Statistica Applicata – ITALIAN JOURNAL OF APPLIED STATISTICS

EDITORIAL TEAM

A.S.A CONTACTS
Principal Contact
Francesco Palumbo (Editor in Chief)
editor@sa-ijas.org

Support Contact
Domenico Vistocco (Editorial Manager)
ijas@sa-ijas.org

JOURNAL WEBPAGE
https://www.sa-ijas.org/ojs/index.php/sa-ijas

Statistica Applicata – Italian Journal of Applied Statistics is a four-monthly journal published by the Associazione per la Statistica Applicata (A.S.A.), Largo Gemelli 1 – 20123 Milano, Italy (phone + 39 02 72342904). Advertising:Cleup sc, via G. Belzoni, 118/3 – 35128 Padova, Italy (phone +39 049 8753496 – Fax +39 049 9865390), email: info@cleup.it.
Rules for manuscript submission: https://www.sa-ijas.org/ojs/index.php/sa-ijas/about/submissions
Subscription: yearly €103.30; single copy €40.00; A.S.A. associates €60.00; supporting institutions: €350.00. Advertisement lower than 70%. Postal subscription Group IV, Milan. Forum licence n. 782/89. CLEUP SC on behalf of ASA, 7 March 2023.

Statistica Applicata – Italian Journal of Applied Statistics is associated to the following Italian and international journals:

QTQM – Quality Technology & Quantitative Management (http/web.it.nctu.ew/~qtqm/)

SINERGIE – Italian Journal of Management

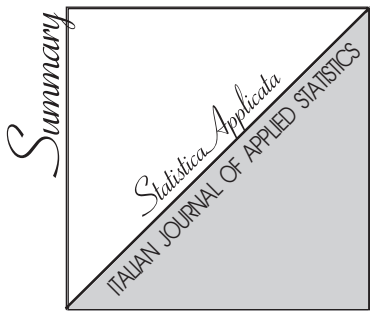
Statistica Applicata – Italian Journal of Applied Statistics (ISSN:1125-1964, E-ISSN:2038-5587) applies the Creative Commons Attribution (CC BY) license to everything we publish.

Pubblished: November 2024

Published by Firenze University Press and Cleup
Powered by Firenze University Press

Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

Cleup sc
'Coop. Libraria Editrice Università di Padova'
via G. Belzoni, 118/3 – Padova Italy
Phone +39 049 8753496 Fax +39 049 9865390
info@cleup.it – www.cleup.it – www.facebook.com/cleup



Vol. 36, Number 3

doi.org/10.26398/IJAS.281

THE USE OF ONLINE JOB ADS TO ANALYSE SKILL CHANGES IN ICT OCCUPATIONS

Adham Kahlawi, Laura Grassini and Lucia Buzzigoli
Department of Statistics, Computer Science, Applications, University of Florence, Florence, Italy

Abstract. *Online job ads are a timely and detailed data source that can be used to get in-depth information about occupations requested by employers, together with the related skills, and to monitor the time evolution of skills demand in the different occupations, even at a detailed territorial level. The paper describes the more recent dynamics in skills demand in Italy for two innovation-related groups of occupations, using a measure of skill change between 2019 and 2021 at the regional level.*

Keywords*: Labour market, Online job ad, Occupation, Skill, Skill change.*

1. INTRODUCTION

Online Job Ads data (OJAs) is gaining popularity in labour market research due to its ability to provide valuable and timely insights into job offers and specific skill requirements across various levels such as territorial or sectoral contexts (Beręsewicz and Pater, 2021).

Of particular significance is the utilisation of OJAs to study and monitor the development of Information and Communication Technology (ICT) occupations and their related job markets, as they provide a channel that aligns well with the interests of potential candidates. Moreover, OJAs provide highly detailed and up-to-date information on specific and innovative skills, such as those in the ICT domain, which are crucial for employers (Aica et al., 2019).

Undoubtedly, ICT has had a profound impact on business processes, tasks, and organisations in various economic activities (López Cobo et al., 2020). The multifaceted roles of ICT professionals, including research, planning, information systems maintenance, and safeguarding data integrity and security, necessitate a wide range of specialised occupational skills. As technological advancements continue to occur at a rapid pace, these skills are becoming increasingly pervasive throughout the economy, making it essential to study the

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.
Competing Interests: The Author(s) declare(s) no conflict of interest.

evolution of skills in these occupations. In Italy, the Observatory of Digital Skills confirms this trend and uses OJAs to analyse the changing landscape of ICT professions in response to new technological trends. The Observatory produces annual reports to track these changes[1].

Eurostat's European survey on ICT usage and e-commerce in enterprises provides official statistics that shed light on the recruitment of ICT specialists in Italy and the European Union[2]. During 2012-2022, the number of employed ICT specialists in Europe increased by 57.8%, almost seven times higher than the increase (8.8%) in total employment. The estimate of employed ICT specialists in Italy as a percentage of total employment is 3.5 in 2019 and 3.8 in 2021, not far from the EU-27 values (respectively 3.9 and 4.5), but much smaller than in northern countries, like Sweden (7.0 and 8.0) or Finland (7.6 and 7.4) (Eurostat, 2023a). On the other hand, the percentage of European Union enterprises recruiting or attempting to recruit ICT specialists in 2021 was 9.5% (4.9% in Italy), and 62.8% of these (61.3% in Italy) had difficulties in recruiting them (Eurostat, 2023b). Cedefop (2023) predicts a strong growth in the demand for ICT professionals up to 2035, while the employment shares of technicians are expected to decline.

The market for these occupations is, therefore, very dynamic, and it is expanding rapidly, also following the significant investments in the digital transition envisaged by the National Recovery and Resilience Plan[3] (PNRR) implemented in Italy through NextGenerationEU funds.

This paper uses the Lightcast dataset[4], which collects online job postings describing occupations and related skills with variables referred to official classifications. The aim is to analyse the more recent dynamics in skills demand in Italy for occupations with the highest concentration of ICT specialists and apply a measure of skill change between 2019 and 2021 at the regional level. Other research contributions have studied OJA data for Italy but are not concerned with comparisons of skill change across regions (among others: Lovaglio et al., 2018; Vannini et al., 2019; Kahlawi et al., 2023; Lucarelli and Righi, 2023).

[1] https://www.anitec-assinform.it/pubblicazioni/studi/. Last access December 2023.
[2] According to the EUROSTAT Glossary (last access May 2023) ICT specialists belong to the ISCO-08 occupation groups 133, 25, 35 and to other unit groups that involve the production of ICT goods and services.
[3] https://www.italiadomani.gov.it/it/home.html. Last access September 2023.
[4] Source: Lightcast™ 2022.

The study extends the line of research initiated by Giambona et al. (2021) in two directions.

Firstly, the temporal comparison involves the pre- and post-pandemic situations (2019-2021). As shown in Unioncamere and ANPAL (2022), the COVID-19 pandemic has forced Italian businesses to adopt and use digital technologies at a faster pace. This has increased the awareness of the need to rethink their business models and move towards greater digitalisation. The community and national-level measures to respond to the crisis have further accelerated the adoption of digital technologies. The Next Generation EU program has explicitly prioritised the digital transition. Despite this, some structural difficulties still hinder a broader digital transformation of the Italian production system. These difficulties are primarily related to production specialisation, governance, and company size. However, Italian companies have been progressing in integrating digital technologies into their business processes, especially recently. Therefore, we can expect a not well-defined picture and, consequently, a not well-defined pattern for skill changes.

Secondly, a deeper analysis of the skill change is presented, putting in evidence negative and positive variations at regional and occupation levels and exploring different types of skills. The results also show the different skill dynamics in two ICT occupation groups: *Professionals* and *Technicians*.

The structure of the paper is the following: the next section discusses the characteristics, pros and cons of OJA data for statistical labour market analysis. Section 3 presents a descriptive analysis of the Lightcast data. Section 4 focuses on methodology. Section 5 shows the results and a discussion, while the final section provides concluding remarks.

2. OJA DATA IN LABOUR MARKET INFORMATION SYSTEMS

Over the past decades, technological advancements, digitalisation, globalisation, and environmental concerns have continuously evolved the labour market (Arregui Pabollet et al., 2019).

In this ever-changing context, information that aids policymakers and decision-makers in tracking market trends and facilitating the alignment of labour demand and supply is precious. Online Job Ads data has garnered increasing attention due to its timely availability and detailed insights into job demand, which enrich traditional labour market data sources derived from

official statistics. Furthermore, OJAs that include education requirements hold significant potential as supportive tools in higher education policy, facilitating the implementation of effective training programs (OECD, 2020).

OJAs are therefore considered a valuable component of labour market information systems, encompassing data on the size, composition, functioning, problems, opportunities, and employment-related intentions of the labour market and its participants (ILO, 2001) and employing intelligent approaches to support operational and decision-making activities (Mezzanzanica and Mercorio, 2019).

However, it is well known that the use of OJA data poses challenges related to their representativeness, reliability, and overall quality (Beręsewicz and Pater, 2021; Cammeraat and Squicciarini, 2021; Cedefop, 2019; ILO, 2020; Fabo and Kurekova, 2022; Napierala, 2022). Creating standardised datasets from individual job ads published online faces technical and contextual complexities, impacting the data's potential applications (Giambona et al., 2021). The need for data cleaning further adds to the complexity, while duplications of the same job announcement across multiple portals introduce noise into the final dataset. Moreover, the unavailability of standards for occupational skill descriptions leads to heterogeneity in employer-provided skills listings.

Notably, OJAs do not fully represent all job openings, as some vacancies may not be posted online. Certain countries tend to overrepresent high-skilled occupations and the sectors requiring them, while low-skilled jobs may be underrepresented (Carnevale et al., 2014; Schmidt et al., 2023). The prevalence of other recruitment methods, regional variations in internet usage for job advertisements, and varying employer preferences further contribute to representativeness issues.

Additionally, not all published job ads correspond to actual vacancies, with companies exploring the labour market without necessarily intending to recruit new resources. Comparisons between web vacancy postings and official vacancy statistics have been explored in previous studies (de Pedraza et al., 2017; Lovaglio et al., 2020), with OJAs also complementing classic survey-based approaches for economic statistics (Turrell et al., 2019). However, European official job vacancy statistics are only available at the national level, and for Italy, only rates are released. In any case, the OJAs remain a valuable source of information for studying specific skills associated with job demand with a granularity that cannot be deduced from official sources. That also justifies the

interest Istat (the Italian National Institute of Statistics) recently expressed for this data type (Vannini et al., 2019; Lucarelli and Righi, 2023).

Contextual factors, such as the features of web job portal systems, the role of public employment services, the level of digitalisation, and demographic and economic structures, are other essential issues to consider while analysing OJA data for territorial analysis. These factors may influence the scope and limitations of international and regional comparisons in countries marked by socioeconomic and digital divides (ILO, 2021). For instance, in Italy, the percentage of individuals using the Internet to look for a job or send a job application[5] is 13.6 in 2019 and 13.7 in 2021, in line with the EURO-27 average (respectively 15.63 and 13.42), but much less than the Nothern European countries (e.g., Denmark presents 37.6 and 36.3; Sweden 32.1 and 34.2).

On the other hand, the OJAs are the only available source that provides detailed information regarding the skills required by employers, although not included in official statistics.

Nowadays, there is great interest in the study of skills (see, above all, the European Agenda for Skills, a five-year plan to help individuals and businesses develop more and better skills and put them to use[6]). The anticipation of skills demanded by the labour market and the definition of consequent intervention strategies can help to identify appropriate measures in the fields of work, education, industrial and local development.

3. LIGHTCAST DATA

In this paper, we use the dataset provided for Italy by Lightcast, previously Burning Glass Technologies and Emsi Burning Glass, which collects online job postings deriving from numerous Internet sources (both online job portals and company websites) that have undergone a cleaning process to remove noise, outliers and duplications (Lightcast, 2022). The dataset contains about 70 variables describing the temporal (opening and closure date of publication) and geographical dimension (job location), as well as the economic activity of the company that posted the ad and the educational level required. Finally, the occupation is described by a list of related skills. Most variables refer to official

[5] EUROSTAT Data Browser. Online data code: ISOC_CI_AC_I. Last access 23/08/2023.

[6] https ://ec.europa.eu/social/main.jsp?catId=1223. Last access September 2023.

classifications (such as LAU and NUTS for geographical areas, ATECO2007 for economic activities), and skills are traced back to the European Multilingual Classification of Skills, Competences, Qualifications, and Occupations[7] (ESCO) taxonomy and the one used in the Occupational Information Network[8] system (O*NET).

For European countries, ESCO facilitates a more comprehensive and networked representation of the labour market, incorporating skills and knowledge dimensions in the definition of occupations (European Commission, 2019). Regarding occupations, the first four levels of the ESCO hierarchical classification of occupations (up to 4-digit ESCO level) coincide with those of the international standard classification of occupations ISCO-08 (ILO, 2012). Regarding skills, ESCO taxonomy distinguishes between two skill types: skills/competence concepts and knowledge concepts that are defined under the European Qualifications Framework: 'knowledge' means the outcome of the assimilation of information through learning (European Commission, 2019). On the other hand, 'skill' means the ability to apply knowledge and use know-how to complete tasks and solve problems (Council of the European Union, 2017). In ESCO taxonomy, action verbs are used when creating the terms for skills while learning outcomes (nouns) to specify what is known (knowledge). Moreover, each concept can be classified as 'occupation-specific', 'sector-specific', 'cross-sectoral' and 'transversal'[9]. This hierarchy, called skill reusability, refers to how specific a skill/knowledge is within occupations or economic sectors and is a relative concept because it depends on the relation between the skill/knowledge and a particular occupation.

