

Online Appendix

Unemployment in Administrative data using Survey data as a benchmark

Cristina Lafuente
University of Edinburgh

June 2019

This appendix provides technical guidance on how to turn the raw csv files from the MCVL into a panel dataset in Stata, similar to the Labour Force Survey (LFS) panelling format. It extends the appendix of the main paper and includes more details on how to combine the different files, create and select new variables and build a panel. Do files that follow these steps can be provided upon request.

The code accompanying this description is freely available to download from the Github repository <https://github.com/crisla/MCVL>. The repository includes detailed instructions on how to request access to the data and how to name and place the files for the code to work. The code is written in STATA.

There are four main blocks: formatting, binding, implementing unemployment expansions 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 García Pérez (2008)) and keeps the representativity of the population in every year.
- The **retrospective panel** uses information from the latest available year only. Although some representativity of the sample is sacrificed, it is easier to study unemployment duration as there are no “cuts” or overlapping spells as in the year-by-year version.

As explained in the main paper, 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 and appropriate for unemployment measurement. The same steps are needed for the retrospective panel, but there is no need to limit spell duration to be within the year they are reported, which makes things simpler.

1.1 Formatting Affiliation files

Open the ASCII files for each year, name the variables according to the MCVL guide. Be careful as the position of variables as it changes through years. Then format the dates of start (*alta*) and end (*baja*) of each spell, which together with the personal identifiers define each spell.

Next proceed to clean the overlapping spells (spells beginning before the end of the previous spell) as in García Pérez (2008) cases a (total overlap) and b (partial overlap). For total overlaps, 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 time to create the most important variable: labour market status during the spell. I chose to make this a string variable - which will be convenient when creating flows - but a labelled numerical value would work as well. For this we need to combine the information of 4 other 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 temporary, “T”.
- *Tipo de contrato*, codes 99-400 correspond to permanent contracts.¹ Mark as permanent, “P”. There is an exception though: code 540 corresponds to partial retirement, so I mark this as out-of-the-labour-force (OLF). Also those whose variable *regimen de cotizacion* is 140 are in early retirement - so I mark them as OLF as well.
- *Regimen de cotizacion*, codes 500-600 correspond to self-employment. I mark them as “A” for Spanish *autonomos*.²

¹Note 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.

² Those cases where *Regimen de cotizacion* is equal to 700-800 and 824-840 should also be counted as self employed. These are the cases of farmers and sea captains.

You may also want to create other auxiliary variables, such as an indicator for 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 identifier.

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.

Ideally the 2011-2012 personal file should be the most up-to-date information as there was a census in 2011. It is recommend to impose the values of the education variables from 2012 onwards over earlier years, whenever possible.

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 the formatted affiliation file and append the pension file to it. sort by date. If the final observation of a person is an entry from the pension file (easily identifiable because all affiliation file variable will be blank) then fill in their labour market status variable as OLF. Then fill in all the information missing in this last entry from the

previous spell (place of residence, for example). Delete all other pension entries if you are only interested in labour market flows.³

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 virtually all cases. For the rare exceptions, I choose to 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 onwards. This would ease the binding process below. Skip this step for the earliest year in the sample if you want keep retrospective information before the start of the sample.⁴ You can keep retrospective information for all years, but that would make the merging years together process more cumbersome. Whenever in doubt, keep all spells.

Important: create a variable called *year* and set it equal to the file year. 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.⁵ Drop all spells starting further than the 31st of December - modify the 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 affiliation file (with pension and personal information). Trim the spells over the end of the year and add the next file. 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 unemployment spells) use the variable *Codigo de Cuenta de Cotizacion Principal (CCCP)* (right after the fiscal identifier variable) to

³For example, Some disability or widowhood pensions can be received while in the labour force. If these are of no interest for the study, they can be committed.

⁴In 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.

⁵ It is recommended the 2005 wave as the absolute earliest to keep the sample representative of the population in that year. However you can use the 2005 file but name an earlier year as your starting point. However the further away from 2005, the worse the representativity of the sample.

identify duplicates spells.

At 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, and this is particularly true with dates.

3. Expansions

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 variable *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, then there was a change in contract. The next variable, *Tipo de contrato inicial*, contains the original contract code of the job. Use the guide in step 1 to interpret its value.

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).⁶

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.

⁶In case you are interested on recalls or temporary contract renewals, you may wish to also create a new entry even if the contract type doesn't change.

