doi.org/10.362653/IJAS.16751

MATCHING AND INTEGRATION OF REGISTRY AND SURVEY DATA ON THE DYNAMICS OF WORK HISTORIES: A PILOT STUDY

Martina Bazzoli[1], Sonia Marzadro
FBK-IRVAPP, Trento, Italy

Abstract. *In the last decade the practice of combining longitudinal surveys with administrative data has been affirmed, enabling the integration of the complementary advantages offered by these two data sources. In this perspective, the paper describes a pilot study conducted in 2015 involving deterministic matching and integration between a retrospective panel survey carried out on a representative sample of households living in the province of Trento (Italy) and the register data from the provincial section of the administrative archive of INPS (the Italian social security agency). The aim was to create a comprehensive database for the study of work histories. Through the survey we address the main limitation of INPS data, which do not cover the universe of workers and jobseekers. Conversely, the administrative data, providing richer and more reliable information on most work episodes and instances of subsidized unemployment, help mitigate one of the most significant sources of errors in the panel surveys, i.e. the distortions caused by memory bias.*

Keywords: *data integration, administrative data, survey data, labour market, work histories*

[1] Corresponding author: Martina Bazzoli, email: bazzoli@irvapp.it
The research has been carried out in the framework of the Agreement Protocol between FBK-Irvapp, the Trentino-Alto Adige Regional Directorate of Inps and Ispat (n. 157/176704109). We are grateful to Antonio Schizzerotto and Ugo Trivellato who coordinated the research project. We also thank Marco Zanotelli and Giovanna Fambri for the support provided by Inps and Ispat respectively in data acquisition and integration, and to Bruno Anastasia, Cinzia Giannelli, Gianluca Mazzarella and Michele Mosca for their comments and suggestions.

1. INTRODUCTION

Are today's youth worse off than yesterday's? The question often arises in the press, in public debate and in economic and social research. However, answering this question is difficult. Indeed, being worse – or better or in the same way – involves comparisons between generations. Such comparisons require dealing with complex methodological issues, but even more, require longitudinal datasets following different cohorts of individuals – today's young people and those of yesterday – for their entire life course or at least for a long part of it.

Usually, many countries collect this kind of data through large multi-purpose household panel surveys using appropriate research designs. In this regard, Understanding Society is the exemplary longitudinal study of the United Kingdom launched in 2009 with annual waves on individuals of about 40,000 families.[2]

In countries - typically Scandinavian - where the production of statistical information is largely based on registers of administrative origin containing comprehensive information on the population or a key sub-population, the alternative to survey data is made up of longitudinal databases for thematic areas derived directly from registers.[3]

Rather than choosing one data source over another, the practice of combining longitudinal surveys with administrative data in such a way as to maximize the strengths and minimize the weaknesses of each source – a process called *data integration* – has been established in the last decade especially in English-speaking countries (Calderwood and Lessof, 2009). The example is still provided by Understanding Society, in which, with the consent of the adult respondents, the pertinent administrative data on health and education are linked to the survey data.

[2] The sample also includes 8,000 households from the previous British Household Panel Survey, which started in 1991. Among other longitudinal household surveys, is worth mentioning the German Socio-Economic Panel (SOEP), conducted since 1994, and the US Panel Study of Income Dynamics (PSID), started in 1968.

[3] Here the Danish experience in the labour market field is paradigmatic: Statistics Denmark has developed and updated annually, IDA (Integrated Database for Labour Market Research), a base of longitudinal data on the population of workers and enterprises since 1980.

Unfortunately, in Italy, longitudinal databases that follow an appropriate group of individuals over a long-time, whether achieved through household panel surveys[4] or derived from administrative sources[5], are not currently available at the national level despite the considerable progress made by the National Statistical Institute (Istat) in terms of integration of administrative and survey data[6]. Moreover, it must be noted that Italy suffers from restrictive legislation on privacy and protection of personal data which limits the process of integrating data for scientific purposes (Trivellato, 2019). Among the few notable Italian experiences, the AD-SILC[7] database stands out as a significant contribution. It was constructed by longitudinally linking administrative data from various National Social Security Institute archives, covering diverse social groups such as public and private sector workers, self-employed professionals, pensioners, and recipients of other insurance-based welfare benefits. These data were further integrated with microdata from IT-SILC, the Italian version of the EU-SILC survey.

[4] The large national cross-sectional surveys with a longitudinal component, IT-SILC (the Italian section of EU-Survey on Income and Living Conditions), the Labour Force Survey and the Survey on Household Income and Wealth of the Bank of Italy cover individuals for a short period of time. Then there are two other multi-purpose household panel surveys but they were conducted between the end of the twentieth and the beginning of the twenty-first century: the European Community Household Panel in which also Italy were included (8 annual waves between 1994 and 2001, and the Italian Longitudinal Household panel (5 biennial waves, from 1997 to 2005).

[5] Administrative data are restricted to segments of population. The most interesting experience in the labour market field concerns longitudinal databases that have been built on management archives of the National Social Security Institute. Among them the WHIP (Work Histories Italian Panel) contains a linked employer-employee longitudinal database. WHIP has been widely used for studies on the mobility of the Italian labour market however his update has stopped in 2004.

[6] Such dataset is, indeed, mainly focused on firms and income variables (Sestito et Trivellato, 2011; Istat, 2009

7 In the first version of AD-SILC, INPS data were merged with the 2015 IT-SILC microdata (Fondazione G. Brodolini, MEF, 2016). In the latest version (Conti et al., 2023), AD-SILC integrates longitudinal data from IT-SILC (covering the period from 2004 to 2017) and INPS data. AD-SILC has been used primarily, but not exclusively (Naticchioni et al., 2016, Fabrizi and Rocca, 2024), as the basis for the dynamic microsimulation model T-DYMM developed by the Treasury Department of the Italian Ministry of Economy and Finance (Conti et al., 2023).