Information in the Lightcast dataset is also organised on an internal taxonomy derived directly from OJAs (Magrini et al., 2023), and that, in the case of digital skills, enriches the ESCO taxonomy with skills tagged from Stackoverflow (in the following called Lightcast skills/knowledge) (O'Kane et al., 2020). The definition of digital and non-digital skills, which is relevant in our case, is not provided. Still, there is a generic reference to skills *from basic digital tools ('Excel') to advanced programming languages ('Python'), to skills and*

[7] This publication uses the ESCO classification of the European Commission.
[8] https://www.dol.gov/agencies/eta/onet. Last access September 2023.
[9] https://esco.ec.europa.eu/en/about-esco/escopedia/escopedia/skill-reusability-level. Last access September 2023.

abilities necessary to use these tools ('Computer literacy') and competencies that rely on digital tools to be carried out ('Data analysis') (Magrini et al., 2023).

Table 1 presents simple summaries of the Lightcast data in 2019 and 2021. In the following tables (unless otherwise specified), # skills refers to skill and knowledge concepts. In the two years, job ads have grown by 53.4%. Consequently, the number of skills requested increased considerably as well (64.4%), while the number of 2-digit and 4-digit ESCO occupations and unique skills remained almost unchanged. There are more 2-digit and 4-digit occupations in the ESCO taxonomy (44 and 426, respectively), but the differences are due mainly to the armed forces occupations. 1,245 unique skills occur in both years; 163 are in 2019, not 2021, and 169 are in 2021 and not in 2019.

The number of unique skills occurring in the dataset is much smaller than in the ESCO taxonomy (13,890), like in other countries (for Germany, O'Kane et al., 2020): some ESCO concepts are too broad to be specified in job postings, while others are too specific and are not used in job postings' language.

The average number of skills by job posting is slightly increasing (from 8.9 to 9.5), although the number of unique skills found is almost stable.

The proportion of Lightcast concepts is nearly 27% in both years.

Table 1: General information on Lightcast data. Years 2019 and 2021

			Avg.	ESCO occupation		#unique skills	
Year	#job ads	#skills	#skills	2-digit	4-digit	Total	Lightcast
2019	1,275,236	11,297,327	8.9	40	420	1408	397
2021	1,956,642	18,568,205	9.5	40	421	1414	379

Table 2 presents the number of job ads and requested skills for 2-digit ESCO occupations: *25 - Information and communication technology professionals* and *35 - ICT Technicians*, which are considered the ones with the highest concentration of ICT specialists.

The percentage increase in the number of job ads and the number of skills between 2019 and 2021 is lower than for the whole dataset (38.0% and 45.3% for ESCO25, 47.8% and 52.4% for ESCO 35). The job ads requesting ESCO 25 occupations are 7.4% of the total in 2019 and 6.6% in 2021 and contain many skills (20.9% of the total in 2019 and 18.5% in 2021). The average number of skills by job ad is therefore high (25.2 in 2019 and 26.5 in 2021).

The distribution of job ads and skills for ESCO 35 is much different: in both years, the number of job ads for technicians is about a fifth of those for professionals, while for the total of skills, the proportion is about 13%. Therefore, the average number of skills by job ad is smaller (16.2 and 16.7) but still much higher than the average calculated on the whole dataset. The higher number of skills for the professional segment is a characteristic finding of job advertisements, which also occurs in non-ICT occupations (Cedefop, 2023).

The distribution between the respective 3-digit ESCO occupations is strongly unbalanced in both years, with groups 251 and 351 prevailing on 252 and 352.

Table 2: Descriptive numbers of Lightcast data for 2-digit ESCO occupations 25 and 35. Years 2019 and 2021

2-digit	*25 ICT professionals*			*35 ICT technicians*		
3-digit	251	252	Total	351	352	Total
# job ads						
2019	81,921	11,973	93,894	17,946	561	18,507
2021	107,850	21,723	129,573	26,352	998	27,350
# skills						
2019	2,079,657	284,909	2,364,566	297,731	1,795	299,526
2021	2,912,534	524,271	3,436,805	452,677	3,861	456,538

In Table 3, which reports the names of the 2, 3 and 4-digit occupations, we identify 251 as *SW & applications developers & analysts* and 351 as *ICT operations and user support technicians*.

Table 3 also shows the differences between the internal 2-digit ESCO occupations 25 and 35 subgroups. In every subgroup, again, one 4-digit occupation prevails over the others with more than 50% job ads and total skills (e.g., *2512 SW developers* in group 251 and *3512 - ICT user support* in group 351). Regarding the average number of skills, as expected, all the occupations show values much greater than the general ones in Table 1, except the ones in group 352.

Table 3: Descriptive numbers of Lightcast data for 2-digit ESCO occupations 25 and 35 subgroups. Years 2019 and 2021

ESCO occupations	% job ads		% skills required		Avg # skills	
	2019	2021	2019	2021	2019	2021
2511 Systems analysts	31.2	34.6	23.8	26.6	19.3	20.8
2512 SW developers	54.8	51.0	60.4	56.8	27.9	30.1
2513 Web & multimedia developers	8.8	8.9	11.0	11.3	31.8	34.5
2514 Applications programmers	3.4	4.3	3.1	4.0	22.8	25.1
2519 N.e.c.	1.7	1.2	1.8	1.2	26.0	27.5
Total 251 SW & applications developers & analysts	100.0	100.0	100.0	100.0	25.4	27.0
2521 DB designers and administrators	7.2	3.8	8.4	4.3	27.6	27.3
2522 Systems administrators	62.9	58.3	59.0	53.5	22.3	22.2
2523 Computer network professionals	13.8	15.0	11.3	14.1	19.6	22.7
2529 N.e.c.	16.1	22.9	21.3	28.1	31.5	29.6
Total 252 DB and network professionals	100.0	100.0	100.0	100.0	23.8	24.1
3511 - ICT operations	20.3	20.4	16.5	15.4	13.5	13.0
3512 - ICT user support	60.4	60.9	55.4	55.0	15.2	15.5
3513 - Computer network and systems	6.0	5.5	7.1	7.3	19.6	22.8
3514 - Web technicians	13.2	13.2	21	22.3	26.4	29.0
Total 351 - ICT operations and user support technicians	100.0	100.0	100.0	100.0	16.6	17.2
3521 - Broadcasting and audiovisual	86.5	83.0	84.5	70.1	3.1	3.3
3522 Telecommunications engineering	13.5	17.0	15.5	29.9	3.7	6.8
Total 352 Telecommunications and broadcasting technicians	100.0	100.0	100.0	100.0	3.2	3.9

4. METHODOLOGY

The measure of skill change is the index proposed by Deming and Noray (2020). We applied it to understand if any changes in the required skills occurred between 2019 and 2021. By defining:

$$\Delta_{os} = \left(\frac{\#\ JobAds_{os}}{\#\ JobAds_{o}}\right)_{2021} - \left(\frac{\#\ JobAds_{os}}{\#\ JobAds_{o}}\right)_{2019} \quad (1)$$

the formula of the skill change index (SCI) for the single occupation *o* is:

$$SCI_o = \sum_{s=1}^{S} |\Delta_{os}| \quad (2)$$

where $\#\ JobAds_{os}$ is the number of job ads for occupation *o*, which require the skill *s,* and $\#\ JobAds_{o}$ is the number of job ads required by occupation *o.* In the present analysis, *o* refers to the most detailed level available, the 4-digit ESCO occupation, in groups 25 and 35.

This index measures the net skill change in each occupation. The higher the index value, the higher the skill change. The single addend assumes the value $\left(\frac{\#\ JobAds_{os}}{\#\ JobAds_{o}}\right)_{2019}$ when $\#\ JobAds_{o,2021} = 0$, and the value $\left(\frac{\#\ JobAds_{os}}{\#\ JobAds_{o}}\right)_{2021}$ when $\#\ JobAds_{o,2019} = 0$. As the index only shows the change in absolute value, we will decompose the total value to discover the contribution of the positive and negative addends in the summation.

Due to the peculiarities of the Italian labour market, which is characterised by a notable territorial specialisation, it may be helpful to report the index value by region to investigate any different geographical patterns in the change of skills. To this aim, we define:

$$\Delta_{rs} = \left(\frac{\#\ JobAds_{rs}}{\#\ JobAds_{r}}\right)_{2021} - \left(\frac{\#\ JobAds_{rs}}{\#\ JobAds_{r}}\right)_{2019} \quad (3)$$

and the formula of the skill change index for the region *r* is:

$$SCI_r = \sum_{s=1}^{S} |\Delta_{rs}| \quad (4)$$

where $\#\ JobAds_{rs}$ is the number of job ads in region *r* requiring the skill *s,* and $\#\ JobAds_{r}$ is the number of job ads in region *r*. Null denominators are treated as in formula (1). We use formula (4) to calculate two indexes separately for the 2-digit ESCO subgroups 25 and 35.

In both skill change indexes, there is also a composition effect. Specifically, the value of SCI_o is affected by the change in the distribution across regions of job ads requiring occupations o. Analogously, SCI_r is affected by the change in the distribution of job ads across occupations required by region r.

Finally, we try to express the heterogeneity of skill changes between occupations (still separately for the ESCO 25 and 35) and Italian regions through the variation ranges of the skill change $SCRO_s$ and $SCRR_s$ defined below:

$$SCRO_s = max(\Delta_{os}) - min(\Delta_{os}) \quad (5)$$

$$SCRR_s = max(\Delta_{rs}) - min(\Delta_{rs}) \quad (6)$$

5. FINDINGS AND DISCUSSION

In this application, unlike Kahlawi et al. (2022), we focus only on the unique skills requested for ESCO 25 and 35 occupations.

That said, the number of unique skills identified and for which the skill change index is computed are 869 and 389, respectively, for ESCO 25 and 35 (Table 4).

Table 4: ESCO and Lightcast unique skills/knowledge involved in the analysis. ESCO 25 and ESCO 35 occupations

Skill/ knowledge	ESCO 25		ESCO 35	
	#	%	#	%
ESCO	548	63.1	284	53.2
Lightcast	321	36.9	105	46.8
Total	869	100.0	389	100.0

The proportion of unique Lightcast skills/knowledge concepts is higher than in the whole dataset (see Table 1), especially in group ESCO 35. In addition, more than 80% of unique Lightcast concepts are requested by ESCO 25 occupations. Those concepts are not directly traced to the ESCO taxonomy; most refer to specific software and applications labelled with their names (e.g., ABAP, Ubuntu, etc.), and we consider them as knowledge. After that, the number of knowledge among the ESCO own skills is just over 50% (54% in ESCO occupation 25 and 51% in ESCO occupation 35), while among the Lightcast skills/knowledge concepts, excluding the missing occurrences, the percentage is

practically 100% in both groups. Therefore, we obtain Table 5, in which the percentage of knowledge type is remarkably high, especially in ESCO 25.

Table 5: Type of skills involved in the analysis. ESCO 25 and ESCO 35 occupations

Skill/	ESCO 25			ESCO 35		
Knowledge	Knowledge	Skill	Total	Knowledge	Skill	Total
Digital	337	71	408	155	52	207
Non-digital	269	180	449	87	88	175
Total*	606	251	857	242	140	382
Total %	70.7	29.3	100.0	63.4	36.6	100.0

* Skill type: 12 missing values for ESCO 25, 7 missing values for ESCO 35

In our opinion, the fact that knowledge is prevalent in the Lightcast taxonomy is not surprising since the job ads analysis algorithm tends to provide more specific information than the ESCO taxonomy, especially in the ICT domain. Here the purpose is to identify the tools and applications used in this area, which are numerous and constantly evolving. For example, the development and spread of cloud computing, data management services such as data protection and cyber security have enlarged the number of knowledge concepts in the digital area.

Figure 1 shows the skill changes SCI_o between 2019 and 2021 for the 4-digit ESCO occupations in groups 25 and 35. Note that each bar is decomposed to show the contribution of the positive and negative values to the skill change index. What is interesting is that, in many cases, the positives and negatives are almost balanced. The proportion of positive values remains between 42% and 65% for all occupations except two. Considering both the numerical contribution to the value of SCI_o and the proportion of positive values, there is no underlying trend towards growth (or decrease) in the demand for skills except for 3522 (*Telecommunications engineering*), which is, however, an occupation requested by a limited number of job ads and, for this reason, sensitive even to minor variations.

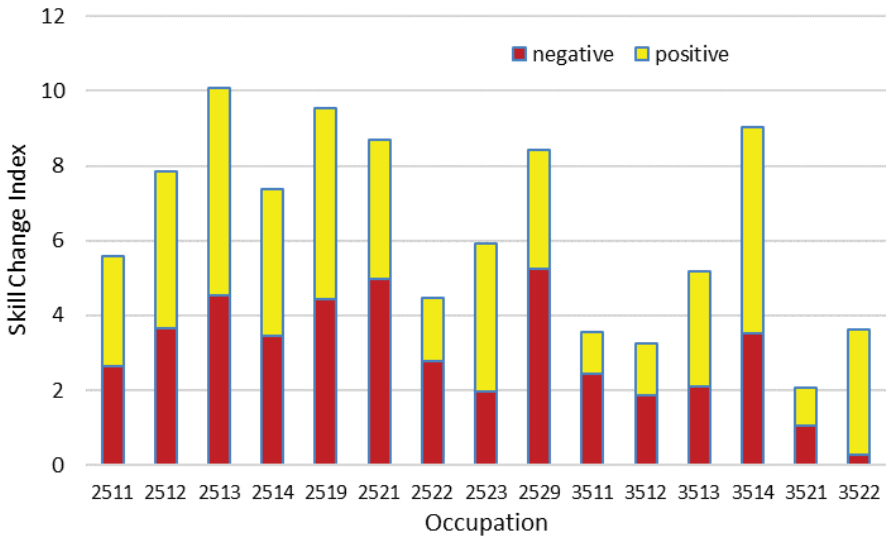


Figure 1: Contribution of negative and positive addends to SCI_o of ESCO 25 and ESCO 35 occupations. Years 2019-2021

We can observe that SCI_o values tend to be smaller for group 35 compared to 25, with one exception (occupation 3514). That likely occurs because of the different number of skills associated, as can be derived from Figure 2, which exhibits a remarkable positive association (correlation coefficient greater than 0.9) between SCI_o values and the average number of skills by job ad (see Table 3).

