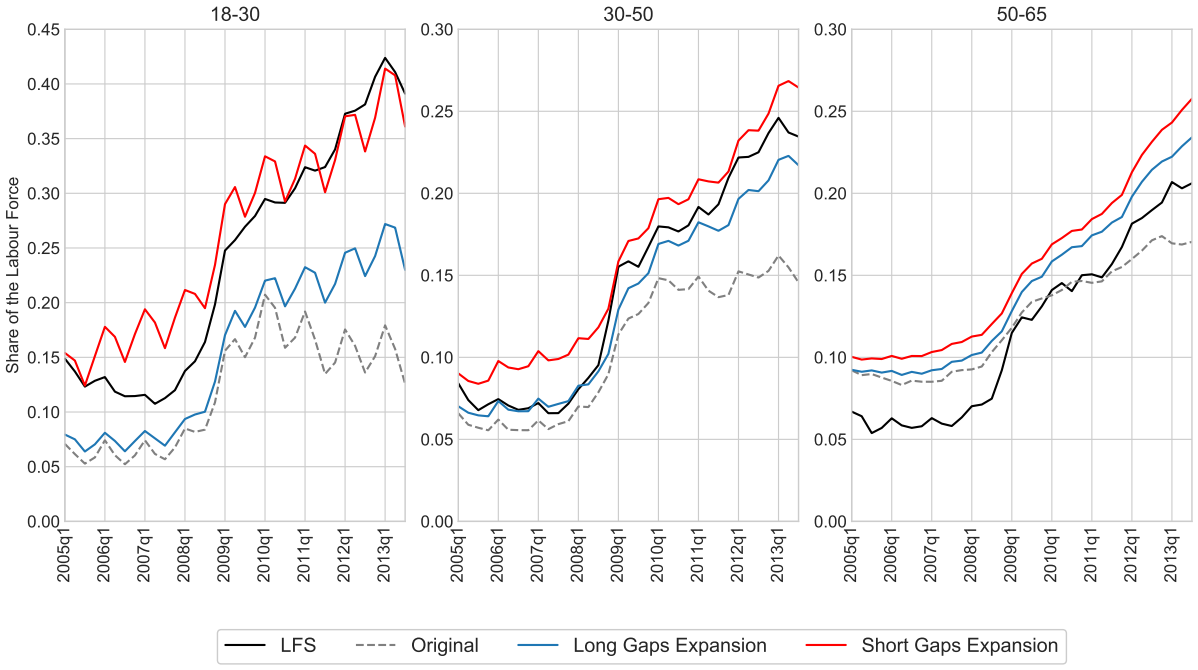


Figure 9: Unemployment rates by age group



expansion improve on the trend of the original, especially for middle-age workers and the Long-Term expansion.

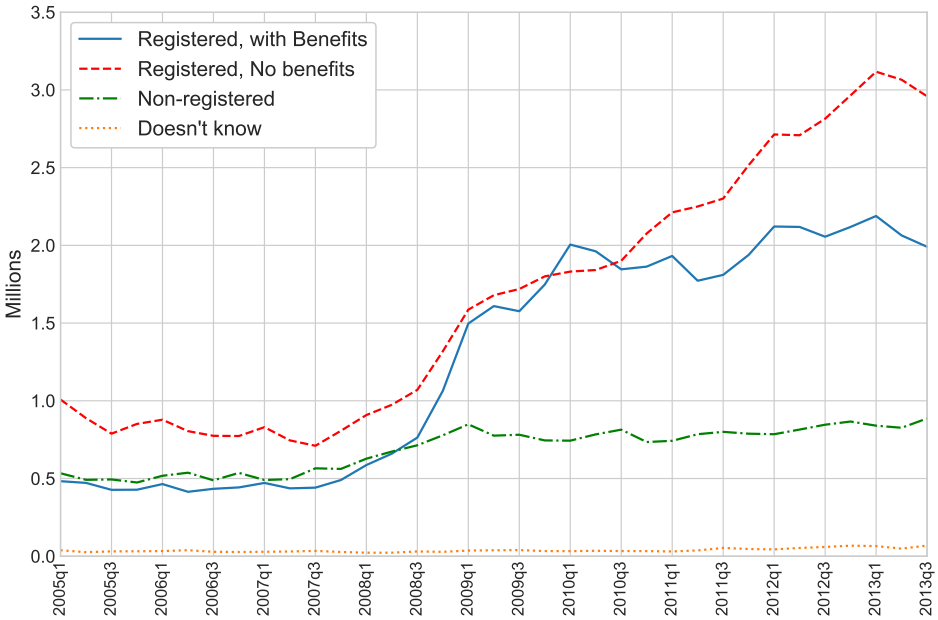
4 Further Robustness checks

In the previous section I created different unemployment series using only the information on the MCVL and institutional setting, then checked the resulting stocks against the LFS. This section provides a further check comparing the extent of self-reported unemployment without benefits in the LFS to the expansions of the MCVL. This check is interesting because it shows that as it was possible to go from the MCVL to LFS definition of unemployment, the reverse path can also be traced up to a certain extent. The methodological differences between the two datasets mean that the match is not perfect, but nevertheless very close.

The fact that the LTU expansion managed to bring the MCVL closer to the LFS was derived directly from the definition of unemployment in the MCVL. But we can use the information in the LFS to check if there is also any clear trend on benefit expiration that would explain why the LTU addition gets closer to data. This check can be done by looking at a variable in the LFS called "Relationship with the Employment Office", that asks workers whether if they are registered as unemployed and if they are receiving benefits. This question is answered by all respondents, as some workers out of the labour force may be receiving some benefit, such as pensions, illness of incapacity. This variable thus classifies workers in four categories, 'Registered, with Benefits', 'Registered, No Benefits', 'Non-registered' and 'Doesn't know'. The first case will be recorded in the MCVL, while the second won't. These unemployed without benefits are the ones that the LTU correction is targeting. The third case corresponds to job seekers that are not registered with the public Employment Office. These cases will not be recorded Social Security, and the the ones the STU seeks to recover, at least in part.