Given this context, in 2015 FBK-IRVAPP, in collaboration with the Regional Directorate of the National Social Security Institute (INPS) and the Institute of Statistics of the Province of Trento (Ispat), conducted a research project aiming at providing a picture, as accurately as possible, of the work career of a sample of individuals living in Trentino. This was achieved through the deterministic matching and integration of a panel survey on households' life conditions (*Indagine sulle condizioni di vita delle famiglie trentine*, hereafter ICFT) carried out on a representative sample of about 2,500 households and register data on all employees managed by the local agency of the National Social Security Institute (hereafter INPS).

The study was conducted on a small-scale context - the autonomous province of Trento - and faced with empirical research issues not based on a well-grounded literature. However, in our opinion, this study holds significant value as one of the rare Italian examples of data integration. It is, as far as we know, the only one that not only merges but fully integrates the employment episodes recorded in administrative data (at that time) with the entire work histories collected through a retrospective survey thereby leveraging the strengths of both sources to provide a more comprehensive and accurate representation of individuals' employment trajectories.

The aim of this paper is to describe the archives involved, the integration process, the various challenges faced, and the characteristics of the resulting dataset. This dataset has been used to explore the main features of the evolution of the work histories of individuals over a time span that extends, overall, from 1974 to 2014. Although this part is not discussed here, a brief summary of the empirical evidence is provided in Section 5 (see Bazzoli et al., 2018 for more detailed findings).

The paper is structured as follows. In Section 2 we provide an overview of important considerations for linking administrative and survey data for research. The two sources of data used are introduced in Section 3, while Section 4 describes the integration procedures. In Section 5, we recall the key findings on the employment dynamics of young people over the last 40 years as studied using the aforementioned database. Conclusions are offered in Section 6.

2. INTEGRATING ADMINISTRATIVE AND SURVEY DATA TO BUILD EVIDENCE

Data sources are an essential part of research and whether administrative and survey data differ significantly is becoming a matter of discussion among researchers. In addition to the tightening of survey budgets, it is well-known that falling response rates and memory recall errors are the two main challenges that threaten the quality of collected survey data thus reducing the confidence in the conclusion drawn from such data. Moreover, there is an increasing need to satisfy a growing demand from users for good quality statistics and above all enabling faster measurement of new phenomena.

Administrative data can be a powerful resource in this respect particularly because of the insights these data might offer into social inequality, human behaviour and the effectiveness of social policies (Einav and Levin, 2014; Connelly 2016; Card et al., 2010). Moreover, they are an authoritative source of data, cover a large number of observations, are regularly updated and thus are available for long periods of time (Künn, 2015). Therefore, they can be used not only to enrich survey data, but also to lower survey costs by requiring fewer questions during interviews, reduce the burden on respondents and lessen survey dropout and item nonresponse rates. As a matter of fact, linking administrative data to survey data has become a promising and innovative strategy which affects the quality and quantity of research and increases the potential of data (Crato and Paruolo, 2019; Künn, 2015; Al Baghal et al., 2014; Calderwood and Lessof, 2009).

However, despite the clear advantages, there are a number of challenges in using and integrating that kind of data. First of all, there are ethical and legal considerations that need to be resolved. Indeed, legal constraints on the use of data and the need for anonymization restrict their access and content (Livraga, 2019; Trivellato, 2019; Sakshaug et al., 2012). Moreover, requesting consent to use the linked data may introduce consent bias (that might occur when consenters differ from non-consenters) or reduce response rates, introducing selection bias (Sakshaug et al., 2012).

Another issue deals with the time and conceptual alignment. A reference period in survey data is commonly the state at the time of the interview (e.g. current activity status) or is collected via retrospective questions (e.g. career history) while the data in administrative archives usually are registered in a fixed

point in time. Finally, variables collected for administrative purposes sometimes may diverge from fundamental economic concepts (e.g. unemployment status) or are missing because they are not of interest to the administrators who collect data (such as attitudes and family relationships).

The linkage of register variables to surveys is quite straightforward if the survey units also have valid identifiers as in the case of survey population drawn from an administrative archive. In this case the data can be directly linked via exact/deterministic matching because both data sources contain a unique personal identifier. In the absence of unique identifiers, statistical matching or modelling approaches have to be used instead (Harron et al., 2017).

Although the primary motivation for linking to administrative data is usually to enhance survey data, this can be done in different ways because it may involve pooling or combining information. On one side, administrative data are used to supplement survey data. In this case they are used as a direct substitute for survey questions (low integration). On the other hand, administrative data are used to complement/validate survey data. This occurs, for example, when administrative data is available only for a subgroup of survey respondents (high integration), like in the present case.

In the next sections we clarify how we have used the two sources of data mentioned before – a multi-purpose panel survey with retrospective part on working life of a representative sample of individuals and register data that provide much more detailed and reliable information on work histories than survey data but on a subsample of workers – in order to create a database that enables to the study of individuals' work histories over a period of nearly four decades.

3. DATA DESCRIPTION

3.1 THE ICFT PANEL SURVEY

The survey used is a multi-purpose longitudinal survey with a retrospective part on working life conducted on a representative sample of about 2,500 households from Trentino, an autonomous province of Northern Italy. It was started in 2005/06 with biennial waves[8] conducted using face-to-face computer assisted personal interviewing (CAPI).