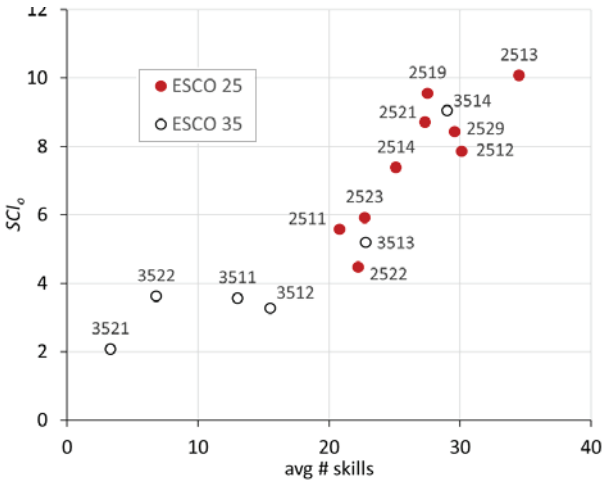


Figure 2: Relationship between SCI_o and the average number of skills by job ad for ESCO 25 and ESCO 35 occupations. Years 2019-2021

Figure 3 reports the values of SCI_r at the regional level still separately for occupations ESCO 25 and ESCO 35. Consistently with Figure 1, ESCO group 35 generally presents lower skill change in several Italian regions. Valle d'Aosta (VA) represents a peculiar case: it has the highest value in both occupations generated by only positive addends.

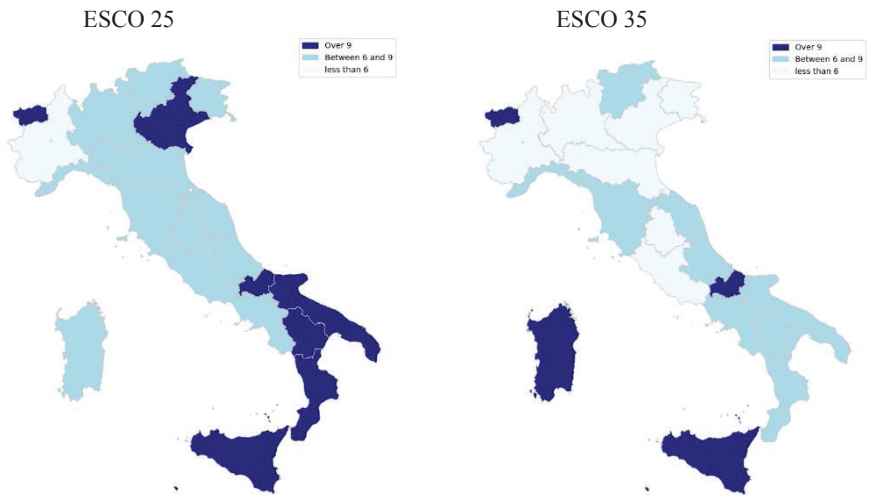


Figure 3: SCI_r for ESCO 25 and ESCO 35 occupations. Years 2019-2021

regions does not occur in the maps: for instance, for ESCO 25, Veneto, with a value greater than nine, belongs to the higher class, like five southern regions, while for ESCO 35, some of the central regions belong to the lower class and others to the intermediate class.

If we also consider the number of unique skills in each region (the ones that are in common in the two years), we find a negative correlation with the SCI_r values for both occupations: -0.64 for ESCO 25 and -0.87 for ESCO 35. This could be due to the different skill weights in the region in relative terms, as in formula (3).

For ESCO 25, there are six northern/central regions where the job ads contain more than 50% of the total unique skills[10]: Lombardy (LOM), Veneto (VEN), Lazio (LAZ), Emilia Romagna (E-R), Piemonte (PIE) and Tuscany (TOS). For ESCO 35, the six regions with the higher percentage are the same as before, but the maximum value (for Lombardy) is 26%. These six regions also have the highest total and ESCO 25 job ads (we refer to 2019 data, as in Kahlawi et al. 2022).

The following figures offer a more detailed view of the situation at the territorial level[11], with the decomposition of SCI_r to emphasise positive and negative changes. Even in this case, the negative and positive components of skill change persist in almost all regions. However, some exceptions exist: Valle d'Aosta (VA) shows all positive addends in both occupation groups, while in Veneto (VEN), negative values prevail in the ESCO 25 group.

Overall, there is a slight agreement between ESCO 25 and 35: the correlation coefficient between the two SCI_r series is 0.72.

[10] The percentage is calculated on the average of unique skills in 2019 and 2021 (see Table 1).

[11] The acronyms used to label the Italian regions are the following: PIE (Piemonte), VA (Valle d'Aosta), LOM (Lombardia), TAA (Trentino Alto Adige), FVG (Friuli Venezia Giulia), VEN (Veneto), LIG (Liguria), E-R (Emilia Romagna), TOS (Toscana), UMB (Umbria), MAR (Marche), LAZ (Lazio), ABR (Abruzzo), MOL (Molise), CAM (Campania), PUG (Puglia), BAS (Basilicata), CAL (Calabria), SIC (Sicilia), SAR (Sardegna).

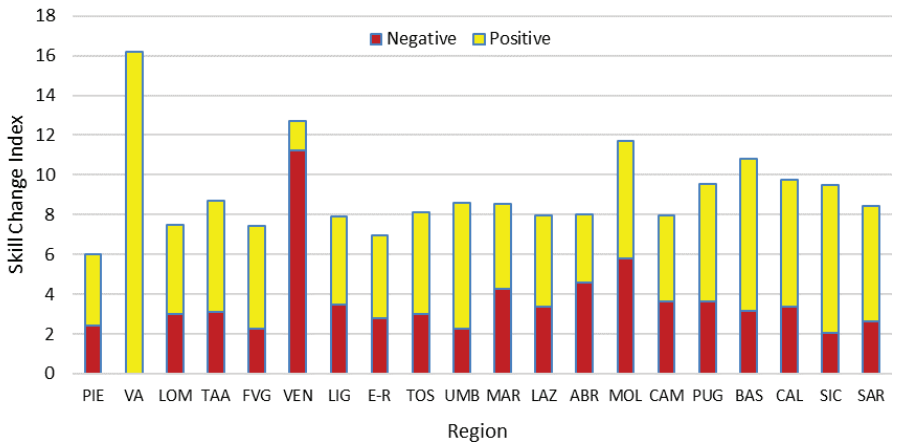


Figure 4: Contribution of negative and positive addends to SCI_r for ESCO 25 occupations. Years 2019-2021

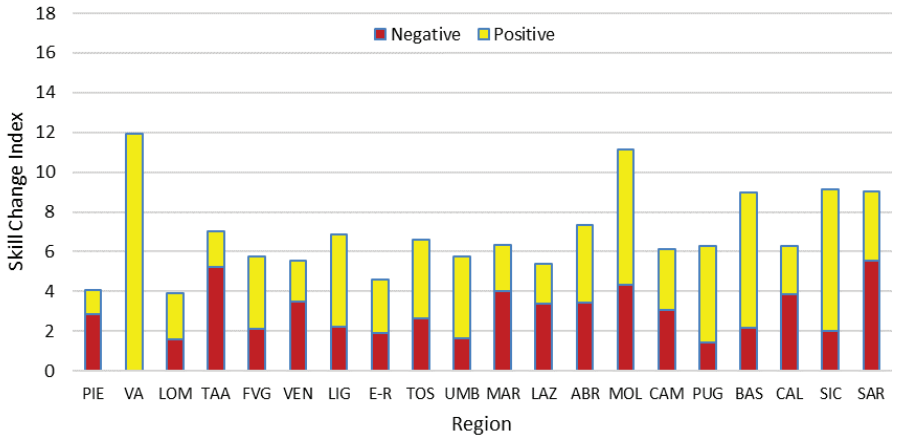


Figure 5: Contribution of negative and positive addends to SCI_r for ESCO 35 occupations. Years 2019-2021

Ultimately, we identify the skills with the highest skill change range across regions and occupations: Figure 6 represents each skill as a point in a Cartesian plane with coordinates $SCRR_s$ and $SCRO_s$ separately for ESCO groups 25 and 35. The lines of the 90th quantile of $SCRR_s$ and the 90th quantile of $SCRO_s$ divide the space into four quadrants. We focus on the skills in the first quadrant, where

$SCRR_s$ and $SCRO_s$ are above the 90th quantile: 49 for ESCO 25 and 25 for ESCO 35.

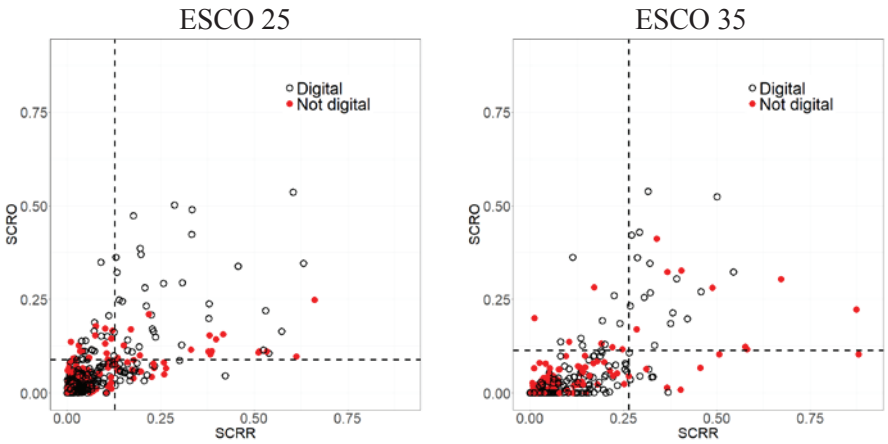


Figure 6: Skill change ranges for ESCO 25 and ESCO 35 occupations. Years 2019-2021

The skills in quadrant I are mostly digital (65% in ESCO 25 and 64% in ESCO 35), while knowledge concepts prevail in group 25 (55%) but not in group 35 (44%). The reusability level of each skill conditioned to the relative occupation is in Table 6: most of the skills are cross-sectoral or sector-specific, and only one is occupation-specific, confirming the relative proportions in groups 25 and 35.

Table 6: Reusability level of skills/knowledge concepts in quadrant I for ESCO 25 and ESCO 35 occupations

	ESCO 25		ESCO 35	
Reusability level	#	%	#	%
Tranversal	9	18.4	7	28.0
Cross-sectoral	15	30.6	8	32.0
Sector-specific	24	49.0	10	40.0
Occupation-specific	1	2.0	0	0.0
Total	49	100.0	25	100.0

To give an idea of the content of quadrant I, Tables 7 and 8 show the ten skills of quadrant I with the highest values for $SCRR_s$ and $SCRO_s$ separately for ESCO

25 and ESCO 35 occupations with the indication of the reusability level and type of the skill/knowledge.

Table 7: Skills/knowledge with the highest values for $SCRR_s$ and $SCRO_s$. ESCO 25 occupations

Across regions				
$SCRR_s$	Digital	Description	Reuse level	Type
0.662	NO	*project management*	sector-specific	knowledge
0.632	YES	*use software design patterns*	sector-specific	skill
0.613	NO	*adapt to change*	transversal	skill
0.605	YES	*analyse software specifications*	sector-specific	skill
0.574	YES	*administer ICT system*	sector-specific	skill
0.539	YES	*business ICT systems*	sector-specific	knowledge
0.534	NO	*work in teams*	transversal	skill
0.531	YES	*computer programming*	transversal	knowledge
0.526	YES	*use a computer*	cross-sectoral	skill
0.514	NO	*teamwork principles*	cross-sectoral	knowledge
Across 4-digit occupations				
$SCRO_s$	Digital	Description	Reuse level	Type
0.536	YES	*analyse software specifications*	sector-specific	skill
0.501	YES	*unified modeling language*	sector-specific	knowledge
0.489	YES	*SQL*	sector-specific	knowledge
0.473	YES	*JavaScript*	sector-specific	knowledge
0.423	YES	*Java*	sector-specific	knowledge
0.386	YES	*process data*	cross-sectoral	skill
0.370	YES	*use spreadsheets software*	transversal	skill
0.362	YES	*implement front-end website design*	sector-specific	skill
0.349	YES	*SQL Server*	sector-specific	knowledge
0.345	YES	*use software design patterns*	sector-specific	skill

Table 8: Skills/knowledge with the highest values for $SCRR_s$ and $SCRO_s$. ESCO 35 occupations

Across regions				
$SCRR_s$	Digital	Description	Reuse level	Type
0.879	NO	*assist customers*	cross-sectoral	skill
0.873	NO	*customer service*	cross-sectoral	knowledge
0.672	NO	*adapt to change*	transversal	skill

0.581	NO	*communication*	cross-sectoral	knowledge
0.576	NO	*work in teams*	transversal	skill
0.545	YES	*use MS office*	cross-sectoral	skill
0.507	NO	*team building*	occupation-	knowledge
0.500	YES	*use a computer*	cross-sectoral	skill
0.488	NO	*provide leadership*	cross-sectoral	skill
0.457	YES	*administer ICT system*	sector-specific	skill
		Across 4-digit occupations		
$SCRO_s$	Digital	Description	Reuse level	Type
0.538	YES	*CSS*	sector-specific	knowledge
0.524	YES	*use a computer*	cross-sectoral	skill
0.429	YES	*implement front-end website design*	sector-specific	skill
0.421	YES	*use markup languages*	sector-specific	skill
0.412	NO	*English*	transversal	knowledge
0.361	YES	*Java*	sector-specific	knowledge
0.346	YES	*business ICT systems*	sector-specific	knowledge
0.326	NO	*solve problems*	transversal	skill
0.323	NO	*create solutions to problems*	cross-sectoral	skill
0.322	YES	*use MS Office*	cross-sectoral	skill

6. CONCLUSION

The statistical analysis conducted in this work examines a measure of change in the demand for skills/knowledge for the ESCO 25 and 35 occupations. Although both refer to ICT, they are of different complexity: ESCO 25 (ICT professionals) is a group of occupations with a high density of skills/knowledge compared to ESCO 35 (ICT technicians) and includes a large part of the Lightcast knowledge concepts. A composition effect influences the skill change indexes SCI_o and SCI_r. SCI_o is affected by the shift in the distribution across regions of job ads demanding occupation o. Similarly, SCI_r is affected by the change in the distribution of job ads across the occupations within region r. In any case, the skill change is still evidence of a movement in the labour market at the level of the required skills.