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.⁷ Remember to reset this counter to zero each time there is a new unemployment spell.⁸

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

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.⁹ There are many cases of workers that received more than one subsidy at the same time (because of illness or family reasons) but they are part of the same unemployment spell.

Second, if there is a gap 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

⁷Before 2013, self-employed time does not count as contributions towards unemployment insurance, so do not count self-employment spells.

⁸Some 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.

⁹Some 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 you may want to skip this step.

sure that the next spell is employment or self-employment, and not retirement. The reason being that we cannot 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.¹⁰ This is crucial to account for all the workers whose benefits have expired 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 expansion meant that the unemployment spell extended over the year of its original wave, duplicate the unemployment spell and set the wave year equal to the next year. If as a result the spell extends over two years, create two copies. For example, if a spell started and ended in 2009 but after the expansion it ends in 2010, create a duplicate of the original spell and modify its year so it belongs to 2010. This way there would be two copies: one in 2009 and one in 2010 - as it would be the case if unemployment benefits wouldn't have expired.

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 days¹¹ 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* is 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 unregistered unemployment.¹² You can 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. I follow this approach in the paper. Set the end and start dates to fill in the gap between jobs.

Expand also unfinished spells that start towards the end of the sample. For my period of study, 2005-2013, I choose to expand all unfinished spells that satisfy the requirements above and start within the last 3 years of the sample. As before, accounting for spells that are not yet done is important to accurately reflect unemployment. See section 4.1

¹⁰See section 4.1 in the main paper for a more detailed discussion on the effect of this extension and some alternatives.

¹¹This threshold is arbitrary. Results do not change much when the limit is put at 10 days or 1 month. Garcia-Perez (2008) also sets 15 days as a reasonable threshold.

¹²See section 3.3 in the main paper for more information on this assumption.

in the main paper for a more detailed discussion on the effect of this extension and some alternative assumptions.

As before, if this expansion takes the unemployment spell over the year of the wave, duplicate the affected observations. There should be one copy for each year the spell takes place in.

Finally, stata creates a new variable to identify copies and originals every time you duplicate observations. 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 in an unknown reference day within the quarter. This is inevitably going to lead to discrepancies in the results, as if the reference day in the MCVL does not coincide with the LFS the answers can be different. The extent of the discrepancy is discussed in section 4.2 . Overall the window of choice does not seem to make much of a difference. The approach here is to select a window period within the quarter (or the month if interested monthly transitions). I chose the 1st to the 15th of the first month of each quarter. That is, the 1st-15th of January, April, July and October. Allowing for 2 week windows 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.

4.2 Create quarterly state variables

Once the window has been chosen, I focus on the spells whose entry date is after the beginning of the window period, but before the end. It is important to clear completely overlapping spells (with the same start and end dates) before proceeding. If there is more than one spell within a window, I chose to keep the one that continues onwards - that is, the last one. Another approach is to take the one with longest duration, but this can prove difficult. For example, take a long employment spell that ends the 22th of January, followed by 6 months of unemployment. The first spell would be selected if we apply the longest duration rule, but the second spell would be selected instead if we keep the continuing spell. This is particularly important if in the next window period the worker is employed again, as not counting unemployment can understate the flows in and out of unemployment. I discuss this criterion in more detail in section 4.3.

Once we have one spell per window, all that is left is to fill in a variable for a spell-

period. In many cases you will have to create copies of the same spell, when it appears in two different windows. For example, an employment spell that features in the first and second quarter would need to be duplicated. The original will be assigned to quarter 1 and the copy to quarter 2.

If you have applied any of the unemployment expansions, you may also want to create a different state for unregistered unemployment spells. For example, if an unemployment spell that originally lasted for a quarter now lasts for two - because of the LTU expansion - then you can label the first observation "U" and the second "0". For this you must use the expansion 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 other spells. We will be left with a panel dataset that mimics the structure of the LFS, with one observation per quarter. In this case however we will have more detailed information - precise tenure and experience data for example. This can be used to calculate unemployment stocks (with or without benefits) and temporary share of employment, for example.

To create flows, just link two consecutive quarters for the same worker. String variables are best for this, as you can simply add two strings to form a new variable. Here you may want to relabel "TUT" or "UTU" flows, conditioning on the duration of unemployment (or the temporary contract). This can also be achieved by following "the longest spell" rule when choosing one observation per quarter. Note that this brings the MCVL flows closer to the LFS, but they are not classification errors - which is the reason these flows are modified when working with surveys (see Elsby et al. (2015) for more details).

References

- 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.
- 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).