[8] For details see Fambri and Schizzerotto, 2008.

The first wave gathered retrospective information on all significant events occurring to the members of the sample in the period between their births and the date of the interview. The purpose of each of the subsequent surveys was to update this information, recording all significant events of life histories that occurred to the members of the sample in the period between the previous interview and the date of the current one.

The most complex part of the interview was the reconstruction of the individual working career, including any periods of unemployment or inactivity. The survey gathered the dates of each episode experienced by the interviewees, as well as the main features (such as labour contract, activity sector, type of work). Each 'work-history' studied is represented by a continuous sequence - of varying duration – of distinct episodes, each at least one-month long. The starting date of each episode (except the first) corresponds to the ending date of the one that precedes it. The career of a respondent is thus represented by a sequence of non-overlapping episodes - of work, unemployment or inactivity - lasting at least one month between the date of the first entry into the labour market and the date of the interview.

The sample for the experience of integration presented in this paper is composed of 5,756 adults interviewed during the 2012 wave.

3.2 THE ADMINISTRATIVE INPS DATA

For each individual belonging to the reference sample, individual information was then acquired in 2015 from the National Social Security Institution (INPS), from "*Cassetto previdenziale del cittadino*", which contains the list of contributions recorded in favour of the worker from the opening of an insurance position.

Contributions belong to different funds: 1) private sector employees; 2) clergy; 3) showbusiness and entertainment; 4) separate management which includes all atypical workers. Unfortunately, public employees[9], professionals registered with their own pension fund and workers employed in the informal economy are not included in the archive. For each spell, INPS collects starting and ending dates and, for private sector employees only, some main features such

[9] The INPDAP, the National Institute for Social Security and Assistance for Public Administration Employees, was abolished, and its functions were transferred to the INPS as of January 2012. At the time of extracting the INPS data (2015), information on public sector employees was not yet available.

as the type of work, salary, weeks of contributions paid. It should be noted that for all contributions belonging to the separate management, starting and ending dates are conventionally fixed at the 1st January and the 31st December respectively. Since 1974, for most of the employees, there is also a unique number identifying the company (*matricola*) that made it possible to link, for each contribution spell, the fiscal code of the company, its economic activity code, address, date of establishment and (potential) closure.[10]

In addition to the contributions paid by employers in the case of employees or by the self-employed themselves, the archive managed by INPS also includes non-working spells called imputed contributions (*contribuzioni figurative*) that include solidarity contracts, unemployment and mobility benefits.

4. THE INTEGRATION PROCESS

4.1 RECORD LINKAGE

The aim of record linkage was to identify pairs of records in the two sources using the individual fiscal code as the matching key.

We successfully matched all the individual records of the wave 2012 of the ICFT survey (N=5,733) with INPS data (Table 1). The result was a combined database of 5,489 adults with at least one working episode, in the survey and/or in the register. The vast majority of the sample (4,551 subjects) was in both databases. Then, 507 individuals in the survey were not found in INPS database since their jobs were not recorded, by definition, in the administrative data and 431 subjects didn't declare any job experience in the survey even though they had one. Finally, 244 subjects were excluded because they didn't have any work episodes.

[10] These data come from the E-Mens flows.

Table 1: Sampling of individuals after the matching

	N
Individuals over 18s interviewed in ICFT survey (wave 2012) [a]	5,733
• With at least one working episode in ICFT survey	
o and at least one episode in *INPS*	4,551
o but not in *INPS*	507
• Without any working episode in ICFT survey	
o But at least one episode in *INPS*	431
o And any working episode in *INPS*	244

[a] Net of 23 individuals with unknown fiscal code.

4.2 DATA PROCESSING

The next step was to integrate the information contained in the two databases. However, the study confronted us with the dual need to address different issues of definition and to develop multiple procedures for handling the data.

In particular, while the ICFT was explicitly designed to study work stories, processing register data was quite demanding as the archive meets the Institute's management purposes and is therefore not immediately usable by researchers. The register included about 158,000 records, 124,000 of which were work records and the rest were non-work records. In order to reconstruct the working life of each individual in the sample, the following steps have been followed.

First of all, the type of contribution was used to identify work episodes and, in particular, to classify all records belonging to a fund under the following employment conditions: "Employees in the private non-agricultural sector", "Artisans, retailers and entrepreneurs", "Freelance professionals", "Temporary workers and other self-employed individuals", "Agricultural workers (employees and self- employers)". Within the periods of imputed contributions, the time segments of unemployment benefits in the strict sense (ordinary unemployment benefits, mobility benefits, ASpI and MiniASpI) were identified[11].

[11] Register data does not provide complete information on non-employment spells but just on periods of subsidized unemployment which may include subsidised unemployment in the strict sense (ordinary unemployment benefit, mobility allowance, ASpI and MiniASpI) and other subsidised non-employment (reduced unemployment,

In order to have adequate statistical data to represent the work histories, it was necessary to carry out a meticulous activity of refinement of the administrative information since, according to the law, the same episode (both of employment and subsidized unemployment) could appear in the archive in more than one record. In particular, we addressed three main aspects:

i) duplicate episodes or events with no contributing weeks;
ii) segmented episodes;
iii) episodes of spurious mobility.

With the first cleaning procedure we ignored all those episodes, both of work and non-work, which, for administrative reasons, were *duplicates* or had *zero contribution weeks* (overall about 4,000 records). In particular, the former operation concerned episodes of non-agricultural employment bearing the identification of the company[12], self-employment (traders, craftsmen, entrepreneurs), agriculture (self-employed and employees) and episodes of subsidised unemployment stricto sensu. In the case of coincidence between start and end dates and, for work episodes, type of contribution and company, only one episode was maintained.