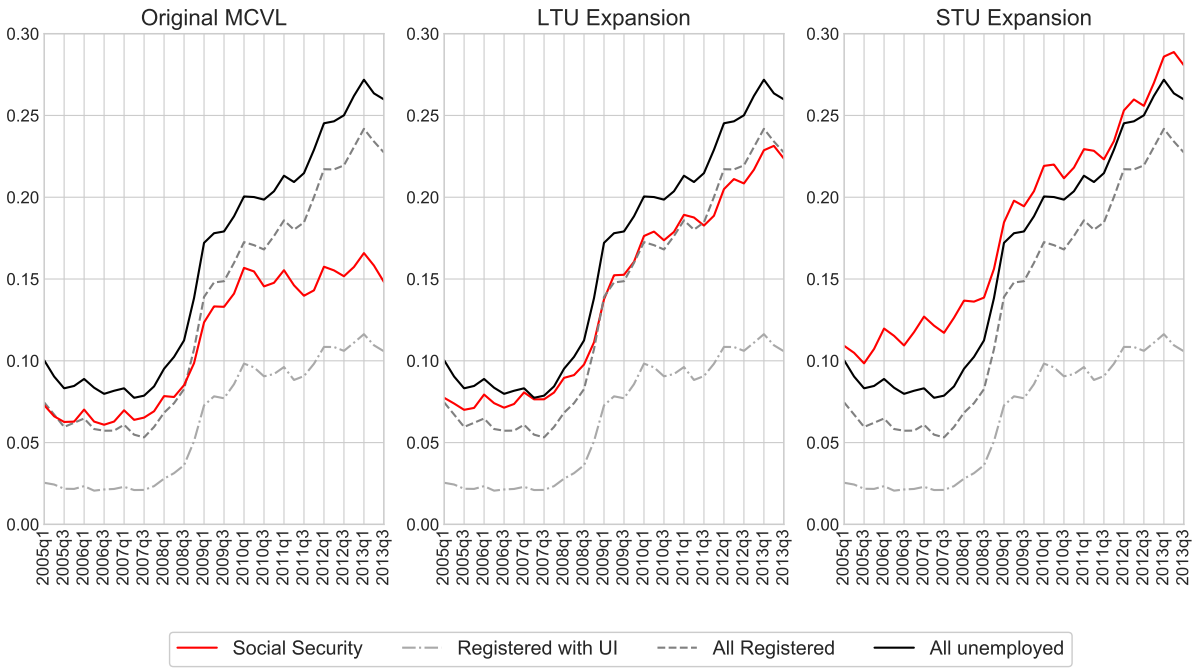
Figure 10 shows the evolution of these stocks (in millions) in the 2005-2013 period. The *'Registered, with Benefits'* line looks very similar to the original MCVL series, which highlights again that these are the only unemployed captured in it. The stock of *'Registered, No Benefits'* on the other hand looks more similar to the expanded series, with a big increase after 2008. The 'Non-registered, No Benefits' stocks does not change substantially in the period, increasing slightly after 2008. These trends are reassuring, as they correspond to the different expansions.

Figure 10: Relationship with the Employment Office



Source: LFS

Figure 11: Alternative unemployment rate series



In figure 11 I use this variable and the labour stocks from the LFS to build alternative measurements of unemployment. The first panel shows the the original MCVL unemployment series (in red) has very similar trend to the LFS series were only the unemployed who are receiving benefits are considered. However, the MCVL is higher and corresponds to only the registered unemployed before 2005. This can indicate that most unemployed workers were receiving benefits before 2005, but not reporting it. The second panel shows that the LTU expansion matches closely with the case where only those registered in the Social Security (with or without benefits) are considered, in particular after 2008. This provides additional evidence that considering only those who are registered in the employment office is not enough to account for unemployment after 2008. The last panel shows how the STU expansion achieves a closer unemployment rate to the LFS when all unemployed are considered after the recession, but overestimates unemployment before. As discussed in section 3.3, this likely reflects that the MCVL captures more frictional unemployment than the LFS, which was more relevant before 2005 than after.

5 Using the MCVL to enrich the LFS

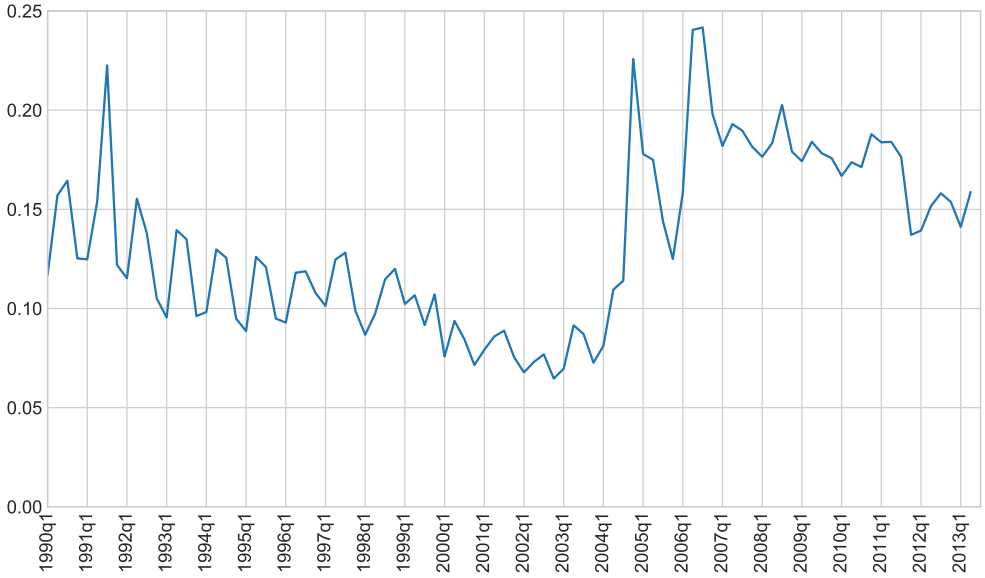
As the unemployment rate from the expanded MCVL is similar to the LFS, we can use this to compare other labour market magnitudes. In this section I will be focusing on labour market flows, who are problematic in the LFS for two reasons: (1) respondents not replying on consecutive interviews, which affects the flows and (2) changes in the design of the survey which changes workers labour market status, in particular in 2005. The aim of this section is to show how the MCVL can clarify these issues.

5.1 Attrition and Labour Market Flows

The LFS is a rotating, panel, such that each household is interview in 6 consecutive quarters. I define *attrition* as a respondent who is not in her last interview fails to report on the subsequent interview. The size of the attrition bias has not been constant over time nor affect all individuals the same. Figure 12 shows the share of respondents who are not in their last interview and report being unemployed any given quarter, but do not respond to the survey in the next quarter. 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 corrects for this problem by changing the weights of the observations each quarter and introducing more people in the sample. This makes stocks consistent over

Figure 12: Attrition of Unemployment Stocks in the LFS



time. However if we want to calculate labour market transitions the weights do not solve the problem. This problem is common among other labour force surveys, and usually resolved as in Silva and Vázquez-Grenno (2013) and Elsby et al. (2015): Taking the stocks as given, but the transition rates are biased because of attrition. Then we can calculate the transition rates that are consistent with the evolution of the stocks.

