Probabilistic matching between LoSaI and CICO. Methodological annex.

Fabio Berton, Lia Pacelli, Roberto Quaranta and Francesco Trentini

This annex describes the construction and validation of the dataset on which the econometric analysis is run. The procedure to implement the probabilistic matching between LoSaI and CICO is published in a public repository by the authors[1]. As a first step, it is important to give an account of the potentiality of the existing data in providing a source of information suitable for our analysis, which is directed to the evaluation of the impact of hiring incentives and reduced firing costs on young workers' probability of obtaining permanent contracts (Table 1). The two data sources, CICO and LoSaI, are disseminated by the Ministry of Labor and Social Policies[2] but are built on different archives. CICO is structured as a register of employment relations and is a 48-date sample of the compulsory notifications that employers, public and private, send to the Ministry of Labor and Social Policies, to which a sample of autonomous workers is added. As far as our study is concerned, the source includes detailed information on hiring incentives since 2011. On the base of the available information, we can identify individuals eligible for such incentives by looking at previous work episodes, which are included in the sample as the sampling procedure is done at the individual level. However, contract transformations from temporary to permanent contracts are not recorded and cannot be identified either. On the contrary, the region where the job is performed+ is documented. Further information on the characteristics of the employment relationship is occupation, qualification, part-time and reason for termination. Half of the sample presents wages, the nominal wage communicated at the beginning of the contract. Individual characteristics are rich: gender, year of birth, education, qualification, citizenship, region of residence**.** The database does not include firm size information and therefore it is not possible to identify the set of firms affected by the changes in employment protection legislation.

Let's now consider the LoSaI dataset. The archive has an event structure. A new event is a change in the employment relationship that is relevant for social security, for instance, a change of contract type, job title, contract, qualification, work area, and so on. Therefore, the database has a panel structure and for each employment relationship there may be more records in the same year. Transformations to permanent contracts can be clearly identified as a new event is generated when the contract type is changed. A date of transformation can also be estimated on the base of the actual days of work that are registered for each episode, given the contract starting date. The database includes detailed information on the hiring incentives**,** in particular a variable that allows to identify the recipients of the latest social security contribution reliefs, separating the three-year incentive established in 2015 (Tipo_politica == 51), the two-year hiring incentive that was then established in 2016 (Tipo_politica == 52, out of the scope of the present analysis) and other forms of hiring subsidies (Tipo_politica == 5).

[1] https://github.com/francesco-trentini/abbinamento-cico-losai

[2] http://dati.lavoro.gov.it/microdati-la-ricerca

Table 1. Summary of the available information in CICO and LoSaI.

Characteristics	CICO Variable	LoSaI Variable
Beginning of work relation (day)	Rapporto_DataInizio	Data_assunzione
End of work relation	dtCessazioneEffettiva	Data_cessazione
Type of job contract	codTipoContratto	Tipo_contratto
Region of work	codRegioneLavoro	-
Salary	RetrMese_INPS (estimated)	Retribuzione_imponibile (final)
Type of incentive	codAgevolazione (up to 2012)	Tipo_politica
Qualification	codQualificaProfessionale	Qualifica
Contract trasformation	-	Not present but inferable
Cause of termination	codMotivoCessazioneCO	Motivo_cessazione
Sex	codGenere	SESSO
Age	AnnoNascita	ANNO_NASC
Education	codTitoloStudio	-
Citizenship	codCittadinanza	-
Region of living	codRegioneDomicilio	REGIONE
Firm identifier	cf_datore_crip	ID_AZIENDA
Year	-	ANNO
Firm size	-	CLASS_DIM
Sector of economic activity	-	ATECO07_2_CALC

Source: own representation.

The whole sample has information on wages, which are actual wages on which social security contributions are calculated. Additional information about the employment relationship is qualification, part-time, reason for hiring and reason for termination. The individual characteristics available are gender and year of birth**.** The database includes information on the size of the firm, organized in classes. Therefore, in addition to the perfect identification of actual and potential recipients of hiring incentives, the information is sufficient to identify the firm exposed to the changes in employment protection legislation. Unfortunately, two fundamental aspects are not covered and impinge on the possibility of fully relying on LoSaI: the region of work and education. The strategy that we design to solve this issue is the following. We use LoSaI as the master dataset because it includes the possibility to identify employment relationships that are eligible for incentives, those incentivized and firm size, expressed as the number of employees. In addition, actual wages can be calculated. Nonetheless, LoSaI misses two important dimensions, namely the region where the job is localized and the highest education level achieved by the worker. Our strategy is to enrich LoSaI with CICO, which registers these dimensions. We opt for a probabilistic matching on individual work relationships on the sample of overlapping reference populations, while residually imputing the region of residence as the region of work for the remaining subset.

Data construction

Each dataset is elaborated and restructured as a panel of work relations uniquely identified by a worker identifier, a firm identifier (both consistent within each source, but not across them) and the start date of the employment relationship. The variables of the two datasets are harmonized. Then we move to the next stage which requires matching each individual in the LoSaI database to a single individual in CICO. In order to describe the method followed in the matching, we present an example. First, we implement a many-to-many matching over the following set of characteristics:

- Work relation starting date (DD/MM/YYYY)
- Region of residence (21 NUTS2 region)
- Year of birth
- Sex
- Contract (Permanent or temporary)
- Time schedule (Full-time, part-time)

Table 2. Example of match and potential issues, drawn from the subsample 2012-2014.

index	Id_losai	id_wr_losai	Id_cico	Start_date
1	L1	1	C1	28/05/2012
2	L1	2	C2	06/09/2012
3	L1	3	C1	03/12/2013
4	L1	3	C3	03/12/2013
5	L1	4	C1	09/12/2013
6	L1	4	C4	09/12/2013
7	L1	5	C5	01/04/2014
8	L1	6	.	05/05/2014
9	L1	7	C6	03/11/2014
10	L1	7	C7	03/11/2014
11	L1	7	C1	03/11/2014
12	L1	7	C8	03/11/2014

Source: LoSaI and CICO statistically integrated sample.