The second intervention concerned the *segmented work episodes*, i.e. the (quite frequent) cases in which a working episode within the same company was recorded as if it were a plurality of different episodes with different but consecutive starting and ending dates.[13] In this case, the records were combined into a single work episode. The number of records was reduced by approximately 80,000 cases.

Finally, we detected the firms' legal transformations to handle spurious movements of workers. As is well known, firms can undergo changes over time: in their identifying characteristics (name or address), structural characteristics (size or prevailing activity) or in cases of mergers, demergers or deaths. In many cases, these changes do not result from real changes but from the adaptation of legal entities to laws and administrative rules. However, all these events, even

agricultural and construction). The latter have not been considered here because, conventionally, annual periods starting in January and ending in December are reported, thus overestimating the actual duration; moreover, the monetary transfer is also compatible with the inactive condition.

[12] No intervention has been carried out on non-agricultural employees with missing information on the company.

[13] Not necessarily from January to December.

those which, in fact, did not generate real discontinuities in the life of companies, were recorded in the administrative archives generating the so-called spurious flows.

For the study of work histories, it was, therefore, crucial to discriminate between movements originating from choices of workers and spurious movements that were induced by changes that had affected companies. The possibility provided by the administrative archive to observe both the individual employees and the companies for which they work, enabled us to deduce the existence of other underlying legal transformations from the simultaneous flow of workers between two companies, in order to discriminate between real movements (i.e. resulting from workers' decisions to change jobs) and spurious movements.[14] If we had ignored the specific nature of these flows, we would have observed a fictitious variation in the number of jobs.

In order to detect the presence of firms' legal transformations we proceeded as follows. Firstly, we considered all the episodes of private sector employees belonging to the regular fund (about 25,000). Then we identified the direct movements from one firm to another (the so-called job-to-job) and then we selected those that occurred over a period of 30 days (about 5,000).

An ad hoc employee/employers' database was then created containing, for each individual with at least one job-to-job within 30 days, the characteristics of the starting company and those of the ending one. Two criteria were then used to detect spurious moves:

i) *Similarity of attributes*: a link was established between two companies with different identification code (matricola) according to the similarity in the values of certain characteristics and, in particular, their *name* and *address*;

ii) Identification of firms' *legal transformations* through workers' movements: the existence of transformations was deduced from the flow of employees.

As regards the first criterion, movements between companies with the same *name* were initially candidates to be considered as spurious. This was the case

[14] The identification of spurious mobility relationships could only be carried out for employees with company identification. Therefore, in addition to working episodes prior to 1975, all episodes that fell under the separate management, in showbusiness and entertainment and clergy (as well as those of self-employment) were excluded. Given their peculiarity, episodes of temporary agency work of usual seasonal work were also excluded. Altogether we ignored from the analysis of spurious mobility about 45,000 episodes.

for companies with several operating units (e.g. Unicredit). Overall, 206 job-to-jobs have been initially identified using this criterion.

In addition to the name, a second characteristic that was used to establish a link between two companies was the *address*. However, considering the fact that the same address can be written in different ways (Via S. Croce 77, Trento vs. Via Santa Croce 77, Trento), we preferred to use geocoding via Google Maps. Based on the consistency between the geographical coordinates, the economic activity code and the difference between the date of cessation of the company of origin and the date of establishment of the company of arrival (delta), the following levels of quality of the combination have been defined:

- High level (5): companies with the same address, economic activity code equal (to the 3rd digit) and delta less than 30 days.
- Medium-high level (4): level 5 without an economic activity constraint.
- Medium level (3): level 5 but with a maximum discrepancy in the geographical coordinates of the two companies of 0.0001 points (corresponding to about half a block: an example is a building with several house numbers).
- Medium-low level (2): level 3 without constraint of economic activity
- Low level (1): maximum discrepancy in the geographical coordinates of the two companies of 0.001 points (about one block), economic activity code at the 2nd digit, time difference < 30 days.

The same ranking was also assigned when the date of cessation of the company of origin and/or the date of establishment of the company of arrival was missing. The table summarizes the number of job-to-job movements identified with geolocation (174) according to the quality of the combination (Table 2). For all of these cases, a manual check was then carried out.

Turning to the second criterion (firms' legal changes), in order to identify the existence of corporate transformations, we first looked at the number of individuals who made the same movement. Since we were observing a sample of about 5,000 workers, if there were, in the sample, at least three individuals with the same movement, it seemed reasonable to assume that it was spurious mobility. All these cases (88) therefore passed directly to manual check.

Table 2: Number of job-to-job according to the level of precision of the geolocalisation and the completeness of the available information

Geolocalisation quality	Cases with complete information	Cases with incomplete information [a]
High	42	69
Medium-high	7	33
Medium	2	1
Medium-low	0	3
Low	9	8
Total	174	

[a] Cases where the date of cessation of the company of origin and/or the date of establishment of the company of arrival were missing.

In cases where job-to-jobs occurred in the same month or on adjacent months (113 cases), a different procedure was followed based on variations in the size of the companies and, more precisely, on the monthly series of employees. The algorithm used for this purpose took into account the change in the number of employees at the time of the change of business: the delta of the company of origin (Δ company_from) corresponds to the difference in the number of employees between time t (where job-to-job occurs) and time t-1; likewise, the delta of the company of arrival (Δ company_at) corresponds to the difference in the number of employees between time t and time t+1.