In all of these cases the stocks are given by the sum of flows in each quarter. For example, 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 equation 1 shows:[24]

$$\lambda^{XY}_{t,flows} = \frac{Y_{t+1}|X_t}{\sum_Y Y_{t+1}|X_t} \tag{1}$$

Assume that there is attrition in this data, but that it does not affect the transitions from X to Y, but the number of stayers $X_{t+1}|X_t$. Then the denominator would be *lower* than what it should be, as the non-respondents are taken out of the sample. Consider instead the the transition rate defined as in equation 2 below: 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.

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

In this way the transition rate would be consistent with the data. In practice, attrition can affect all of the rates out of state X, so the resulting bias of $\lambda^{XY}_{t,flows}$ is ambiguous. We can consider the case of $\lambda^{XY}_{t,stocks}$ as the extreme case when all of the attrition comes from stayers. That is, the non-respondents are not transiting to any other state in the next quarter. Figure 13 shows the evolution of $\lambda^{XY}_{t,flows}$ and $\lambda^{XY}_{t,stocks}$ from 1987 to 2013. There is not much difference between the two except in the flows between unemployment and temporary contracts. Here the gap is very noticeable in the 2005-2008 period, which coincides with the attrition 'jump' in figure 12. The gap is also noticeable for the temporary to unemployment (TU) rate after 2008.

The MCVL does not suffer from this bias, as we can observe with more precision and up to daily frequency the changes in labour status of workers. The definitions of unemployment are different in both, as discussed, but given that the expansions get them closer 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 due to attrition.

Figures 14 - 15 compare the flows resulting for the LFS to the MCVL. *LFS (flows)* shows the transition rates from the LFS calculated as in equation 1 (the denominator being the sum of transitions) while *LFS (stocks)* shows it as in equation 2 (the denominator being the stock).[25] The blue lines correspond to the LTU expansion of the MCVL and the red line to the LTU expansion. Given the increase in attrition of unemployed workers in 2005, I have taken back the MCVL to 2003 to have a larger window for comparison.[26]

In general, the level and trend of the flows is close between the two datasets. The MCVL series have both higher seasonal variation, which is due to the higher frequency of the data, that can capture short employment spells that the LFS can't, due to its quarterly structure. This leads to smoother series. Notice as well that the flows version of the LFS is always higher than the stocks version, which would be consistent with the non-respondents being unemployed the next quarter as well.

In the left panel of figure 14 shows that there are important differences between the stocks and flows version of the LFS in the 2004-2008 period. This could be indicative that

[25] 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.

[26] The observations from before 2005 are taken from the 2005 file, so there might be some representatively issues. I'm assuming these are not substantial for 2 years earlier in the sample.