The study at the 4-digit occupation level shows that the occupational profiles examined do not have a stable configuration regarding the changes in skill/knowledge demand, as the value of the skill change index is composed of an almost equivalent number of positive and negative variations, with a few exceptions.

A similar pattern occurs even at the regional level, with a heterogeneous picture of Italy. Still, no clear north-south divide emerges, and the results are difficult to interpret, given the numerous underlying factors that can influence the comparisons, as anticipated in Section 2. The study of this kind of data at the territorial level proves to be a complex task, as already pointed out in other empirical analyses (ILO, 2021; Kahlawi et al., 2022; Kahlawi et al., 2023, among others).

In any case, the analysis of ICT occupations seems to be a good training ground for assessing the information content of job ads. In particular, the skill perspective, derivable from OJAs, could be helpful to provide a more focused, comprehensive and detailed understanding of a context characterised by the rapid obsolescence of skills/knowledge and the emergence of new ones. In fact, OJAs specify occupations in terms of the required skills (verbs) and knowledge (nouns) according to the principles of ESCO taxonomy. However, the fact that most non-ESCO skills turn out to be knowledge concepts would suggest that the text of the job ads does not contain action verbs or, better, that the text mining algorithms simplify the statements, extracting only the keywords and decontextualising them. Still, the emphasis on knowledge contents is relevant in ICT occupations, where the tools to complete tasks continuously evolve. Furthermore, the reference to knowledge can be seen as a sound strategy to implement effective planning of education and training policies to limit the risk of skill mismatch.

Paper Funding: This work was supported by the Italian Ministry of University and Research (MUR), Department of Excellence project 2018-2022 - Department of Statistics, Computer Science, Applications - University of Florence.

REFERENCES

Aica, Anitec-Assinform, Assintel, Assinter Italia (2019). *Osservatorio delle competenze digitali*. https://www.assintel.it/assinteldownloads/osservatorio-delle-competenze-digitali-2019/. Last access: 19/01/2024.

Arregui Pabollet, E., Bacigalupo, M., Biagi, F., Cabrera Giraldez, M., Caena, F., Castano Munoz, J., Centeno Mediavilla, C., Edwards, J., Fernandez Macias, E., Gomez Gutierrez, E., Gomez Herrera, E., Inamorato Dos Santos, A., Kampylis, P., Klenert, D., López Cobo, M., Marschinski, R., Pesole, A., Punie, Y., Tolan, S., Torrejon Perez, S., Urzi Brancati, C. and Vuorikari, R. (2019). *The Changing Nature of Work and Skills in the Digital Age*. EUR 29823 EN. Publications Office of the European Union, Luxembourg.

Beręsewicz M. and Pater R. (2021). *Inferring Job Vacancies from Online Job Advertisements*. Publications Office of the European Union, Luxembourg.

Cammeraat, E. and Squicciarini, M. (2021). Burning glass technologies' data use in policy-relevant analysis: an occupation-level assessment. *OECD Science, Technology and Industry Working Papers*. 2021/05.

Carnevale, A.P., Jayasundera, T. and Repnikov, D. (2014). *Understanding Online Job Ads Data: a Technical Report*, Georgetown University, Technical Report.

Cedefop (2019). *Online job vacancies and skills analysis. A Cedefop Pan-European Approach*, The European Centre for the Development of Vocational Training, Thessaloniki.

Cedefop (2023). *Skills in Transition. The Way to 2035*. Publications Office of the European Union, Luxembourg.

Council of the European Union (2017). *Council Recommendation on the European Qualifications Framework for Lifelong Learning and Repealing the Recommendation of the European Parliament and of the Council of 23 April 2008 on the establishment of the European Qualifications Framework for Lifelong Learning*. 2017/C 189/03.

de Pedraza P., Visintin, S., Tijdens, K. and Kismihók, G. (2017). Survey vs scraped data: Comparing time series properties of web and survey vacancy data. *AIAS Working Paper*, 175, Universiteit van Amsterdam.

Deming, D.J. and Noray, K. (2020). Earnings dynamics, changing job skills, and STEM careers. *Quarterly Journal of Economics*. 135(4): 1965–2005.

European Commission (2019). *ESCO Handbook*, Directorate-General for Employment, Social Affairs and Inclusion.

Eurostat (2023a). ICT specialists in employment. *Statistics Explained.* https://ec.europa.eu/eurostat/statistics-explained/SEPDF/cache/47162.pdf. Last access: 07/12/2023.

Eurostat (2023b). ICT specialists – statistics on hard-to-fill vacancies in enterprises. *Statistics Explained*. https://ec.europa.eu/eurostat/statistics-explained/SEPDF/cache/40327.pdf. Last access: 07/12/2023.

Fabo, B. and Mýtna Kureková, L. (2022). Methodological issues related to the use of online labour market data. *ILO Working Papers* 68. International Labour Office, Geneva.

Giambona, F., Khalawi, A., Buzzigoli, L., Grassini, L. Martelli, C. (2021). Big data analysis of Italian online job vacancies data. In: B. Bertaccini, L. Fabbris, and A. Petrucci, Editors, *Statistics and Information Systems for Policy Evaluation*. FUP, Firenze: 117-120.

ILO (2001). *The Public Employment Service in a Changing Labour Market*, International Labour Organization, Geneva.

ILO (2012). *International Standard Classification of Occupations*. International Labour Office, Geneva.

ILO (2020). *The Feasibility of Using Big Data in Anticipating and Matching Skills Needs*. International Labour Organization, Geneva.

ILO (2021). *Public Employment Services Pressing Ahead with Digitalisation Should Be Aware of the Digital Divide*. ILO Policy Note.

Kahlawi A., Buzzigoli L., Grassini L., Martelli C. (2022). Skill similarities and dissimilarities in online job vacancy data across Italian regions. In: A. Balzanella, M. Bini, C. Cavicchia, and R. Verde, Editors, *51st Scientific Meeting of the Italian Statistical Society-Book of Short Papers*, Pearson Italia: 284-291.

Kahlawi, A., Buzzigoli, L., Giambona, F., Grassini, L., Martelli, C. (2023). Online job ads in Italy: A regional analysis of ICT professionals, *Statistical Methods & Applications*, DOI: 10.1007/s10260-023-00735-9.

Lightcast (2022). *European Data Methodology*. Lightcast.

López Cobo, M., Rohman, I.K., De Prato, G., Cardona, M., Righi, R., Samoili, S., Vázquez-Prada Baillet, M. (2020). *ICT Specialists in Employment. Methodological Note*, Seville: European Commission, JRC119846.

Lovaglio, P.G., Cesarini, M., Mercorio, F., Mezzanzanica, M. (2018). Skills in demand for ICT and statistical occupations: Evidence from web-based job vacancies, *Statistical Analysis and Data Mining*. 11:78–91.

Lovaglio P.G., Mezzanzanica, M., Colombo, E. (2020). Comparing time series characteristics of official and web job vacancy data. *Quality & Quantity*. 54: 85–98.

Lucarelli, A., Righi, A. (2023). Enriching job vacancy official information with online job advertisements: chances and limits. In A. Bucci, A. Cartone, A.Evangelista, and A. Marletta, Editors, *IES 2023 - Statistical Methods for Evaluation and Quality: Techniques, Technologies and Trend*s (T^3), Ed. Il Viandante, Pescara: 119-125.

Magrini, E., Pelucchi, M. and Brown, D. (2023). How is the digital transformation changing demand for skills in apprenticeships-typical occupations? *Cedefop / OECD Symposium on apprenticeships and the digital transition, 15 - 16 June 2023.*

Mezzanzanica, M., Mercorio, F. (2019). *Big Data for Labour Market Intelligence: An Introductory Guide.* European Training Foundation, Turin.

Napierala, J., Kvetan, V. and Branka, J. (2022). Assessing the representativeness of online job advertisements. *Cedefop Working Paper*, No 17. Publications Office of the European Union, Luxembourg.

O'Kane, L., Narasimhan, R., Nania, J., Burning, Taska, B. (2020). *Digitalisation in the German Labor Market*, Bertelsmann Stiftung, Gütersloh.

OECD (2020). *What Skills Do Employers Seek in Graduates? Using Online Job Posting Data to Support Policy and Practice in Higher Education*. OECD Education Working Papers, No. 231.

Schmidt, J., Pilgrim, G. and Mourougane, A. (2023). What is the role of data in jobs in the United Kingdom, Canada, and the United States? A natural language processing approach. *SDD Working Papers.* No. 119, OECD.

Unioncamere and ANPAL (2022). *Le competenze digitali. Analisi della domanda di competenze digitali nelle imprese. Indagine 2022*. https://excelsior.unioncamere.net/sites/default/files/pubblicazioni/2022/B5-Competenze_Digitali_2022_DEF.pdf. Last access: 18/01/2024.

Vannini, I., Rotolone, D. and Di Stefano, C. (2019). Online job vacancies in the Italian labour market. *Statistica Applicata - Italian Journal of Applied Statistics*. 31(1): 97–118.

doi.org/10.26398/IJAS.279

TRENDS IN THE LABOUR MARKET: ISSUES STILL OPEN FOR THE ANALYSIS

Caterina Marini[1], **Vittorio Nicolardi**[2]
Department of Economics and Finance, University of Bari Aldo Moro, Bari

Abstract *The availability of dynamic databases to analyse the Italian labour market is still unsatisfactory because of many constraints that impede research on causal correlations between employment trends and economic-juridical background. The Italian labour market normative scheme was totally renovated at the beginning of the 21th century but the reform effects on employment and production are still partially known. In this paper, the economic statistical effects of the promulgated labour market reforms on the workforce dynamics are investigated through a graphical-matrix approach that uses ISTAT official data and administrative microdata from Veneto Labour Office. It is proved the undisclosed data from administrative databases alongside the official statistical information can allow to depict a picture of a complex phenomenon, as the Italian labour market is.*

Keywords: *Labour market, Administrative data, Official statistical data, Transition matrix.*

Acknowledgements *The authors are grateful to Veneto Lavoro for providing the database Mercurio for our Ven.DB.*

[1]caterina.marini@uniba.it
[2]vittorio.nicolardi@uniba.it

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.
Competing Interests: The Author(s) declare(s) no conflict of interest.

1. INTRODUCTION

The third millennium, the same that was welcomed with fabulous celebrations all over the world, began unexpectedly with important disruptions to the global equilibrium in all aspects of the human life, and the world economy is still struggling to resist default and social fail, or recover from hard economic and social challenges. In fact, the innumerable economic structural breaks that occurred since the beginning of the 21th century have marked a turning point in the development of most economies around the world, both advanced and emerging/developing: in order, the US early recession and the Twin Towers attack in 2001 provoked global uncertainty; the great and long US recession in 2007 caused a global financial crisis because of the subprime mortgage crisis; the EU economic and financial crisis in 2008/2009 as a consequence of the American downturn; the EU sovereign debt crisis that started in 2009 and marked deeply the economy of some peripheric EU countries causing the Greek default and severe control policies of debt in all countries; the pandemic of SARS-Cov2 in 2020 tested again the global economy stability and, finally, the Russia-Ukraine conflict in 2022 provoked a deep increase of inflation all over the world.

Among all the serious consequences that the above picture yielded, the labour market of most countries suffered the economic depression and Italy was not beneath the others. In particular, for the period 2008-2013 the statistical data on the Italian labour market sketch an alarming depiction of the situation: the sharp drop in jobs; the increase of unemployment up to around 1.5million workforce, almost half in 2012; a rise of youth unemployment (+14.5%) that contributed 80% to the total unemployment rate over the period (+12.1%).