The underlying logic is that the more similar the Δ were, the more likely it was that there were spurious mobility due to corporate transformations (such as spin off, company acquisitions, company branch acquisitions).

The cases identified (87) varied depending on the number of employees in the company of arrival (Nempl_at):

- |(Δ company _from – Δ company _at)|=< 2 if Nempl_at < 20
- |(Δ company _from – Δ company _at)|=< 5 if Nempl_at between 20 and 49
- |(Δ company _from – Δ company _at)|/ Nempl_at =< 0.1 if Nempl_at >= 50

Finally, cases of closure of the company of origin were isolated if the number of employees of the company of origin (Nempl _from) increased from N to 0 (6 cases):

- Δ company _from = Nempl _from

To sum up, the similarity of the attributes made it possible to identify 380 cases of probable spurious mobility, 206 on the basis of the name and 174 via

geolocation, respectively. On the other hand, 181 cases were identified from the analysis of workers' flows.

Considering that some cases were identified by more than one criterion, 476 company movements have been checked manually by consulting the websites of the companies involved (Table 3). Overall, 91.4% of the cases were positive, i.e. they confirmed the hypothesis of spurious mobility. The identification of 435 spurious company transfers made it possible to correct (and therefore not consider in the calculation of work mobility) 951 work episodes, corresponding to 3.8% of all the episodes in the private sector belonging to the regular fund.

Table 3: Number of possible spurious movements according to the criterion used to identify them and the result of the manual check.

Criterion	Manual check		N
	negative	positive	
Denomination	0	142	142
Geolocalisation	26	105	131
N. of employees	15	106	121
Denomination + Geolocalisation	0	22	22
Denomination + N. of employees	0	39	39
Geolocalisation + N. of employees	0	18	18
Denomination + Geolocalisation +N. of employees	0	3	3
Total	41	435	476

In conclusion, in order to create a database that would meet the purposes of analysing work histories, the procedures described above led to a reduction of about two thirds of the number of episodes that were originally in the administrative archive (from 158,000 to about 55,000).

4.3 THE MATCHING

As stated in Section 2 several content issues need to be considered while working with administrative and survey data (i.e. data definitions, consistency between different time periods, coverage in the administrative data system that may be subject to discontinuity due to changes in the legislation or in administrative practices).

In our case, the first important discrepancy concerned the *employment status* as the ICFT survey recoded all employment conditions, while INPS did not

include civil servants, freelancers with their own social security fund and, obviously, illegal workers (Figure 1).[15]

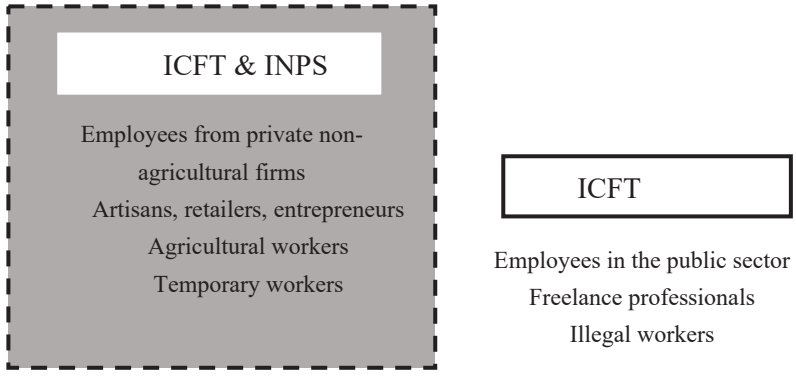


Figure 1: Type of occupations in the two sources of data

The second discrepancy concerned the *unemployment situation*. Determining unemployment episodes was problematic, for two reasons: the first one involved the difficulty to identify unemployment, as highlighted nearly fifty years ago by Shiskin (1976) in a paper with the paradigmatic title – [measuring] the Doughnut [of employment] or the Hole [of unemployment]? – and recurrent in the economic literature up to Brandolini and Viviano (2016). The second reason concerned the narrow perimeter of subsidized unemployment and the approximate way in which unemployment is recorded in the ICFT survey. Unlike in INPS, in the survey all those who declared themselves in search of a job were classified as unemployed (Figure 2). The survey therefore covered both those who were looking for their first job and those who were looking for a new job but without receiving any benefit.

The third factor of dissimilarity between the two sources concerned the *timing* of the observations: in INPS the episodes were documented with units of time at the day level (except for what has been said for the members of the separate fund), while in ICFT the segments of work histories declared by the interviewees were recorded with units of time by the month.

[15] Another possible underestimation is that of co.co.co and co.co.occ since for the former until January 1996 there was no obligation to register and for the latter it depends on the self-declaration of the worker.

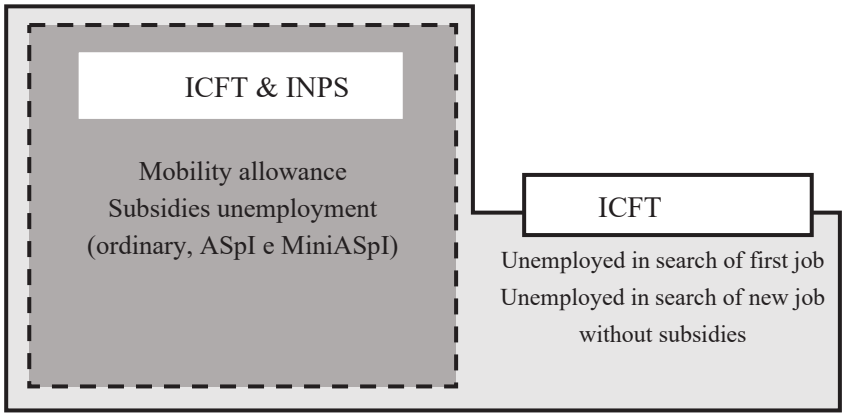


Figure 2: Non-working conditions in the two sources of data

The fourth and last source of distortion was linked to the fact that, while in the survey the prevailing condition was recoded excluding the *possibility of overlap* between episodes, in INPS it was possible to find (partial or complete) overlapping episodes.