Figure 13: Labour Market Flows in LFS

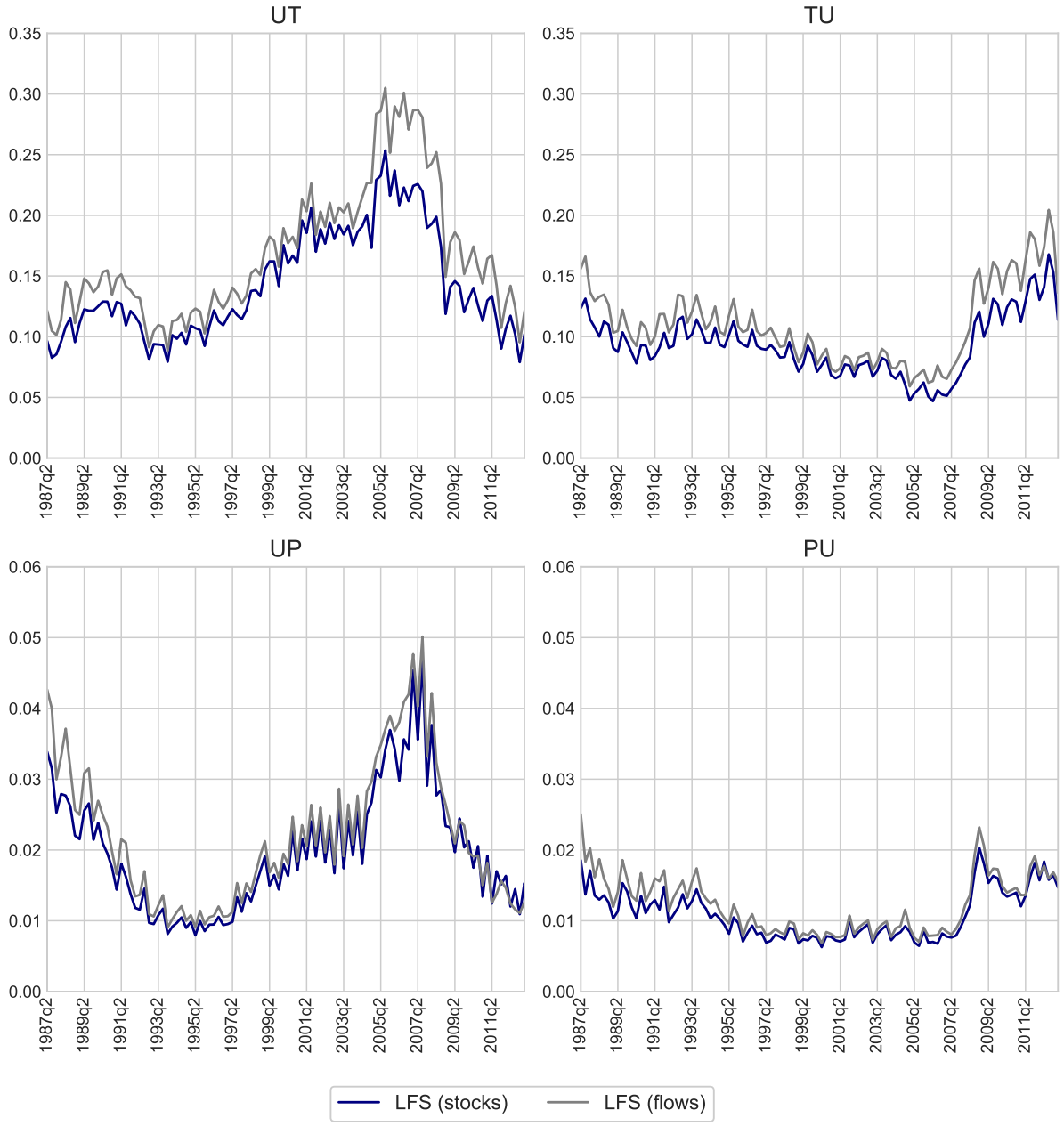


Figure 14: Flows out of unemployment

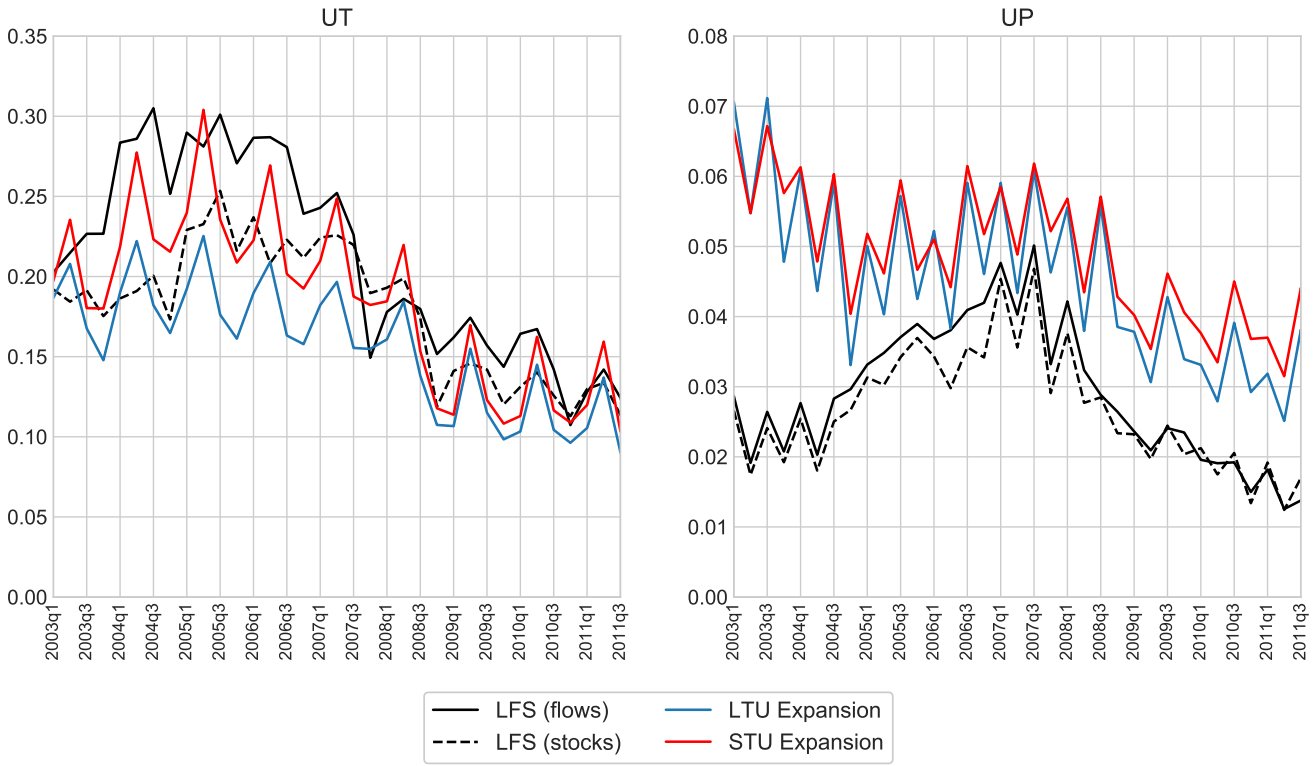
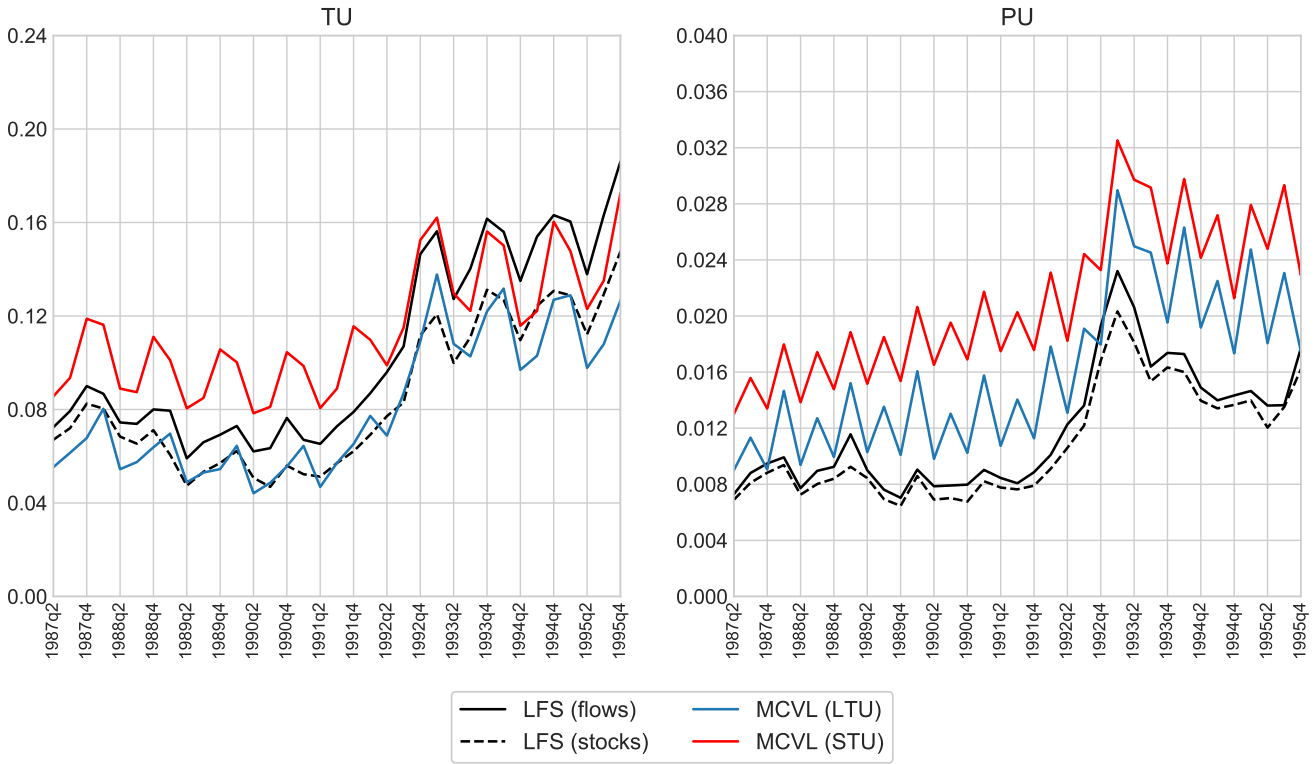


Figure 15: Flows into unemployment



the UT flow is not affected by attrition, but the denominator is. In this way, the non-respondents are not likely going to come from workers that take in temporary jobs. This discrepancy coincides with a higher discrepancy between the LTU and STU expansions of the MCVL. The differences between the expansions are clear: the STU expansion is capturing more movements into and out of temporary contracts, corresponding to more frictional unemployment. After 2008 however all series converge. This suggests that the attrition bias is partly driven by the unemployed without right to claim benefits flowing in and out of temporary contracts, which are not adequately captured in the LFS due to its frequency.