Therefore, in Italy the long recession of the economy and the complex situation of the labour market forced the Italian governments to address the problem to strengthen the manufacturing sector on one hand, and restrain the haemorrhage of jobs on the other. It is starting from 1997 that the Italian government gradually reformed the antiquated regulatory framework of the Italian labour market, still inappropriate to face the new economic challenges at that time. The cornerstone of the problem was focused on the labour market and its functioning: job creation, living standards and social cohesion depend on its ability to well operate. Therefore, the labour reforms that were implemented since the beginning of the new millennium turned the Italian labour regulation upside-down and the flexibility was the real challenge. It is important to consider that the economic context when all the mutations happened is characterised by continuous changes into the global productive process, more vulnerable, competitive and variable than in past, and

the Italian enterprises and companies needed to face the new globalised framework to save or gain important market shares. And in this context, a new regulation of the Italian labour market was extremely important to guarantee higher job flexibility on one hand, and a well-defined juridical protection of workers' rights on the other. Furthermore, it is important to bear in mind that, towards the end of the 20th century, the job flexibility was not only a milestone for Italy but also an important issue of the European agenda because rigidities of the labour markets were an obstacle to the development of many European countries and debates over probable solutions were in progress since the 1980s. However, it is not intention of this work to report and analyse the juridical intentions of the Italian legislator in his decisions of jurisprudence to revise the system of rules for the labour market. The framework of rules and norms that were approved in 18 years to update the Italian labour market is undoubtedly central to explain trends and magnitudes, but the economic statistical effects on the labour forces of the governmental interventions are under analysis, instead. In literature, the Italian labour market is extensively analysed not only to depict the status of the Italian workforce but also to evaluate the effects that the juridical and financial interventions have on its dynamics.

Some authors examine the effects that flexibility produces on the employment rate and firm productivity. Battisti and Vallanti (2013), in their analysis on the consequences that the decentralised wage schemes and the temporary job contracts have on worker/firm performance, demonstrated that the large flexibility in the working environment, namely the presence of many temporary contracts, implies a reduction in workers' motivation and effort, and lower probabilities of dismissal for workers with permanent contracts. Boeri and Garibaldi (2007) similarly argued that productivity experiences lower growth rates when firms hire temporary workers. Many researchers analyse the consequences of flexibility from the social point of view and its impact on the household economy. Tomelleri (2021) analysed the relationship between the temporary employment and the individual earnings to estimate the impact of the labour market reforms on income inequalities. And, in regard to this aspect, Hoffmann and Malacrino (2019) proved that the employment period length is the most important element of the increase in the Italian earnings. Many other scientists focused more on the impact that the Italian labour market reforms had on the Italian employment trends. And in particular, the last reform still effective and approved in Italy in 2014, implemented in 2015, named JOBS Act, is under deep investigation still now. Cirillo et al. (2017), for instance, tried to describe the first effects on the employment of the JOBS

Act and the extraordinary economic and financial measures that were planned to stimulate employment and, although the availability of data was limited and information aggregated, they concluded with first evidence of their negative effects on employment. Sestito and Viviano (2018) pointed out other important aspects of the JOBS Act and proved that the new firing rules as regulated in its normative framework stimulated the open-ended employment more than fixed-term. However, the complexity of the analysis of the Italian labour market, a key issue indeed, arises from the long reform phase that embraced 18 years since 1997 and, as we show next, experienced the overlapping of rules and economic-financial interventions that complicates the comprehension of trends and outcomes from the demand side. Furthermore, the limited availability of open data, both official and administrative, is a further complexity. Therefore, all researches conducted on the labour forces' dynamics and employment/unemployment in Italy are complementary analyses with all the others. In this paper, we analyse the evolution of employment in Italy from the juridical-economic perspective. In particular, through the utilisation of official statistical data and administrative microdata, we focus our attention on one of the Italian regions that is very active in terms of GDP, i.e. Veneto, and prove that the combination of several and variegated data sources as compiled by different subjects can allow to depict the labour market as a whole and solve some issue still unsolved, both statistical methodological and legal/pragmatical. As we already pointed out in a previous work (Marini and Nicolardi, 2021a), the administrative data are a precious source of information, but they are not exempt from many issues related to their nature: nonstatistical but administrative purposes; material and human errors; very large dimension, huge occasionally; variety of data primary sources; juridical because of the individual privacy. In particular, the availability of dynamic databases referring to the labour market is not satisfactory for the most various analytical purposes. In fact, in most cases, the problem is caused by the general data protection regulation that impedes the use of many precious data and, therefore, the analysis of the real effects of the labour market reforms. In theory, hence, there is a vast availability of data in many data formats, but they are not usable and mergeable with other data both official and administrative of various types. At the same time, the values and the different data sources are complete autonomously to yield a structural and complete analysis of the labour market because they are the outcome of administrative or institutional activities or surveys that are part of the work of the corresponding institution or administration or company. Furthermore, data that are more complete in terms of analytical potentialities are exclusively on the labour

demand side or labour supply side and, therefore, not appropriate for a dynamic analysis of the labour market that would need an alignment of both information. Therefore, the enormous potentialities of the administrative data collide with their ambiguities when the latter are not solved and statistically managed to harmonise and align the administrative data with official statistical data as yielded by the national institutes of statistics. The approach we use in this work is a graphical-matrix method to investigate the labour mobility: transition matrices jointly used with a graphical representation of the labour market flows. The use of the transition matrix is scientifically consolidated but not prevalent in the analytical context (Albisinni and Discenza, 2004; Bernardi and Zaccarin, 1984, 1991; Blumen et al., 1995; Ward-Warmedinger and Macchiarelli, 2013) and allows to grasp all changes in the job positions and the probability that each employee move from his working status to another. In the framework of the radical reforms that involved the Italian labour market, the analysis of the employment flows through the graphical-matrix method is a positive experiment to describe trends and the interconnected effects of a complicated scheme of rules and financial interventions to recognise all the opportunities and potentialities. In Marini and Nicolardi (2018) a first evaluation of the transition probabilities in the transition matrices was yielded as a proxy based on the official data as provided by the Italian National Institute of Statistics (ISTAT, hereafter) to estimate the change of job status from temporary to permanent job positions as a consequence of coming into effect of the last reform in 2015. This work is the progression of that first a pproach: information is integrated with administrative data, though referred to a single Italian region, and the stage of the analysis is sufficient to suppose the effectiveness of integrated data in examining complex social-economic phenomena.

The paper is organised as follows: a brief description of the process that reformed the Italian labour market is reported in Section 2; Section 3 describes the dataset contents and the outcome of the analysis; some concluding remarks are reported in Section 4.

2. THE LABOUR MARKET REFORM SCHEME

The labour market policies that were pursued in Italy between 1997 and 2014 had the great task to renovate and liberalise the employment policy, still obsolete and constrained by dated regulations. The revision process of the labour market policies was addressed to the active population in all ages and all status, both employed and unemployed/first-job seekers, both young and middle age. The process started in 1997 but the bulk of major reforms was introduced at the beginning of

the 21st century. In 1997 the Treu Reform (law n. 196/1997) took the first steps towards the job flexibility reducing the constraints that prevented firms from using fixed-term labour contracts and part-time labour contracts. In 2001, in compliance with the European directive n.1999/70/CE, the Italian government introduced the fixed-term causal labour contract that reduces slightly more the constraints on its use adducing adequate reasons related to selected manufacturing sectors, job titles and well-defined contingencies (legislative decree n.368/2001). At the beginning of the 21st century, in the light of a weak socio-economic context and a worrying employment stagnation, the semi-liberalisation of the fixed-term contract was a positive equilibrium between the entrepreneurial needs of job flexibility and the workers' necessity of socio-economic stability. In 2003, the Biagi Reform (legislative decree n.276/2003) continued along the path towards flexibility, reformed some pre-existing short-term contracts and introduced a wide variety of other short-term para-subordinated and non-standard employment contracts in the Italian labour market: on-call contract (lavoro intermittente), job sharing contract (lavoro ripartito), occasional employment (lavoro occasionale), project-job contract (lavoro a progetto), coordinated and continuous collaborations (co.co.co.), outsourcing or staff-leasing contract (somministrazione) and entry position or training contract (contratto di inserimento).

A first interesting summary of the effects that were yielded by the introduc-tion of the short-term and non-standard employment contracts in the Italian labour market still unprepared for the great change can be find in Tealdi (2011): to prove the discrimination in terms of the rights experienced by the short-term workers, the author analysed the characteristics of all contracts (both permanent and fixed-term and temporary) that were considered in the Treu Reform (1997) and Biagi Reform (2003), how they changed between the two reforms and how their use and worker's benefits were affected by the two reforms over time. Others (Daruich et al., 2023) focused their analysis on the legislative decree n.368/2001 that, as explained before, introduced new rules for the fixed-term causal labour contract: to evaluate the impact that the new policy of the labour market had on the Italian employment as a whole, on jobs, firms, and workers and across different sectoral collective bargaining agreements, the authors used longitudinal data combining matched employer-employee data with firms' financial records and proved that firms experienced great advantages from the policy, while young workers were more penalised.

In 2012, the alarming employment stagnation and the high uncertainty about the economic-financial situation forced the Italian government to reform again

rules and norms of the Italian labour market. The Fornero Reform (law n.92/2012), the big and severe reform of the Italian labour market, promoted the open-ended contract, enhanced the apprenticeship contract for young, widened the range of applicability of the fixed-term contracts although restricted their use (only one renewal) and duration (max 36 months) and, more important and controversial, reduced the effectiveness of the worker protection in case of illicit layoff (Articolo 18). In 2014, the Poletti Decree (legislative decree n.34/2014) and the Jumpstart Our Business Startups Act, known as JOBS Act (law n.183/2014) were passed. The Poletti Decree enhanced the fixed-term contracts considerably reducing at minimum the constraints on their use and introduced the sole obligation that the ratio of fixed-term contracts to open-ended contracts would not exceed 20%. Instead, the JOBS Act (JA, hereafter), the last and still effective reform, the controversial and debated law, introduced in the jurisprudential scheme of the Italian labour market considerable innovations: a new revision to the fixed-term contract (duration up to 36 months plus one renewal of 12 months) and the removal of all causality constraints; the abrogation of Articolo 18; the contested revision of the open-ended contracts that reduced the protection against just cause dismissals; the introduction and promotion of the new open-ended contract known as "increasing protection" contract (contratto a tutele crescenti) that provides only an economic compensation in case of illicit layoff; the promotion of the apprenticeship contract; the revision of the outsourcing contract and the removal of the project-job contract.

The three major labour market reforms described above (i.e. Biagi Reform, Fornero Reform and JA) included clauses of some monetary incentives taking the form of some discount in firms' social contribution burden per employee to enhance employment of vulnerable segments of workforce, namely women and young, but JA was the reform among them that could also rely on two extraordinary governmental financial plans addressed to all labour forces. In fact, just before the promulgation of JA (March 2015) and the introduction of the "increasing protection" contract in the new regulation of the Italian labour market, important and significant financial measures, as provided in the 2015 Stability Law (law 190/2014, implemented in January 2015), planned monetary extraordinary incentives as a 100% discount in social contribution burden per employee for a 3-year period for those firms that in 2015 hired people with permanent job contracts or transformed temporary job contracts/fixed-term contracts in permanent job contracts for people already employed. The financial measure was replicated in the 2016 Stability Law (law 208/2015, implemented in January 2016) although the

monetary incentives amounted to a 60% discount in social contributions for a 2-year period. Though the two extraordinary governmental financial plans covered two different periods and the magnitude of discounts in firms' social contribution burden per employee is different, all monetary incentives ended in December 2017.

3. THE DATA

In this work, the datasets used to defend our assumption about the potentiality of an integrated information from the demand and supply side are fundamentally two.

The first is the data warehouse of ISTAT, named I.Stat, that is easily available online. The time series of the section Labour Offer contains data from the Labour Force Survey referred to annual, quarterly and monthly information on labour forces for the whole country and the levels NUTS 1, 2 and 3. We name this database LFS.DB. At the time of this work, the up-to-date information was referred to the 2nd Quarter 2022 and we decided to end the period under analysis in 2020, without affecting our outcomes. That was a pragmatic analytical decision to avoid the structural break that occurred in all socio-economic time series because of the SARS-Cov2 pandemic.

The second is the database of the Italian Ministry of Labour and Social Policies for the Region Veneto. The Veneto database (Ven.DB, hereafter) is the only administrative database easily available under request that provides a detailed and complete information of the individual working status over time. Ven.DB contains flow data that derive from the mandatory communications that, based on the regulation of the Italian labour market, employers have to compile when a new employee is hired or changes occurred on the existing contracts. The mandatory communications contain many information on the employer, employee, typology of contract, duration of the contract, and any change about the contract such as transformations, extensions and termination (Marini and Nicolardi, 2017). The great job that Veneto Labour Office made to provide undisclosed microdata referred to all regional individual jobs and contracts is the main reason why its database is available for deep analyses of the regional labour market. As many administrative databases, Ven.DB is not exempt from the problems that characterise the non-official statistical information although the Veneto Labour Office tried to solve many issues. Therefore, the procedure to deal with this type of data, as shown in Marini and Nicolardi (2021a), has been applied to homogenise information and use the same in the best way. Furthermore, it is important to consider

that a further issue is the dimension of Ven.DB that is quite large to be considered in the Big Data scenario (Marini and Nicolardi, 2021b). Finally, as well known, Veneto is one of Italian regions where the performance in terms of employment is great and significant. Therefore, in our experiment, we are considering Veneto as a territorial proxy to evaluate the informative contents of the administrative data for the whole country. At the time of this work, the up-to-date information was refereed to 2017 and, however, this time is adequate to analyse the first effects of all labour market reforms and monetary incentives as described in Section 2. Table 1 shows the numerical structural characteristics of the key fields in Ven.DB at the time of our analysis.

Table 1: Veneto Lavoro database contents.