In order to overcome the differences described above, when integrating the information taken from the two sources, a set of criteria was followed that privileged the employment condition and, overall, gave prevalence to the administrative source[16] (Figure 3).

As regards the working histories, the INPS source was treated as:

a. exclusive in the case of non-agricultural private employees.[17] It is worth mentioning that the number of episodes of this type detected in the survey was approximately one third of the total episodes.

[16] This in order to limit the possible response errors related to the survey.

[17] It should be noted that for 71 people who declared themselves to be non-agricultural private employees in the survey, no such episodes were found in INPS. In particular, 23 cases do not even appear in the administrative file, so their episodes of employment have been maintained but labelled as "illegal work". In the remaining cases, it was decided to eliminate only those episodes of non-agricultural private employment detected by interview if, in the same year, there was an episode of work in another category in INPS (thus assuming the existence of classification errors in the survey).

b. initial source, supplemented by ICFT, in the case of agricultural workers (self-employed and employees), craftsmen and traders and entrepreneurs (assuming that their absence was due to non-payment of contributions).

In the remaining cases, we resorted to:

c. ICFT data as an exclusive source for work episodes not to be found, by definition, in INPS (workers in public administrations and illegal workers);
d. Priority to ICFT for episodes of freelance professionals and temporary workers: (d1) the former because, if they had their own cash, they were present only in the ICVT; (d2) both because they could be found in INPS with significant information limits on the start and end dates.[18]

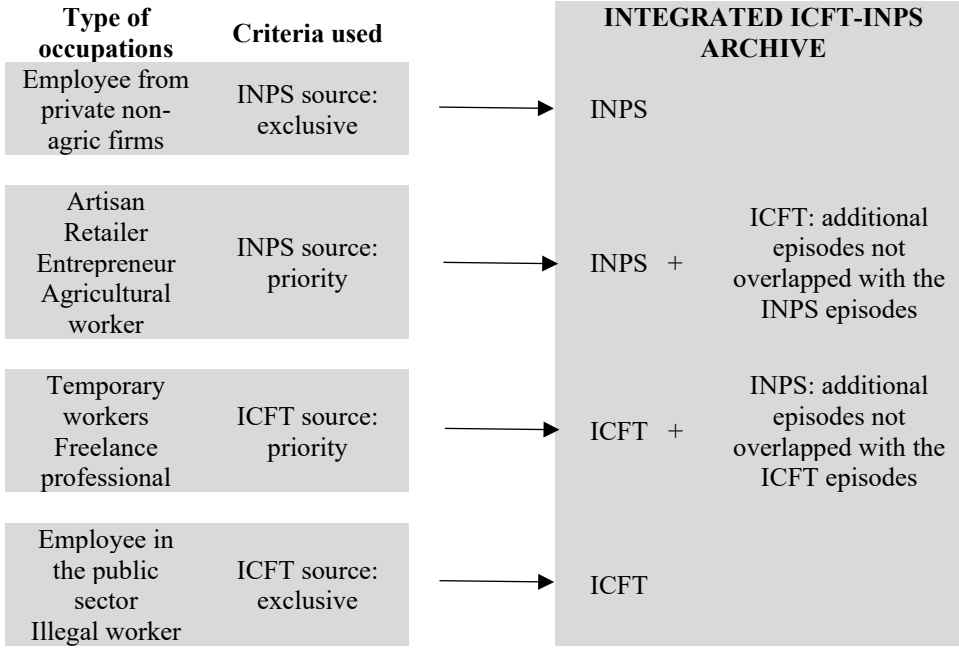


Figure 3: Data integration: exclusiveness or priority criteria used

[18] It should be noted that all the episodes belonging to the separate management, the beginning and the end, are conventionally equal respectively to 1 January and 31 December.

As for the unemployment episodes, the INPS priority criterion was used giving priority to the subsidised unemployment episodes. The integration with ICFT data was then used only if – or, in any case, for the part in which - the episode did not overlap with the periods of work as described above and with the periods of subsidized unemployment. In particular, two situations were identified. In cases (1,764) in which declared unemployment fell during a period in which neither work nor non-work episodes were found in INPS, a new episode of unemployment was added. In cases (193) where a period of declared unemployment was adjacent to a period of subsidised unemployment, the latter was extended to include the non-subsidised part.

Overall, the ICFT-INPS integrated archive comprised 5,489 individuals with at least one working episode. The working episodes were 48,580, of which 42,408 from the administrative archive, and 6,172 from the survey (Table 4). The unemployment episodes were, in total, 6,804:5,021 from INPS and the remaining from ICFT.