The right panel of figure 14 shows that the unemployment to permanent flows are higher in the MCVL, but the differences are small - notice that the scale is only from 0 to 8%. The divergence of the MCVL before 2004 can be partly explained by contract modifications adjustment not being marked before 2005 and it is a reminder that the data can't be take retrospectively without major problems. As for figure 15, the same conclusions carry over: the STU expansion is adding some short spells from temporary contracts that otherwise would be taken as job-to-job transition. The LFS does not capture well these quick changes and has a tendency to smooth them out, so both LFS series are below the STU expansion. This changes after the recession, as the volume of turnovers increases considerably. The LTU expansion matches quite closely the LFS (stocks) series.

In conclusion, the expanded MCVL is very close to the LFS, and helps explain that the attrition bias of the LFS is in part related to short-term unemployment spells coming from temporary contracts. Further research would have to confirm this, but the comparison with the MCVL points in this direction.

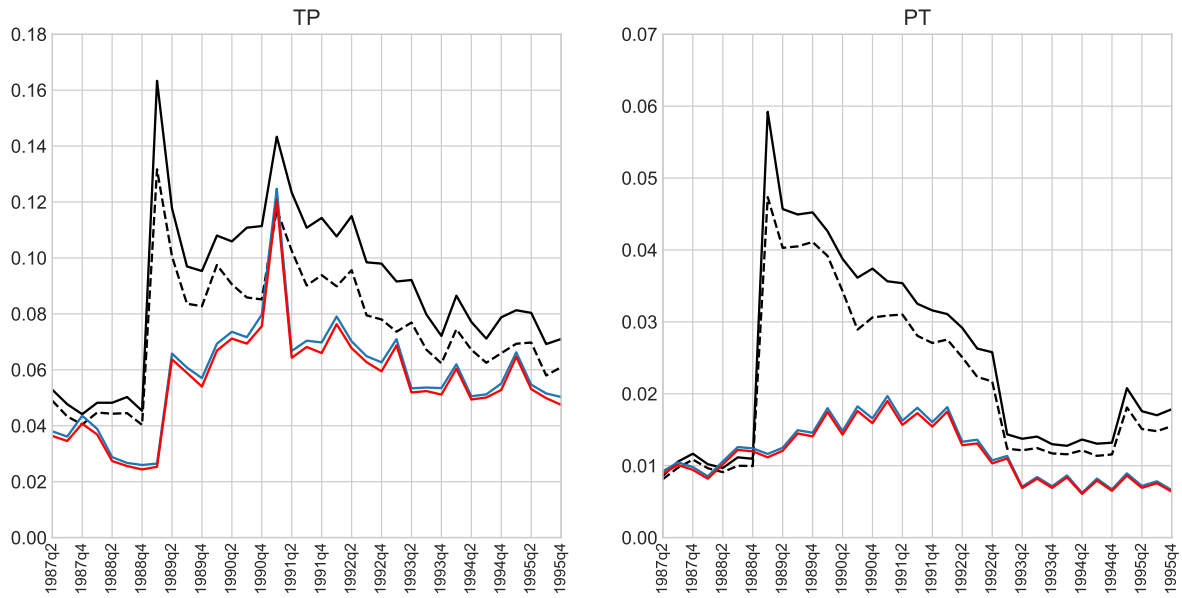
5.2 Changes in Survey Design

Attrition is not the only challenge when computing flows with the LFS: changes in the structure of the interview have also cause several discontinuities. These breaks are not present in stocks, because the National Institute of Statistics ensures that the stocks are consistent over time. Figure 16 shows one of the main breaks in the flows between different types of contracts (TP and PT).[27]

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

[27]Other flow rates that suffer breaks relate to inactive workers. But since the MCVL can't speak for them, then there is nothing administrative data can add to that question.

Figure 16: Quarterly Flows: Between contract types



Petrongolo (2007) for example). After the break it shoots up to 12% (16% following the flows calculation), almost a 200% increase. There is another spike in 2006 but it is explained by a labour market reform that happened at that time.[28] The MCVL flows, on the other hand, suffer no break in 2005, but it does have a peak at the same time as the LFS. This could be used as supporting evidence that the TP conversion rate increased after 2005, but at a smaller rate (close to 6%). The change in the LFS must be due to the change in the survey, that implied some spurious classification errors in order to get consistent stocks.

The right panel of figure 16 shows an even more stark case, where the permanent to temporary flow (PT) increases from 1% to 6% and slowly come back to previous levels. In the MCVL in contrast no such increase is resisted, only increasing to 1.7% before falling after the recession. In this way the MCVL helps to interpret the results coming from the LFS as coming for spurious transitions from the change in survey. The attrition problem of unemployed workers can be partly due to the same change.

[28] 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 is suggestive of firms using temporary contracts instead of permanent contracts because the former are cheaper. 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.

6 Conclusion

Administrative datasets are a great source of information for economists, but they also present some challenges. In this paper I analyse the case of the Spanish *Muestra Contiua de Vidas Laborales* (MCVL), a rich administrative dataset of working histories of a representative sample of the Spanish workforce. This dataset is rich in information, but in its original format it has important shortcomings how it records unemployment spells. In this paper I present a simple, systematic method to expand the original dataset by including two kinds of unemployment that the original MCVL struggles to cover: long term and short term unemployment.

Workers whose unemployment benefits expire 'drop' out of the sample, in that the days between benefit expiration and next employment spell are missing in the MCVL. What I label the LTU Expansion adds these missing days. Given the rise in long term unemployment in Spain after the Great Recession, failing to include the unemployed whose benefits have expired underestimates unemployment and presents spells that are artificially shorter.

Some workers do not have the right to claim unemployment benefits. Although this is not the case of most of the unemployed, it is common among workers with short employment tenures and always the case with self-employed workers and quits. By using the information in the MCVL, I identify these cases and add the gaps between employment spells that correspond to these. I show that doing this helps match the unemployment rate from the LFS after the recession, but it is above the LFS before. I argue that given the nature of these spells the LFS would struggle to capture them, as a quarterly survey its smooths out much of this frictional unemployment. I provide further robustness checks using the LFS variable coding the relationship with the employment office and taking a closer look to the added unemployment spells.