Job positions	17,604,175
Employees	3,692,529
Employers	892,084
Working locations	6,954
NACE codes (ATECO 2007)	919

Therefore, the starting year of the period under investigation is the same for both databases, i.e. 2011, but the ending period is different, i.e. 2017 in Ven.DB and 2020 in LFS.DB. Some clarification is necessary to explain the reasons behind the differentiated period and the starting date, as described above. First, 2011 is statistically the right year to start the evaluation of the effects of the two most debated reforms (as described in Section 2) that deeply changed the Italian labour market. Second, the LFS.DB data consent to grasp the effects of both the monetary incentives and the new rules (as defined in JA) on employment owing to the length of its time-series, while Ven.DB data are important to grasp the transitions between different working positions mainly owing to the monetary incentives because data are referred to a time interval, though shorter then LFS.DB, that covers the period involved. It is important, however, to underline that the conclusion on the assumption we are defending in this paper is not affected by the two different lengths of the time intervals.

In the analysis, we decided to focus our attention on the two contracts of employment that were the main targets of the last two labour market reforms, namely Fornero Reform and JA, trying to prove the effectiveness of the new juridical assets over time. Therefore, the open-ended contract and the fixed-term contract are

under investigation through LFS.DB and Ven.DB.

The first information that is important to highlight is referred to the magnitude of the two contracts as occurred in Italy. Figure 1 shows the distribution of employees with open-ended contracts (light grey) and fixed-term contracts (dark grey) over the period 2011-2020, ISTAT data. As expected, the open-ended contract is the primary status of employment in the Italian labour market all over the period: 14.8 million of employees (85.3%) versus 2.5 million of fixed-term contract employees (14.7%), on average over the whole period.

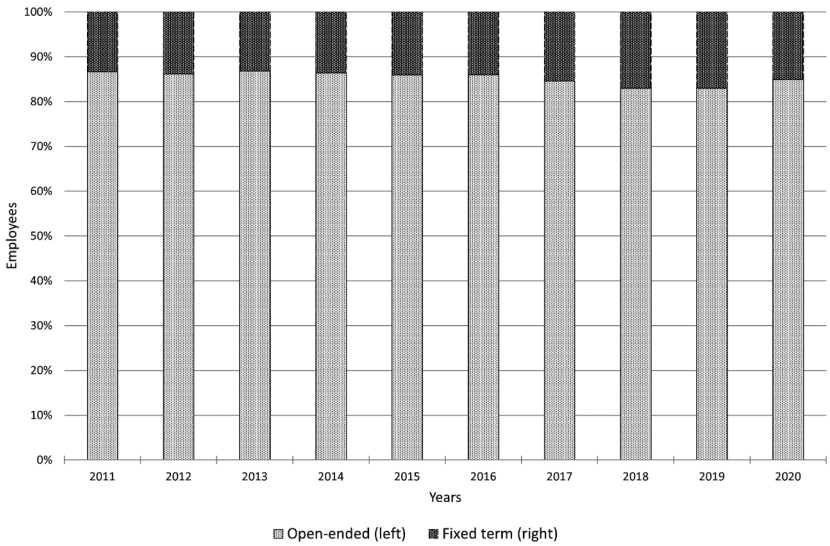


Figure 1: Distribution of employees with fixed-term contracts and open-ended contracts. Years 2011-2020.

Figure 2 shows more clearly in a double axis chart the ISTAT employment trends in the two contracts: open-ended contracts (continuous line) and fixed-term contracts (dash line). Trends of both contracts are quite similar although their paths appear to be differently affected by the political and juridical interventions starting from 2015, when the JA and the monetary incentives were planned to support the Italian labour market.

The analysis of the annual rates of change in the use of the two contracts under investigation, over the whole period, as shown in Figure 3, allows to suppose that the juridical and economic interventions effectively succeeded in the intention of the legislator to stimulate the permanent employment more that the temporary

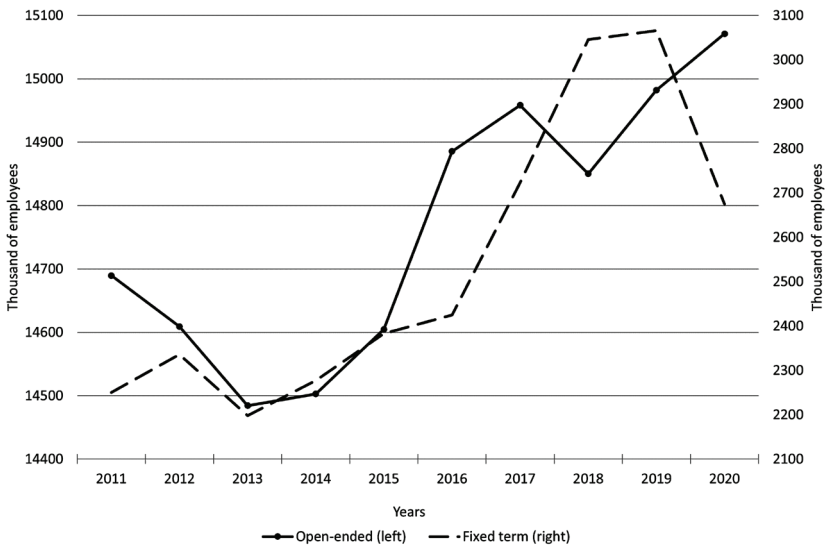


Figure 2: Italian employment trends in open-ended and fixed-term contracts. Thousand of employees. Years 2011-2020.

when they were simultaneously effective. At the same time, the end of the monetary incentives (year 2017) to permanently hire people or transform their contracts from fixed-term to open-ended slowed down the permanent employment (Fig. 3) until 2018, when the same decreased, while the fixed-term employees started increasing sharply in 2017 until 2019, right before the SARS-Cov2 period[3]. This last increase in the fixed-term contracts can be the outcome of the liberalisation of the fixed-term contract use, which the JA regulates, due to the removal of all causality constraints previously effective. It is important to note that the Fornero Reform had little effects on the employment for both types of contracts mainly because of the negative economic conjuncture caused by the EU sovereign debt crisis.

The implementation of a linear regression model in time series that involves economic variables is worthy to sustain hour hypothesis. Let PE_t denotes the amount of permanent employees by quarter t, SPI_t denotes the seasonal adjusted industrial production index at quarter t as a proxy of the Italian economic perfor-

[3]The year 2020 represents a structural break in the world economy because of the effects of the lockdown caused by the SARS-Cov2 pandemic. This is the reason why we preferred to not compare this value with the rest of the time series.

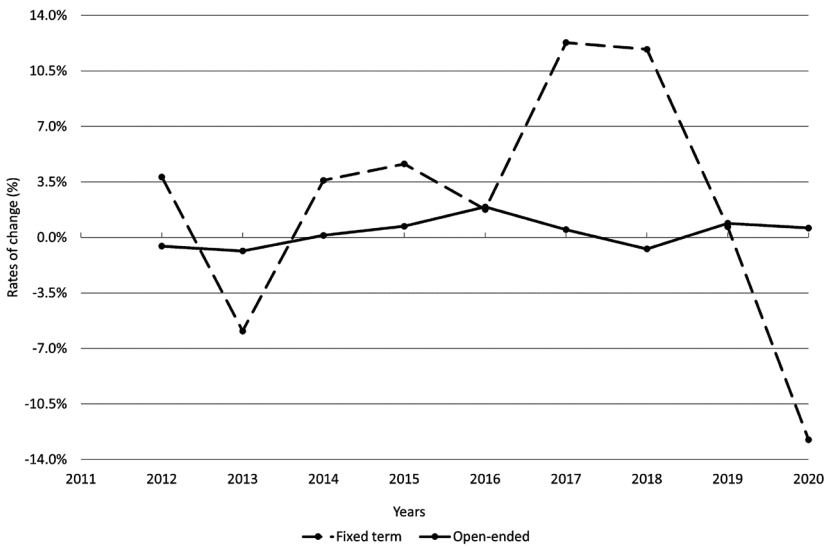


Figure 3: Italian employment trends in open-ended and fixed-term contracts. Annual rates of change. Years 2011-2020.

mance, $dEMI_t$ is a dummy variable at quarter t equal to 1 when the extraordinary monetary incentives are applied and t denotes the time trend variable. Then, the model is defined as follows:

$$PE_t = \alpha + \beta_1 \, PE_{t-1} + \beta_2 \, SPI_{t-1} + \beta_3 \, dEMI_t + \beta_4 \, t + \beta_5 \, t^2 + u_t \qquad (1)$$

Table 2 summarises the OLS estimates of the model, where standard errors are in parentheses and stars denote $p-values$. As expected, based on the assumptions of our hypotheses, it is confirmed that the significance of the implementation of the extraordinary monetary incentives produced, over the period considered, a positive effect on the amount of permanent employees, higher than the effect deriving from the economic performance. The specification analysis confirms the conclusion of our experiment in terms of econometric aspects.

When the focus of the analysis is circumscribed to Veneto, because of the regional data availability from the administrative side, based on ISTAT data it apparently appears that the regulatory framework experienced a major role in the employment performance than the financial-economic incentives. In particular, the regulation that the JA introduced, referred to the fixed-term contracts, led to

Table 2: OLS linear model regression results for the extraordinary monetary incentives of JA

α	7744.2***	(1610.1)
β_1	0.441577***	(0.113)
β_2	4.69393**	(2.023)
β_3	134.510***	(32.801)
β_4	-14.6054***	(5.135)
β_5	0.469966***	(0.127)

*** p-value < 0.01
** p-value < 0.05

a significant increase in the number of temporary employees, (55 thousand, i.e. 27.4%, ISTAT data) and only a modest growth of the number of permanent workers (38 thousand, i.e. +2.8%, ISTAT data) over the period 2015-2017. Figure 4 is a double chart that depicts the Veneto employment trends: above, ISTAT employment trends in open-ended contracts (continuous line) and fixed-term contracts (dash line); below, Ven.DB employment contracts trends that report the open-ended contracts (continuous line), both new and transformed, and the new fixed-term contracts (dash line). The two graphs apparently show opposite trends in the two contracts but the mistake is driven by the two different typologies of data that LFS.DB and Ven.DB contain: the first are always stock measurements and provide the number of employees; the second are flow measures and provide the amounts of new contracts that are signed over the period.

Therefore, the ISTAT graph (Figure 4a) shows trends roughly aligned with the corresponding national trends, although more accentuated because of the characteristics of the Veneto economic system, always more dynamic than the national over the period under investigation, as often highlighted in the series Economie Regionali (that is Regional Economies) of the Bank of Italy and the series Rapporto Statistico (that is Statistical Report) of the Veneto Region[4].

Ven.DB (Figure 4b) highlights, instead, a significant increase in the new open-ended contracts in 2015 and a steady increase in the new fixed-term con-

[4] An overview of the economic sectors based on data for Veneto from LFS.DB shows that, over the whole period, most of the regional workers are in the service activities (namely, all NACE Sections between J and U) that employed, on average, 42% of people, to whom it is necessary to add the employees in the turistic and trade activities (NACE Sectors between G and I) that are around 20%, on average; the manifacturing, mining and quarrying activities (NACE Sectors between B and E) employed 28%, on average, of workers in Veneto.

tracts over the period 2014-2016.

The outcome from the Veneto Labour Office data confirms that the monetary incentives, as in the 2015 Stability Law, had a surprising effect on the open-ended contracts (+77.7%) on the demand side, i.e. entrepreneurs, as the legislator prefigured in his economic intervention, but it is important to suppose that many can be transformations from temporary to permanent positions. Although the Ven.DB fixed-term contract trend can prove our hypothesis (Figure 4b), we need the transition flows between the two job positions to confirm that. Ven.DB data are helpful for this proof and they are used to compute the transition values we need. Figure 5 shows a bar chart where the transition flows from temporary contracts towards permanent contracts are depicted, including also other popular short-term contracts as regulated by Fornero Reform and JA to stimulate the youth employment, namely outsourcing, training and apprenticeship contracts.

Figure 5 shows the confirmation of our assumption and, in particular, the proof of the extraordinary increase of open-ended contracts in 2015 that was strengthened by the exceptional monetary incentive that definitely supported the JA. Table 3 shows the transition flows from the fixed-term contract to the open-ended contract, by age-class for the whole period. The transition flows indicate the number of persons changing their labour status between two time periods, that is changes between temporary and permanent job positions in our analysis. As expected, the magnitude of transitions is significant in 2015 (56,747) and represents the 34.3% of the total open-ended contracts and the 12.4% of the total fixed-term contracts signed in the same year. Another valued outcome to highlight is referred to the distribution by age of the transition flows. In particular, most of the transformations from fixed-term positions to permanent occurred among employees at age 26 or more, over the whole period.

Table 4 shows the transition flows (in percentage) from the fixed-term contract to the open-ended contract in relation to total amount of fixed-term contracts, by age-class for the whole period. In particular, as shown in Table 4, in 2015 and 2016 the age-class 26-30 experienced more advantages (15.2% and 10.6%,respectively) from the monetary incentives to transform the employment contract from fixed-term to permanent than the other age classes.

Another valuable outcome that we wish to highlight is related to the apprenticeship contract. The apprenticeship is considered the most important contract to facilitate the youth employment in both reforms, and the legislator provided for monetary incentives to also transform the apprenticeship contract into open-ended contract. As shown in Figure 5, between 2013-2017 transformations of

Table 3: Veneto transition flows from fixed-term contracts to open-ended contracts. Years 2011-2017.