Table 4: Employment and unemployment episodes in the combined archive

Type	With no overlapping spells			Including overlapping spells		
	INPS	ICFT	Total	INPS	ICFT	Total
Employee from private non-agr. firms	29,648	0 [a]	29,648	34,959	0[a]	34,959
Employee in the public sector	n.d[b]	3,172	3,172	n.d[b]	3,367	3,367
Artisan, retailer and entrepreneur	1,379	188	1,567	1,465	204	1,669
Freelance professional	231	122	353	324	127	451
Temporary worker and other self-employed	1,743	249	1,992	3,205	263	3,468
Agricultural worker (employee & self-employer)	2,210	507	2,717	2,455	592	3,047
Illegal worker	n.d[b]	1,553	1,553	n.d[b]	1,619	1,619
Total employment spells	35,211	5,791	41,002	42,408	6,172	48,580
Total unemployment spells	5,000	1,783	6,783	5,021	1,783	6,804

[a] For the episodes of non-agricultural dependent work, a criterion of exclusivity of the administrative data was chosen, ignoring the episodes of that type declared in the survey.
[b] The episodes of public employment and "illegal" work were not included in INPS.

The table shows the number of episodes including overlapping spells and also those net of any overlaps between episodes of employment (double jobs) and between employment and unemployment. The criteria followed to solve overlapping spells were the following:

i) Where a working episode was completely contained in another more extensive working episode, the first was deleted.
ii) In cases of partial overlap between working episodes, a hierarchical order was first established between the types of work: private employee work (including agricultural work), public administration employee work, self-employment (including agricultural work), collaboration and undeclared work. On the basis of this, the starting or ending date of the lower hierarchical episode was changed.
iii) In the case of (non- part-time) employment overlapping with subsidised unemployment, unemployment had prevailed. The work episode was removed only if it was fully included in the unemployment episode. If, on the other hand, the unemployment episode was partially overlapped with a work episode, the latter was broken into two separate episodes. In case of partial overlap to the right or left of the work episode, the end or start date of the work episode was changed.
iv) In the case of overlap between part-time work and subsidised unemployment or mobility, work prevailed.[19]

Through the resolution of cases of overlap between episodes, the number of episodes was reduced by about 7,600 cases.

Regardless of whether or not the episodes overlap, it is worth mentioning that in addition to the private non-agricultural employees for whom the administrative data was considered an exclusive source, the contribution of the INPS data in the final integrated database was also dominant in cases where the survey data was considered a priority, and particularly in the case of collaborations and the liberal professions (Table 3).

The reasons may be twofold. First of all, as mentioned above, the survey recorded the prevailing working condition, therefore, in the case of double jobs (for example, an employee with a collaboration or the case of a teacher with

[19] The logic followed is the same as that described in the previous note with the only difference that, in this case, what is modified is the unemployment episode.

VAT) the respondents tended to declare the most important or the most stable over time.

The second reason is due to the fact that in the survey the work career was reconstructed retrospectively so the further away the event to be remembered, the more difficult it was for the respondent to identify short-term or occasional events.

The integration with the administrative data therefore had the undoubted advantage of correcting, at least for some types of episodes, the timing errors, completing the omissions of events and, in general, enriching the information base for the study of the dynamics of the working histories.

5. HIGHLIGHTS OF KEY FINDINGS

While the focus of this paper was to describe the construction and features of the data archive, it is worth briefly recalling some of the main findings derived from its analysis. These findings are explored in greater detail in a separate contribution, to which we refer for further discussion (Bazzoli et al. 2018).

The archive allowed to study changes in the work histories of young people over the last forty years. By comparing two cohorts - one composed of individuals born between 1959 and 1966, and the other of those born between 1975 and 1982 – we found an overall deterioration in working conditions for younger generations.

The average age of labour market entry increased from 18.2 years in the older cohort to 19.3 years in the younger one. This increase is particularly marked for women, who enter the labour market almost two years later than men in the most recent cohort. This delay is closely linked to longer education pathways, which have seen a marked improvement in educational attainment. In the younger cohort, women are on average more educated than men, reversing the trend observed in the previous generation.

Despite the improvement in education levels, the stability of first jobs has drastically decreased. The average duration of the first job declined from 23.6 months in the older cohort to 13.7 months in the younger cohort, with a particularly pronounced contraction among university graduates. The latter face the greatest difficulties in maintaining their first job, a phenomenon attributable to both their higher expectations in terms of job quality and pay, and the limited growth in demand for highly skilled labour. This reduction in first-job stability

affects both genders, although it is more pronounced for men, consistent with the changes in the sectoral distribution of first jobs.

Work histories of the younger cohort are also more fragmented and characterized by greater job mobility. During the first eight years of their careers, young people in the recent cohort experience a higher number of employment episodes compared to the older cohort, but with shorter durations. Pathways are particularly unstable for graduates and high school diploma holders, who face a growing number of fixed-term contracts and more limited access to stable positions. However, this increased mobility does not translate into higher career mobility, which remains low for both cohorts, highlighting persistent structural rigidity in the Italian labour market.

The comparison between the cohorts also reveals a polarization in types of employment. In the younger cohort, over 80% of young people begin their careers in the private sector, while the share of those entering public employment or self-employment has drastically declined. Women, compared to men, continue to have a higher likelihood of starting their careers in the public sector, although this tendency has diminished in the most recent generation. At the same time, the younger cohort faces a significantly higher risk of losing their first job, with a 57% increase compared to the older cohort. This risk is particularly high for more educated youth and those from families of dependent workers, whereas the children of self-employed workers demonstrate greater stability.

In summary, the comparison between the two cohorts highlights a deterioration in working conditions for younger generations. Despite higher levels of education, young people in the recent cohort face more unstable, fragmented, and precarious career paths, with increasing vulnerability in the labour market. This scenario reflects growing difficulties in aligning young people's qualifications with available opportunities, underscoring the structural challenges of a labour market characterized by polarization and inequalities.