I then use the MCVL to improve on the LFS in two main aspects: attrition bias from unemployed individuals failing to respond two consecutive quarters and changes in the survey that generate spurious labour market flows. The flows from the MCVL closely match the ones from the LFS, and where there are noticeable differences (for example in the unemployment to temporary transition rate) they help understand the sources of bias of the LFS.

Applying the proposed extensions to the MCVL makes it ready to be used for research on unemployment. In the rest of this thesis I use this dataset for two such applications relating to long term unemployment.

References

Alvarez, F. E., K. Borovickova, and R. Shimer (2015). A nonparametric variance decomposition using panel data. Technical report, Mimeo, University of Chicago.

Arranz, J. M., C. G. Serrano, et al. (2011). Are the mcvl tax data useful? ideas for mining. *Hacienda Pública Española 199*(4), 151–186.

Bonhomme, S. and L. Hospido (2017). The cycle of earnings inequality: evidence from spanish social security data. *The Economic Journal*.

Card, D., R. Chetty, and A. Weber (2007). The spike at benefit exhaustion: Leaving the unemployment system or starting a new job? *American Economic Review 97*(2), 113–118.

Couch, K. A., N. A. Jolly, and D. W. Placzek (2011). Earnings losses of displaced workers and the business cycle: an analysis with administrative data. *Economics Letters 111*(1), 16–19.

Elsby, M. W., B. Hobijn, and A. Şahin (2015). On the importance of the participation margin for labor market fluctuations. *Journal of Monetary Economics 72*, 64–82.

Fujita, S. and G. Moscarini (2017). Recall and unemployment. *American Economic Review 107*(12), 3875–3916.

García Pérez, J. I. (2008). La muestra continua de vidas laborales: una guía de uso para el análisis de transiciones. *Revista de Economía Aplicada 16*(1).

Güell, M. and B. Petrongolo (2007). How binding are legal limits? transitions from temporary to permanent work in spain. *Labour Economics 14*(2), 153–183.

Katz, L. F. and B. D. Meyer (1990). The impact of the potential duration of unemployment benefits on the duration of unemployment. *Journal of public economics 41*(1), 45–72.

Krueger, A. B. and A. Mueller (2011). Job search, emotional well-being, and job finding in a period of mass unemployment: Evidence from high-frequency longitudinal data. *Brookings Papers on Economic Activity 2011*(1), 1–57.

Laborda, A. M. (2013). La temporalidad en el mercado laboral español: nuevas aportaciones a la compresión del fenómeno.

Lapuerta, I. (2010). Claves para el trabajo con la muestra continua de vidas laborales.

Moffitt, R. (1985). Unemployment insurance and the distribution of unemployment spells. *Journal of Econometrics 28*(1), 85–101.

Rebollo-Sanz, Y. (2012). Unemployment insurance and job turnover in spain. *Labour Economics 19*(3), 403–426.

Silva, J. I. and J. Vázquez-Grenno (2013). The ins and outs of unemployment in a two-tier labor market. *Labour Economics 24*, 161–169.

Sullivan, D. and T. Von Wachter (2009). Job displacement and mortality: An analysis using administrative data. *The Quarterly Journal of Economics 124*(3), 1265–1306.

Tattara, G. and M. Valentini (2010). Turnover and excess worker reallocation. the veneto labour market between 1982 and 1996. *Labour 24*(4), 474–500.

Appendix

A.1 Step-by-Step guide to work with the MCVL

This guide provides with technical guidance on how to turn the raw csv files from the MCVL into a panel dataset in Stata. It extends Section 2.2 and includes more details on how to combine the different files, treat them and build a panel. Do files that follow these steps can be provided upon request.

The procedure is divided in four parts parts: formatting, binding, unemployment extensions and panel formatting.

1. Formatting

There are two ways of formatting the MCVL: year-by-year panel and retrospective panel:

- The **year-by-year panel** uses the information in all waves of the MCVL separately. This allows to take into account workers who are in some waves but not in others (see Garcia-Perez, 2009) and keeps the representatively of the population in every year.

- The **retrospective panel** uses information from the latest available year only. Although some representative of the sample is sacrificed, it is easier to study unemployment duration as there are no "cuts" as in the year-by-year version. However the previous waves are needed for wages, as the fiscal annex only has information relating the same fiscal year as the wave.

As described in the main body of the chapter, there may be applications where one or the other is preferable. In what follows I describe the year-by-year panel approach, as it is more complex. The same steps are needed for the retrospective panel, but there is no need to condition by year or censored spells to the year they are reported, which makes things simpler.

1.1 Formatting Affiliation files

Open the ASCII files for each year, name the variables (careful with the position of variables as it changes through years) and format the dates of start (*alta*) and and end (*baja*) of each spell.

Then proceed to clean the overlaps that may happen as in GP, cases a (total overlap) and b (partial overlap). Here the approach is to keep the longest continuous spell and drop smaller spells that happen at the same time. For partial overlaps, make the continuing spells start when the previous ends.

It is also a good idea to create a few auxiliary variables, namely a variable for labour market status. For this we need to combine information of different variables:

- *Tipo de relacion con la seguridad social*, codes 700-800 correspond to unemployment benefit claimants. Mark as unemployed ("U").
- *Tipo de contrato*, codes 400-900 correspond to temporary contracts. Mark as "T".
- *Tipo de contrato*, codes 99-400 correspond to permanent contracts.[29] Mark as "P". There is an exception though: code 540 corresponds to partial retirement, so I mark this as out-of-the-labour-force. Also those whose variable *regimen de cotizacion* is 140 are in early retirement.
- *Regimen de cotizacion*, codes 500-600 correspond to self-employment. I mark them as "A" for Spanish *autonomos*.[30]

You may also want to single out part-time contracts. For this you can refer to the accompanying .do file or refer to the official guide. The variable for this is *Tipo de contrato*.

Important: the variables *Empleador (forma Judirica) - Letra NIF de la Entidad Pagadora* and *Identificador (NIF/CIF) anonimado de la entidad pagadora* uniquely identify firms both in the affiliation files and in the fiscal file. If you want to use wage information, make sure to create a variable that joins both into one string variable. I call this variable *firmID* and move it right after the worker identified.

1.2 Formatting pension files:

Name the variables according to the official guide. If using for labour market flows, most of this variables are irrelevant, but keep the dates (and format them accordingly) and the personal identification number.

1.3 Formatting personal information files:

Name the variables according to the official guide.

Special care should be taken with the variable *fecha de defuncion* that marks the death date of some workers who passed. The birth date should also be considered carefully as there are some likely mistakes - most famously a worker who was supposedly born in 1906 and was still working in 2005 - likely a coding error for 1960.