	Years						
Age	2011	2012	2013	2014	2015	2016	2017
< 21	1,142	1,264	943	836	1,651	1,077	886
21-25	5,027	5,579	4,443	3,901	7,382	5,102	4,279
26-30	6,868	7,203	5,851	5,328	9,654	6,950	5,528
31-35	7,432	7,253	5,852	5,198	8,731	6,179	4,931
36-40	6,737	6,859	5,691	4,810	8,499	5,811	4,523
41-45	5,779	5,921	4,971	4,205	8,025	5,569	4,472
46-50	3,974	4,301	3,640	3,264	6,408	4,747	3,687
51-55	2,204	2,453	2,251	1,957	4,054	3,043	2,505
56-60	927	992	995	868	1,863	1,334	1,106
> 60	240	319	297	287	480	375	329
Total	40,330	42,144	34,934	30,654	56,747	40,187	32,246

apprenticeship positions into permanent are relevant, although relatively little in their amount. In fact, the apprenticeship contract is on average 3.7% of all job positions over the whole period, and the great part of them involved young at age between 19 and 21. Table 5 reports the transition flows from the apprenticeship contract to the open-ended contract, by age for the whole period.

The outcome that is important to underline is that the transition between the two job positions hastened over the period and involved the youngest apprentices towards the end (Table 5).

4. CONCLUDING REMARKS

Two key issues are addressed in this paper: 1. the effectiveness of political interventions that occurred in Italy over a period of 18 years (starting from 1997) to radically reform the Italian labour market to stimulate employment and production; 2. potentialities and opportunities of an integrated information where official statistical and administrative data are aligned to examine complex socio-economic phenomena. In this work, the data warehouse I.Stat from ISTAT and the administrative database provided by Veneto Labour Office are used to analyse the Italian labour market over the period 2011-2020, which embraces the promul-

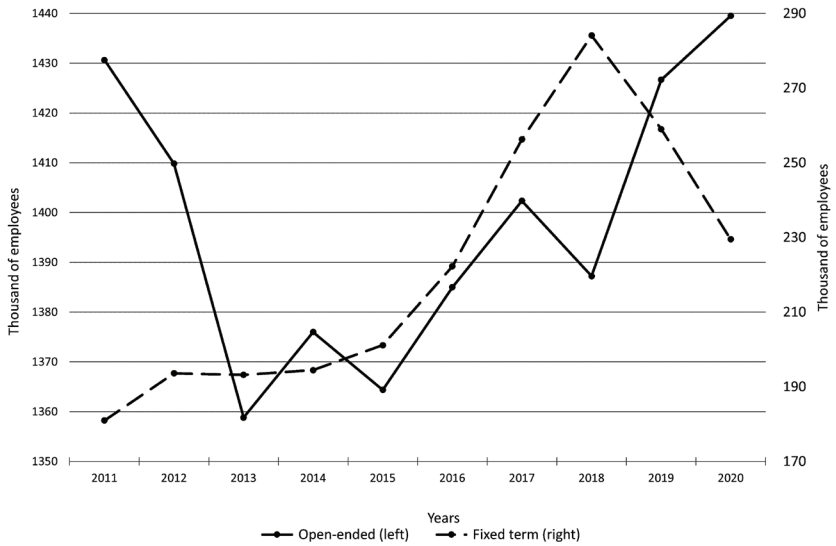
Table 4: Veneto transformation flows from the fixed-term contracts to the open-ended contracts compared to the fixed-term contracts (per cent values). Years 2011-2017.

	Years						
Age	2011	2012	2013	2014	2015	2016	2017
< 21	6.0%	7.0%	5.4%	4.6%	8.3%	5.0%	2.9%
21-25	9.1%	10.6%	8.1%	6.7%	13.0%	8.3%	5.7%
26-30	10.4%	11.5%	9.8%	8.4%	15.2%	10.6%	7.2%
31-35	10.6%	11.0%	9.0%	7.6%	13.8%	10.2%	7.3%
36-40	10.3%	10.9%	8.9%	7.1%	13.1%	9.7%	6.9%
41-45	10.1%	10.4%	8.5%	6.7%	12.7%	9.2%	6.6%
46-50	9.1%	9.6%	7.5%	6.1%	11.9%	9.0%	6.1%
51-55	8.0%	8.5%	7.1%	5.4%	10.4%	7.7%	5.3%
56-60	6.1%	6.1%	5.8%	4.5%	8.8%	5.9%	3.9%
> 60	2.6%	3.4%	3.0%	2.7%	4.1%	2.7%	1.5%
Total	9.4%	10.1%	8.2%	6.7%	12.4%	8.8%	5.9%

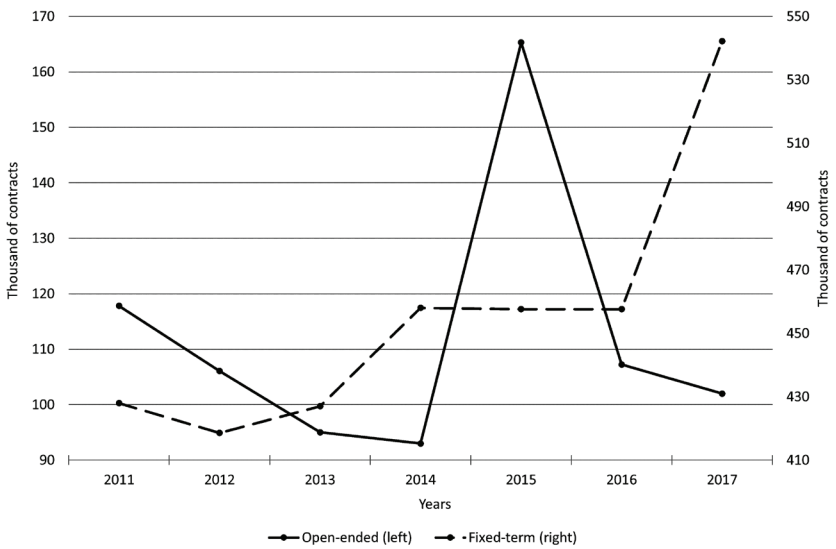
gation of the two more debated Italian reforms, the Fornero Reform (2012) and the JOBS Act (2015). The approach we used in this work is a graphical-matrix method to investigate the labour mobility: transition matrices jointly used with a graphical representation of the labour market flows. The outcome we obtained allowed to conclude that the analysis of the Italian labour market is detailed when the undisclosed data from administrative databases, as the Veneto Labour Office case, are available to complete the official statistical information, as provided by the national institutes of statistics. This would allow to highlight the interconnected and overlapped effects of any juridical and monetary interventions that differently are not affordable. This work is still a preliminary experiment conducted to test opportunities to solve some statistical methodological issues still open in the scientific debate referred to the integration of administrative data and official data to analyse complex social p henomena. Longitudinal statistical data on labour market provided by ISTAT, recently available on request, are definitely a chance to resolve the information gap that the available open data experience. Further investigations will be conducted in this sense to update this our first analysis alongside a restricted causal analysis of data from Ven.DB to investigate the connection between the labour market policies and the employment trends.

Table 5: Veneto transition flows from the apprenticeship contract to the open-ended contract. Years 2011-2017.

	Years						
Age	2011	2012	2013	2014	2015	2016	2017
< 17	1	21	28	14	13	3	8
17	15	64	131	102	135	81	43
18	36	104	192	267	411	267	186
19	84	209	445	673	973	655	458
20	135	299	701	1,089	1,664	1,140	985
21	147	344	577	995	1,329	1,003	892
22	181	355	461	721	1,032	812	716
23	144	304	419	605	855	750	636
24	134	246	337	528	848	695	666
25	111	228	355	539	839	849	691
26	103	214	342	478	799	735	643
27	80	188	228	423	644	566	556
28	62	134	201	332	534	516	422
29	38	102	146	211	377	414	350
30	22	55	73	100	195	224	217
> 30	21	31	8	9	10	20	6
Total	1,314	2,898	4,644	7,086	10,658	8,730	7,475



(a) data source: LFS.DB database.



(b) data source: Ven.DB database.

Figure 4: Veneto employment trends in open-ended and fixed-term contracts. Years 2011-2020 LFS.DB (a): thousand of employees. Years 2011-2017 Ven.DB (b): thousand of contracts.

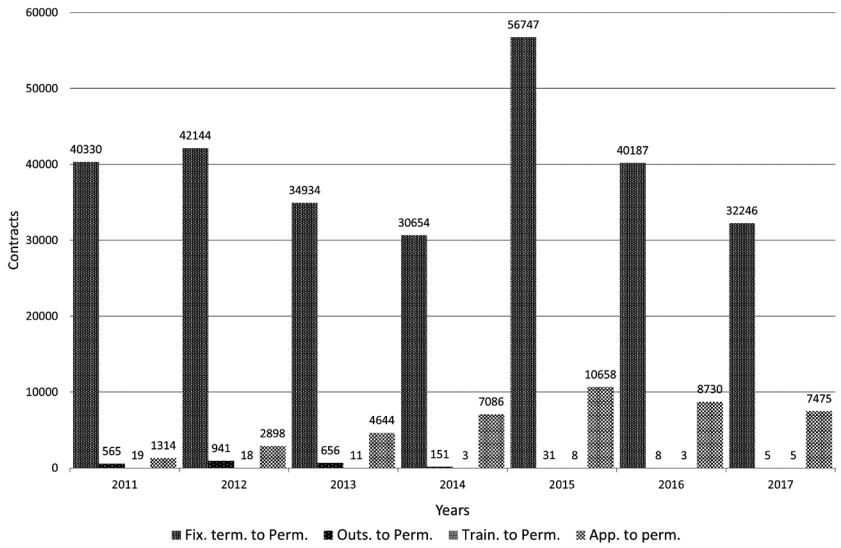


Figure 5: Veneto transition flows from temporary job contracts to permanent job contracts. Years 2011-2017.

References

Albisinni, M. and Discenza, A. (2004). *La mobilitá dell'occupazione e della disoccupazione dalla seconda parte degli anni Novanta*. Sistema Previdenza, XXI. N. 3, pp. 47-67.

Battisti, M. and Vallanti, G. (2013). Flexible wage contracts, temporary jobs, and firm performance: Evidence from Italian firms. In *Industrial Relations: A Journal of Economy and Society*. 52: 737-764.

Bernardi, L. and Zaccarin, S. (1984). *Indicatori di mobilitá: applicazione di un modello markoviano ai dati della rilevazione trimestrale delle FL.* In Schenkel, M. (a cura di), L'offerta di lavoro in Italia. Venezia: Marsilio.

Bernardi, L. and Zaccarin, S. (1991). *La stima dei flussi e di matrici di transizione*. In U. Trivellato (a cura di), Forze di lavoro: disegno dell'indagine e analisi strutturali. Annali di Statistica, Serie IX, Vol. 11. Roma: ISTAT.

Blumen, I.M.K., and McCarthy, P. (1995). *The Industrial Mobility of Labor as a Probability Process*. Cornell Studies in Industrial and Labor Relations. Vol. 6.

Boeri, T. and Garibaldi, P. (2007). Two tier reforms of employment protection: A honeymoon effect? In *The Economic Journal*. 117: 357-385.

Cirillo, V., Fana, M., and Guarascio, D. (2017). Labour market reforms in Italy: evaluating the effects of the jobs act. In *Economia Politica*. 34: 211-232.

Daruich, D., Di Addario, S., and Saggio, R. (2023). The effects of partial employment protection reforms: Evidence from Italy. In *The Review of Economic Studies*. 90: 2880-2942.

Hoffmann, E.B. and Malacrino, D. (2019). Employment time and the cyclicality of earnings growth. In *Journal of Public Economics*. 169: 160-171.

Marini, C. and Nicolardi, V. (2017). Database del mercato del lavoro a confronto: possibile integrazione per una analisi dinamica dell'occupazione. In *Metodi e analisi statistiche*. 127-149.

Marini, C. and Nicolardi, V. (2018). Livelli e probabilità di transizione del mercato del lavoro: alcune evidenze degli effetti del jobs act sull'occupazione in italia. In *Economia, Istituzioni, Etica e Territorio. Casi di studio ed esperienze a confronto*, 55–78. FrancoAngeli.

Marini, C. and Nicolardi, V. (2021a). Administrative database and official statistics: The case of the real estate analysis. In *Statistica Applicata - Italian Journal of Applied Statistics*. 33: 83-95.

Marini, C. and Nicolardi, V. (2021b). Big data and economic analysis: The challenge of a harmonized database. In P. Mariani and M. Zenga, eds., *Data Science and Social Research II*. Springer International Publishing, Cham: pp 235-246.

Sestito, P. and Viviano, E. (2018). Firing costs and firm hiring: evidence from an Italian reform. In *Economia Politica*. 33: 101-130.

Tealdi, C. (2011). Typical and atypical employment contracts: The case of italy. In *Occasional Papers 39456, Munich Personal RePEc Archive*.

Tomelleri, A. (2021). Temporary jobs and increasing inequality for recent cohotrts in Italy. In *LABIUR*. 35: 500-537.

Ward-Warmedinger, M. and Macchiarelli, C. (2013). Transitions in labour market status in the eu. In *IZA Discussion Papers 7814*. Institute for the Study of Labor (IZA), Bonn.

doi.org/10.26398/IJAS.286

INTERVIEWING ADMINISTRATIVE RECORDS

A CONCEPTUAL MAP FOR THE USE OF BIG DATA FOR ECONOMIC RESEARCH

Roberto Leombruni
Department of Economics and Statistics "Cognetti de Martiis", University of Torino, Italy

roberto.leombruni@unito.it

ORCID: 0000-0001-8816-2407

INTERVIEWING ADMINISTRATIVE RECORDS

A CONCEPTUAL MAP FOR THE USE OF BIG DATA FOR ECONOMIC RESEARCH

***Abstract**. Businesses, academia and official statistics are turning more and more to novel data sources besides traditional sampling surveys, but the debate about their defining features and the challenges they pose for research is still open. In this paper I propose a conceptual map of what data are in the field of empirical economic research to clarify what are the conditions and possible strategies to fully grasp their opportunities, particularly in the case of big data of administrative origin. The conceptual map is inserted into a recent literature addressing the clarification of the very notion of data, and exemplified using three well-know cases of failures and best practices in the use of large data samples. The conceptual map is then used to discuss the case of labour market research based on social security data.*

***Keywords**: Big data for research purposes; Notion of data; Administrative data on the labour market.*

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.
Competing Interests: The Author(s) declare(s) no conflict of interest.