6. CONCLUSIONS

As a result of the combination of restrictive interpretations of legislation already restrictive in itself and, even more so, the absence of an adequate scientific and political culture, the huge set of information contained in the archives of the Italian public administration has long remained unproductively hidden and only recently has begun to be used, albeit sporadically.

However, it should also be stressed that administrative archives alone are not always able to answer the research questions of scholars. Quite understandably these archives contain information about individuals that is segmented and filtered in the light of institutional needs. Moreover, they lack other significant data which, in many ways, are crucial in order to understand the configuration of what is contained in the administrative archive and, above all, to be able to use that data in order to carry out analyses and to evaluate public policies. To overcome the problem of segmentation of information, a way that has long been practiced in the Scandinavian countries already mentioned, but which is difficult to implement in Italy, is to combine complementary administrative archives. However, even this procedure, if it is truly feasible, is not always adequate to answer the research questions that arise from scientific research and the political sphere. An innovative and promising procedure for overcoming these possible further inadequacies is to integrate administrative archives with large-scale sample surveys using IT tools. This is the context in which the project on the working histories presented in this paper has started.

The integration between the INPS data and the ICFT panel survey as a tool for socio-economic analysis has confronted us with the multiple sources of potential distortion in the data of the two sources, the difficulties in identifying them, as a result of the myriad obstacles encountered in the integration at the micro scale. Often simplified solutions have been adopted (but plausible for the analysis of the changes also because it is reasonable to believe that the distortion is invariant - or exhibits low variability - over time). The analysis, however, has provided important insights into the deterioration of labour conditions for younger generations, the increasing mismatch between education levels and labour market opportunities, and the structural rigidity of the Italian labour market.

While challenges remain, this study represents a meaningful advancement in the integration of data for research and policy analysis in Italy. It contributes valuable methodological insights and demonstrates the potential of integrated datasets to improve our understanding of socio-economic dynamics.

REFERENCES

Al Baghal, T., Knies, G., Burton, J. (2014). Linking administrative records to surveys: differences in the correlates to consent decisions. *ISER, Understanding Society Working Paper Series* 9.

Bazzoli, M., Marzadro, S., Schizzerotto, A.,Trivellato, U. (2018). Come sono cambiate le storie lavorative dei giovani negli ultimi quarant'anni? Evidenze da uno studio pilota. *Stato e mercato*, 38(3), 369-418.

Brandolini, A., Viviano, E. (2016), Behind and beyond the (head count) unemployment rate. *Journal of the Royal Statistical Society*: Series A, 179(3): pp. 657-681.

Card, D., Chetty, R., Feldstein, M. S., Saez, E. (2010). *Expanding Access to Administrative Data for Research in the United States*. American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas.

Calderwood, L., Lessof, C. (2009). Enhancing longitudinal surveys by linking to administrative data, in P. Lynn (Ed.), *Methodology of Longitudinal Surveys*, Chichester: Wiley: Ch. 4.

Connelly, R., Playford, C.J., Gayle, V., Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research,* 59, September: 1-12.

Conti, R., Bavaro, M., Boscolo, S., Fabrizi, E., Puccioni, C., Ricchi, O., Tedeschi, S. (2023). The Italian treasury dynamic microsimulation model (T-DYMM): Data, structure and baseline results. *Government of the Italian Republic (Italy), Ministry of Economy and Finance, Department of the Treasury Working Paper*, (1).

Crato, N., Paruolo, P. (2019). *Data-Driven Policy Impact Evaluation: How Access to Microdata Is Transforming Policy Design* (p. 346). Springer Nature.

Einav, L., Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1), 1-24.

Fabrizi, E., Rocca, A. (2024). NEET Status Duration and Socio-Economic Background. *Socio-Economic Planning Sciences*, 101986.

Fambri, G., Schizzerotto, A. (2008), a cura di. *Le condizioni di vita delle famiglie trentine. Rapporto di ricerca*, Quaderni della Programmazione n. 21, Provincia Autonoma di Trento e Università degli Studi di Trento, Trento: Edizioni 31.

Fondazione G. Brodolini, MEF "IESS Improving Effectiveness in Social Security Final Report" https://www.dt.mef.gov.it/export/sites/sitodt/modules/documenti_it/analisi_progammazione/analisi_programmazione_economico/final-report-IESS.pdf

Harron, K.L., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, L.M, Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, July–December: 1–12.

Istat (2009). Integrazione di dati campionari Eu-Silc con dati di fonte amministrativa. *Metodi e Norme* n. 38, Roma.

Künn S. (2015). *The Challenges of Linking Survey and Administrative Data*. IZA World of Labour, 214.

Livraga, G. (2019). Privacy in microdata release: Challenges, techniques, and approaches. *Data-Driven Policy Impact Evaluation*. Springer, 67-83.

Naticchioni, P., Raitano, M.,Vittori, C. (2016). La Meglio Gioventù: Earnings gaps across generations and skills in Italy. *Economia Politica*, *33*, 233-264.

Sakshaug, J. W., Couper, M. P., Ofstedal, M. B., Weir, D. R. (2012). Linking survey and administrative records: Mechanisms of consent. *Sociological Methods & Research* 41(4): 535–569.

Sestito, P., Trivellato, U. (2011). Indagini dirette e fonti amministrative: dall'alternativa all'ancora incompiuta integrazione. *Rivista di Politica Economica*, luglio-settembre: 183-227.

Shiskin, J. (1976), Employment and Unemployment: The Doughnut or the Hole?. *Monthly Labour Review*, 99(2): 3-10.

Trivellato, U. (2019). Microdata for social sciences and policy evaluation as a public good. In Crato N. and Paruolo P. (Eds).