[29] Note that some of these contracts may be *fijos discontinuos*, that is, permanent workers that only work for part of the year. They are different than temporary workers because they don't have a contract expiration rate and are protected by severance payments. If the reserchers wants to treat them differently, their contracts correspond to codes 300-400.

[30] If you want to be really precise, you should mark those whose *Regimen de cotizacion* is equal to 700-800 and 824-840 as self employed. These are the cases of farmers and sea captains.

Ideally the 2011-2012 personal file should be the most up-to-date information as there was a census in 2011. I recommend when joining all the years later to give preference for the education variables in 2012 onwards over earlier years, for all those that have more recent information available.

There are some exceptional cases of repeated personal identifiers. I chose to keep the youngest of the two, but whatever criteria you use, keep only one so it can merge easily with the affiliation file.

2. Binding

2.1 Binding the files together: pensions

Year by year, open affiliation and append the pension file. sort by date. If the final registry of a person is a pension registry (easily identifiable because all affiliation file variable will be blank) then fill in their labour market status variable as out-of-the-labour force. You can also fill in all the affiliation file info from the last spell if needed. Delete all other pension entries if you are only interested in labour market flows.

2.2 Binding the files together: personal information

Merge the formatted personal file and the affiliation file (with pension information) together, using the personal identifier as joining variable. It should match all cases, or almost. I keep the affiliation registered without personal information but drop the personal information entries without a matching affiliation entry.

Now it is a good time to drop all spells that happen before the current wave year - so 2006 only has spells active in the period 1st January 2006 until 31st December 2006 or beyond. This would ease the binding process below. Skip this step for the earliest year in the sample if you want to have some retrospective information before the start of the sample.[31] If you choose to keep retrospective information for all years you can, but that would make the merging years together process a bit more cumbersome. In case of doubt, keep all spells.

Important: create a variable called *year* and set it equal to the year in the wave. This will help you identify the information that each wave brings to the unified panel. Save the enriched affiliation files.

2.3 Binding all years together

Start with the earliest year in the dataset (it is recommended the 2005 wave as the absolute earliest). Drop all spells starting further than the 31st of December - modify the

[31] In practice this trimming will affect all observations that are active in later years, so if you are using many waves together keep all retrospective information and drop the repeated cases later, during the **Biding all years together** phase.

end date of the spell so it is 31st December. This right censoring should ensure the years are well matched together, so the 2006 file brings only spells active in 2006, the 2007 in 2007 etc. Append the next year affiliation file (with pension and personal information). Trim the ends as before and add the next year. Continue until you are left with the last wave. Do not censored this last year.

If you choose to keep retrospective information in each year, as you append each year erase duplicated spells. Here having created a variable for each year will come in handy, as it would help to identify identical spells but in different years (waves): they must share the same start date and the same *firmID* value. For the rare cases where firmID is not available (for example in unemployed spells) use the variable *Codigo de Cuenta de Cotizacion Principal (CCCP)* (right after the fiscal identifier variable) to identify duplicates spells.

Up until this point you should have one unique affiliation file with all waves joined together and no duplicated spells except from those that last beyond a calendar year - for example, a job that starts in May 2006 and ends in June 2009 should have 4 entries: one each for 2006,2007,2008 and 2009.

Depending on what you are interested in, it may be a good idea to create a *effective end date* variable that matches the latest end date for each spell - in the example above, set the effective end date as June 2009 in all entries. This way if you want to get statistics on tenure, you can either consider tenure up to the current year or total tenure in the sample. In the previous example, the first variable would be 8 months (May-December 2006) and the second 3 years (May 2006- June 2009). As a general rule it is better to create new variable than modify old ones, in particular with dates.

3. Extensions

3.1 Contract modification adjustment

In many cases contracts change across the years - this is the case of temporary workers promoted to permanent contracts. The way these cases are recorded in the MCVL is not easy to deal with. Ideally, for the purpose of job market flows we would like to have separate entries for each kind of contract.

Look at the varaible *Fecha de modificacion del tipo de contrato inicial o del coeficiente de tiempo parcial inicial*, towards the end of the affiliation file variables. If this variable is filled with a date, there was a change in contract. Now look at the next variable, *Tipo de contrato inicial*: this is the original contract code of the job. You can use the guide in step 1 to interpret this coefficient.

Create an indicator variable that is equal to 1 if (1) the current wave year equals the year of the contract modification date AND (2) the type of contract is not the same as the original type of contract (for example, if there is a change from temporary to permanent

(or vice-versa) or to part-time).

Duplicate the spell in which the indicator variable is equal to 1. Change the type of contract of the first copy to be the original type of contract. Change the end date of this first copy to coincide with the modification date, and change the start date of the second copy to the modification date. Depending on how you want to treat tenure, you may want to extend this last change to the start date of all the other entries in posterior years to the contract modification. Now you have two spells for each job: one before the contract change and one after.

Repeat these steps with the variable *Fecha de modificacion del tipo de contrato segundo o del coeficiente de tiempo parcial segundo* and *Tipo de contrato segundo*. This is the second contract modification variable.

Be careful when recording the length of each spell before and after the contract change. In some applications you may be interested in the whole period (for example for tenure) but if you want to count temporary and permanent job experience separately you may want to treat the two contracts differently.

3.2 Unemployment Expansions

Before proceeding, sort all spells by worker id, labour market state and date (in this order). Number the spells in separate variables for each state - so for example, if a worker was unemployed in two separate periods, create a variable called *number of unemployment spell* (NoU) and set it equal to 1 for the first one and 2 for the second. Or if a worker had 9 temporary jobs, create a variable NoT and number them chronologically.

Sort again the sample by id and date of entry and exit. Fill in all the blanks in NoU equal to the previous NoU value and set 0 for all spells before the first unemployment value. Using this variable (NoU), create another variable counting the days the worker is employed at each year in between unemployment spells. This will give us the total number of days contributed to the social security, which we will use to calculate unemployment benefit entitlements.[32] Remember to reset this counter to zero each time there is a new unemployment spell.[33]

Create a variable equal to the end of each spell (call it *original ending*) that will be of use later to calculate the extension period.

[32]Self-employed workers do not contribute to the social security so **do not count self-employment spells**.

[33]Some workers can choose to "save" part of their unconsumed unemployment benefits for next unemployed period, in which case the time contributed by the next job won't count towards the total. By resetting after each unemployment spell by default we make sure we only count the minimum possible time a worker could have contributed to the social security.

The LTU expansion

First, join consecutive unemployment spells *within the year*: if both unemployment spells came from the same wave, and one starts immediately after the other, I consider them one single spell.[34] There are many cases of workers that received more than one subsidy (because of illness or family reasons) and thus I don't want to record them as separate spells.