The many-to-many procedure allows more than one record in the using file (CICO) to be matched to the same record in the master file (LoSaI) and vice versa. An example is given in Table 2, where individual X in LoSaI is matched with 8 individuals in CICO.

The third work relation (id_wr_losai==3) is matched to two records in CICO, associated to id_cico C3 and C1. The latter is present three more times in the career, associated with id_wr_losai 1, 4 and 7. In the first case the match is perfect, while in the others more than one work relation is found. The work relations id_wr_losai 2 and 5 are associated with other CICO ids, while id_rl_losai is not associated with any CICO work relation. Therefore, the individual id_cico C1 is associated to the id_losai L1 in five out of seven valid cases.

We build an indicator to quantify the quality of the match. We measure the precision of the match between LoSaI and CICO at the individual level. For each individual in LoSaI, we count the number of observed work relations and the number of corresponding CICO individuals. The indicator is calculated as the number of recurrences of a CICO id over the total number of work relations observed for each individual in LoSaI:

$$PrecisionLoSai_{ij} = \frac{numberofmatches(idLoSai_j, idCico_{ij})}{numberofworkrelationsforidLoSai_j}$$

In the example presented above the id_cico with the highest number of matches is C1, with 4 matches on 7 work relations and a precision of 0.5714. The lowest possible precision (0) corresponds to the missing case, while the average precision is 0.1746 and informs us that we observe many low-volume matches with a high number of id_cico.

Table 3. Summary measures of the precision of the match of L1, subsample 2012-14

Id_losai	Id_cico	Number of matches	id_wr_losai_ max	Precision	Precision min	Precision max	Precision avg
L1	C1	4	7	0.571429	0	0.571429	0.174603

Source: LoSaI and CICO statistically integrated sample.

The single procedure on the whole sample is repeated iteratively. At each repetition, the individuals in LoSaI and CICO with precision 1 are excluded. The next iteration features a reduced set of individuals in both datasets and allows to use the same criterion to identify new full-precision matches. After a certain number of repetitions, no matches survive. In this case, we need to vary the parameters of the match, namely the list of features over which we perform the match and the threshold of precision. We opt for changing both parameters subsequently: for each level of precision (1, 0.75 and 0.5) we loop over four different keys, presented in Table 4, so that once we end iterating on the key for a level of precision, we lower the precision threshold and repeat the procedure over the same set of keys.

Table 4. Sets of keys used for matching

	Key
1	rapporto_datainizio [Work_relation_start_date], regione_abitazione [region_of_residence], codgenere [sex], annonascita [year of birth], contratto [permanent/ temporary], fulltime [fulltime/part-time]
2	rapporto_datainizio, codgenere, annonascita, contratto, fulltime
3	rapporto_datainizio_s, codgenere, annonascita, contratto, fulltime
4	rapporto_datainizio_m, codgenere, annonascita, contratto, fulltime

Note: rapporto_datainizio_s and rapporto_datainizio_m refer to the start date, coded to weekly bins or monthly bins respectively. The rationale is that we let the constraint be gradually less stringent by widening the time window on which work relations are matched.

Validation

We use official INPS publications as a benchmark to assess the quality of our dataset in representing stocks of workers affected by the hiring incentive. As we have seen above, the origin of our master database is LoSaI, which is a sample of social security records, simply restructured from a person-event structure to a work-relations structure. The subsequent enrichment of the information by means of statistical matching does not modify it. Therefore, having our sample the same theoretical reference population of the data used by INPS for the annual report 2018 (INPS 2018: Table 1.28, pages 67-68), we replicated their reported results. More specifically, we focused on the phenomena of interest, *i.e.*, contract activations, conversion, and incentives in 2015. Considering that LoSaI is a sample while INPS data are the whole population, we provide computed values for the population and a 24-dates sample respectively. Sample and population figures are then directly comparable. The comparison is presented in Table 5. The table reports the number of permanent contracts activated in 2015, except for apprentices, separating new hires and conversions from temporary contracts. We further compute the number of incentivized work relations in the group identified above and its prevalence. The table clearly shows that our attempt correctly estimates the number of incentivized work relations among new hires while overestimating the number of new activations. On the contrary, conversions are correctly estimated both in the number of incentivized and total numbers.

References

INPS (2018), *XVII Rapporto annuale*, Roma, INPS [link]

Table 5. New permanent contracts and conversions of temporary in 2015. Highlighted cells report computed values

	Hires					**Conversions**					**Total**			
	Our database		**INPS**			**Our database**		**INPS**			**Our database**		**INPS**	
	n	**N**	**n**	**N**		**N**	**N**	**n**	**N**		**n**	**N**	**n**	**N**
Total number of incentivized work relations	76,557	1,158,930	74,765	1,121,469		29,573	443,595	27,022	405,326		106,835	1,602,525	101,786	1,526,795
Total number of new work relations	148,598	2,328,495	130,545	1,958,181		39,231	588,465	35,585	533,770		194,464	2,916,960	166,130	2,491,951
% incentivized work relation on total		50%		57%			75%		76%			55%		61%

Source: LoSaI and CICO statistically integrated sample and INPS (2018)