Second, if there is a difference between the end of an unemployment spell and the beginning of the next job, extend the end date of the unemployment spell as to join the two. Make sure that the next spell is employment or self-employment, and not retirement. The reason for this is that we can't be sure that these workers are looking for a job - if they transition to retirement probably they were out of the labour force to start with. I choose to extend the spells of workers whose last entry is unemployment to the end of the sample.[35] This is crucial to account for all the workers whose benefits and are still unemployed at the end of the sample. If your final year is beyond 2009, you should definitely do this as the number of unemployed workers without benefits reaches 50% in 2012.

Third, if the previous extension meant that the unemployment spell extended over the year of its original wave, duplicate the unemployment spell and set the wave year to the next one. If as a result it extends over two years, create two copies.

The STU expansion

In addition to the previous expansion, create a new unemployment spell if there is a gap between two jobs that lasts more than 15 days[36] and **at least one** of the following conditions are met:

1. the first job was self-employment
2. the first job ended in a quit (if the variable *Causa de Baja en Afiliacion* equals 51)
3. by the end of the first job, the worker hasn't accumulated 12 months of continuous employment

In all of the previous conditions, the worker is not legally entitled to unemployment benefits, and thus we can interpret the period between jobs as unemployment.[37] You can

[34]Some authors want to make distinctions between unemployment benefits and unemployment subsidies - the latter referring to reduced amounts that some long term unemployed workers receive after running out of unemployment insurance. If so be careful when applying this step.

[35]I restrict this expansions to the cases when the end date of unemployment benefits is within the two years prior to the end of the sample.

[36]This threshold is arbitrary. Results do not change much when the limit is put at 10 days or 1 month. García Pérez (2008) also considers 15 days as a reasonable threshold.

[37]See section 3.2 for more information on this.

further restrict these conditions by imposing that the firm identifiers of the two firms are different, so the worker is not being recalled to the same firm.

Set the end and start dates to fill in the gap between jobs. If this expansion takes the unemployment spell over the year of the wave, duplicate as in the previous case.

Finally, stata creates a new variable every time you duplicate observations to identify all duplicates. If your software of choice does not do that, make sure you have an indicator variable for these unemployment spells so you can identify them later.

4. Panel Formatting

4.1 Select the window

The LFS runs interviews during the reference quarter, and so it gets its answers from replies to an unwound reference day within the quarter. This is inevitably going to lead to discrepancies in the results, as if the reference day in the MCVL doesn't coincide with the LFS the answers can be different. The extent of the discrepancy would depend on the frequency of flows: if there are more transitions within the month than within the quarter, then the probability of discrepancy is higher. The approach here is to select a window period within the quarter (or the month if interested monthly transitions). I chose the 15th to the 30th of the first month of each quarter. That is, the 15th-30th of January, April, July and October. Avoiding the first of these months is important, especially in the case of January as many jobs start after the Christmas break - which in Spain can last up until the 6th of January. Several robustness checks can be done by increasing the window, but the results don't substantially change - except in the Christmas season as noted.

4.2 Create quarterly state variables

Once the window has been chosen, we look at the spells that fall within it, and restrict our attention to the spells whose entry date is after the beginning of the window period, but before the end. If there is ore than one spell within a window, I chose to keep the one that continues in time - that is, the last one. Another approach is to take the one with longest duration, but this can prove difficult. For example, a long employment spell that ends the 22th of January. The spell would have to count if the longest duration rule would prevail, but unemployment would prevail in the case of continuing spell. If the next window sees the worker employed again, not counting unemployment can understate the flows in and out of unemployment. Because unemployment would likely lose if the longest spell counts, I choose the continuing spell approach.

Once we have one spell per window, all that is left is to fill in a variable for spell-period, and this can easily imply that copies of the spells need to be taken if they feature

in two different windows. For example, an employment spell that features in the first and second quarter would be duplicated. The original will be assigned to quarter 1 and the copy to quarter 2.

If using any of the expansions approach, you may also want to create a different state for extended unemployment spells. For example, if an unemployment spell that originally lasted for a quarter now lasts for two - because of the LTU expansion - then we can label the first observation "U" and the second "0". For this we can use the extension date variable we created before modifying the start and end dates of spells. All unemployment spells generated from the gaps expansion can also be labelled differently.

4.3 Create stocks and flows

The last step is to only keep spells that feature in a quarter and discard all of the other spells. We will be left with a panel dataset that mimics the structure of the LFS. This can be used to calculate unemployment (with or without benefits) and temporary share of employment, for example.

To create flows, just link two consecutive quarters for the same worker. Here it could be a good idea to clean "TUT" or "UTU" flows, conditioning on the duration of unemployment (or the temporary contract). Note that this can also be achieved by following a different rule when choosing one state per quarter.

A.2 STU additions

Table A.1 provides summary statistics for the unemployment spells broken by modification. The first to notice is that most unemployment spells (37%) are not modified, but the number of modified spells is also substantial (30% for the LTU expansion and 32% for the STU). It is worth noting that not all of these spells appear in the unemployment rates pictured in figures 4 and 7, only those that correspond to the state of the worker in a given quarter. The statistics in table A.1 correspond to all spells. The first thing to notice is that unemployed workers with a STU expansion spell are younger and have less experience in all type of employment and unemployment. In particular, they have been unemployed an average of 0.7 years in their working lives. The spells are also shorter than those of the LTU expansion (205 versus 430) but the original spells are even shorter. This is because of two reasons: First, a requirement on STU spells is that they last longer than 15 days (which is not required for registered unemployment spells). Second, many registered unemployed workers with long spells run out of benefits, which means that long spells are in the LTU expansion category. Figure A.17 illustrates this point by showing the histogram of spell duration by extension.

Table A.1: Summary Statistics of Unemployment Spells

	No Modification	LTU Expansion	STU Expansion
Observations	755,413	625,973	661,626
Percentage	0.370	0.306	0.324
Age	39.79	36.68	31.00
	(10.79)	(10.15)	(10.12)
Female	0.462	0.445	0.449
	(0.499)	(0.497)	(0.497)
Duration (days)	148	430	205
	(322)	(494)	(286)
Experience PC	7.376	5.638	2.011
	(8.502)	(7.108)	(3.928)
Experience TC	3.104	2.702	1.435
	(2.633)	(2.311)	(1.653)
Experience Unemp	2.053	1.761	0.714
	(2.227)	(2.003)	(1.349)

Sample is all unemployment spells in the 2004-2013 period. Averages with standard errors in parenthesis. Experience is measured in years,duration of the spell in days.

Figure A.17: Histogram of Spell duration, by Extension