Ask not what you can do to the data,
but what the data can do for you
Zvi Griliches

1. INTRODUCTION

The use of novel data sources besides traditional sampling surveys is gaining a primary role in official statistics worldwide. In Europe, an important step in this direction was taken in 2014 with the European Statistical System vision 2020, which identified in a wider and better use of new sources such as administrative and geospatial data a key strategy to answer to the challenges that official statistics was facing. The rationale of this tendency has indeed to do with the cost reduction implied by the re-use of data already collected. From a statistical point of view, however, the main interest rests on the benefits side, and in particular in the possibility of exploiting the wealth of information held by public institutions, e.g. in their fiscal or social security registers, or by private organisations such as telecommunication companies or financial institutions. The unprecedented size and timely amount of information characterising these data sources has the potential to allow the publishing of population-based real-time statistics, on themes and/or with an accuracy that in many cases are out of reach for traditional surveys.

In a somehow parallel way to what is happening in official statistics, the exploitation of administrative data is more and more characterising also academic research. Already ten years ago Raj Chetty was noting that in a leading economic journal such as the American Economic Review the share of articles using micro-data based on surveys went down from about 60% in 1980 to 30% in 2010, mirrored by an increase in articles based on administrative data from 30% to 60% of all publications. A swap in relative importance even larger in the case of the Quarterly Journal of Economics, with survey based articles going down from over 90% to 10% and administrative data based ones from about 10% to 70% (Chetty, 2012). In Italy, first examples of this are the studies carried out already starting in the mid-Eighties exploiting the National Social Security Administration archives to produce novel evidence on previously unexplored matters such as firm demography, job creation and destruction and earnings differentials (Contini and Revelli, 1986, 1987). A back of the envelope estimate says that currently about 70% of all empirical studies published about

Italian labour market are based on INPS administrative data, versus 30% on ISTAT's labour force survey[1].

The trend in using "alien" data sources with respect to statistical surveys further accelerated in the last decade with the advent of the so called big data. Two of the three "V" that are usually quoted to define them (volume and velocity) are indeed a key feature also of administrative archives of public and private organizations. To them, the "V" pointing to the variety of new and different kind of information, such as image- and textual data, took the hotspot in businesses and also in economic research, with applications ranging from the use of satellite imagery to obtain local area estimates of poverty and industrial development (Engstrom *et al*, 2021; Soren and Fisker, 2018), to the use of Twitter posts to predict labour market flows (Antenucci *et al*, 2020) or regional unemployment rates (Llorente *et al*, 2015), to the large and growing literature using on-line platforms for job advertising to estimate the dynamics and skill composition of labour demand (see e.g. CEDEFOP, 2019, and Khaouja *et al.*, 2021, for a review).

It is fair to say, however, that the surge in the use of big data – including administrative ones – has not been accompanied by a matching methodological literature investigating not only the opportunities they offer but also the challenges they pose. Still in 2007, in one of the first systematic studies on the use of administrative registers within official statistics, the authors noted that no well-established theory in the field existed (Wallgren and Wallgren, 2007). Their point was that while statistical surveys are a well-known object in research, thanks to established methodologies grounded on probability theory and inference, no comparable terms or principles were available to provide grounds to a systematic theory on statistical systems based on registers. From this point of view, the current integration of new sources of data within the production process of official statistics is providing the best setting for their study, and indeed we are witnessing many advances in the literature,

[1] Searching in EconLit all papers published in the last 10 years about Italy containing any of a large set of labour-related keywords (labour market, work, employment, unemployment, job, workforce, wages, retirement, pension system, welfare system), combined with either "INPS or Administrative Data" or with "LFS or Labour Force Survey".

particularly in the field of data integration, data quality and total error estimation (see e.g. the essays in Hill *et al*, 2020).

Outside official statistics, however, the situation is more blurred. The possibility of linking administrative data to statistical samples of a well-defined target population would offer also to researchers a solid base for statistical inference. But, also in view of current regulations about data protection and integration, this kind of anchorage is seldom available to the researcher. If one had to review in more detail the kind of data used in academic research, most of the studies would be probably classified as stand-alone applications of administrative data, i.e., as a direct use of an exerpt of administrative archives without their integration into a statistical survey. In Italy, two exceptions to this are the studies on socio-economic inequalities and health based on the linkage of administrative data on health to ISTAT's surveys (see e.g. Ardito *et al*, 2020; Petrelli *et al*, 2022) and the T-DYMM microsimulation models of Italian social security system based on the linkage of the Italian section of EU-SILC to INPS administrative data (see e.g. Conti *et al*, 2023).

From the other side, the recents advancements notwithstanding, the literature about the challenges posed by big data is still scant – on average only about 5 in one thousand articles on big data deals with epistemological aspects (Balazka and Dario Rodighiero, 2020) – and far from having reached maturity even about the very definition of its subject matter. Recent reviews of the literature actually reported the existence of different and sometimes contrasting views about what big data are (see e.g. Connelly *et al*, 2016; Fosso Wamba *et al*, 2015; Al Sai *et al*, 2019). But even besides the issue of identifying where the boundary lays between different typologies of data sources, what is somewhat surprising is the lack of an accepted consensus even about the very notion of "data" also without the "big" part.

Actually, the lack of an agreement about what the term "data" means was the point of departure of a highly quoted article in the information science literature already thirty years ago, which proposed a definition of data as a triple <entity, attribute, value> which is still today a common one in database theory (Fox et al, 1994). Point is, more than ten years later a panel composed of information science scholars reported a list of 42 different definitions of information, data, and their mutual relationship (Zins, 2007). The definitional issue is still today the focus of a large and interesting debate within the larger

Literary and Information Sciences (LIS) literature, part of which will be discussed in a section below as a background for the current contribution.

The aim of this paper is more limited in scope with respect to a fully discussion of different definitions of data and big data. It is to propose a conceptual map of what data are in the field of empirical economic research, as a basis to discuss what are the challenges for the use of big data in this specific field and what are the conditions and possible strategies to fully grasp their opportunities. To exemplify my argument, in next section I will present three examples of "going big" in applied research. I will then present the conceptual map, locating it in the current literature about the epistemological nature of data (Sections 3 and 4). In Section 5 I will provide some examples of the proposed strategy in the case of labour market research based on social security data. The final section will propose some concluding remarks.

2. OLD AND NEW BIG DATA STORIES

Technological aspects usually play a big role in the narrative about big data, as their use is strictly linked to several innovations we witnessed to in last decades in computing power, storage capabilities and cloud technologies (Al-Sai *et al.*, 2019). Besides the current hype on the technological innovations, however, at least the "V" referring to the volume of large datasets is out there from pretty some time. At the very beginning of modern statistical enquiry, we may say that it is more the US Census that triggered improvements in computing technologies – such as the use of punched cards for the storage and processing of data – than the reverse.

Also the first story I'm proposing here is an old one, recounted in many statistics handbooks: The large scale election poll by the *Literary Digest* (LD) during the US presidential campaign in 1936. The LD was a popular magazine which had actually been succesfull in predicting all previous presidential elections. In 1936 they decided to sample an astounding 10 million US citizens, collecting 2.4 million answers, which is an amazing sample size also for today's standards. Their prediction was a huge victory for Alfred Landon, the republican candidate, over the incumbent president Franklin D. Roosevelt, but the result was simply the opposite: a landslide in favour of Roosevelt, who gathered 98.5% of the electoral votes – the largest victory ever in US history.

The number-one suspect for this epic fail was the sampling frame, which was based on lists of telephone and automobile owners, which resulted in the sampling of wealthier than average, pro-republicans voters. Also nonresponse bias was identified as an important factor: a recent reassessment of the matter found that pro-Landon voters were more keen to participate to the LD poll with respect to pro-Roosevelt ones (see Lusinchi, 2012, and Lohr and Brick, 2017). With a bit of a joke, an article titled *Digest Digested* appeared in the *Times* magazine in 1938 reporting the epilogue of this story: the *Literary Digest* ended its publishing history being absorbed by the *Time* magazine itself.

The next two examples are not relative to social inquiry or economics, rather to the rising field of epidemic intelligence, but they too are illustrative of the promises and perils of "going big".

The first is another well-known example, probably the first time a modern big data analytics approach gained the highest stand in scientific research, thanks to a paper published in *Nature* predicting influenza epidemics using search engine query data, the so-called *Google Flu Trends* (GFT, Ginsberg *et al.*, 2009). Building on Polgren and co-authors (2008), who already detected a correlation between virological surveillance data and search queries containing the words "flu" or "influenza", they aggregated historical logs of web searches for 50 million of the most common queries in the US, to build a system which consistently predicted epidemic outbreaks 1-2 weeks ahead of the official surveillance reports. Just four years later, however, the GFT was closed, since it was predicting almost double the number of doctor visits subsequently realised. There have been several explanations for this, including changes in user behaviours and in search engine functioning (see Shin et al., 2016, for a review). Lazer and co-authors proposed also an overfitting issue – the 50 millions records were used to fit as few as 1,152 real observations – plus two considerations. The first is a general point which is sometimes forgotten in what they called the "big data hubris": It's simply not just about the size of data. The second anticipates some aspects I'll be dealing with in the next section: "All empirical research stands on a foundation of measurement. Is the instrumentation actually capturing the theoretical construct of interest? Is measurement stable and comparable across cases and over time?" (Lazer *et al.*, 2014, p. 1204).

The last story – a successful one – is about the Artemis project of the University of Ontario, aimed at preventing disease spread in neonatal intensive

care units (NICU, see Blount *et al.*, 2010; McGregor *et al.*, 2011). The project was based on the real-time analysis of patients' data streams to identify conditions preceding the onset of medical complications. The data gathering included measures such as ECG readings, respiratory rate, blood-oxygen saturation and blood pressure metrics, for about three million data points per hour for each infant in the NICU. The interesting point here is not only the quantity and velocity of information gathered. The "on-line analysis" of data, through the automatic application of clinical rules pointing to possible medical complications, provided clinicians with a decision support based on a stream of data too large to be assessed with a traditional, "off-line" scrutiny of the same information. The deployment of the entire data gathering and analysis pipeline resulted in early warnings of infection spreading 24h before with respect to the traditional approach.

3. THE NOTION OF DATA

Tim Harford, in a lecture given in 2014 at the Royal Statistical Society, discussed the LD and GFT examples proposing two readings of the challenges posed by big data (Harford, 2014). The first is the "theory-free" risk somehow embedded in going big. As a representative example of this attitude he proposed a quote of the provocative *Wired* essay by Anderson (2008), *The End of Theory*: "With enough data, the numbers speak for themselves". This is actually a rather general point and a recurrent discussion in economics and statistics, that we could view as a modern reprise of the Koopmans *versus* Vining "Measurement without Theory" debate in the 1940s and 1950s (Koopmans *et al.*, 1995). The second is more specific to our theme, and is the emphasis he put on one of the defining features of big data, which he labels as the "digital exhaust" of web searches, mobile trackings and credit card payments, proposing "found data" as a (more gentle) way of identifying one of their common characteristics[2].

[2] The term "found data" was actually already common in the AI literature, where the availability of large corpora of "found", textual and audio data were key to the first achievements in speech recognition and natural language processing long before the term "big data" was even introduced (see *e.g.* Gauvain *et al*, 2000; Hirschberg and Manning, 2015). See also the quote by Griliches below.

The fact that the information used by a researcher has not been "made" on purpose for statistical use, but "found" somewhere as the outcome of a process with different purposes – being them the ones of a public administration or of an individual browsing the internet – is clearly relevant to the point. It does not help however to discriminate between the examples we presented: The Artemis project, a brilliant application of a big data approach, is by no means "found data"; the *Literary Digest* was based on "made" data, but it was wrong; the GFT was based on data collected by Google itself, so in this case the very difference between "made" and "found" is a bit fuzzy.

Griliches, anticipating this very theme, already used the term "found data" to note that most of the work in econometrics is based on data that have been collected and assembled by somebody else, often for quite different purposes, including statistical ones (see e.g. Griliches 1984, 1986; see also Triplett, 2007). He noted that this is not a bad *per se* but, rather, a defining feature of much of economic research. When data were perfect the very discipline of econometrics would possibly not exist: the "existential problem" of econometrics is "life with imperfect data and inadequate theories" (Griliches, 1984). His main point – as in the Lazer and co-authors comment quoted above – is again a stress on the issue of measurement. It is because data are imperfect that it is important to consider at least two different data generation processes: the economic model describing the behaviours of economic actors and the measurement model describing how this behaviour is recorded. "While it is usual to focus our attention on the former, a complete analysis must consider them both" (Griliches 1985, p. 198).

If from the one side Griliches is noting that in many cases data is "found", we can rephrase his point about the importance of the measurement process by saying that actually all data is "made". As we now will see, the prominence of the "made" aspect is one of the points which is actually emerging in the current literature trying to clarify the very notion of data, particularly in these years where more and more social and economic events are leaving a "digital exhaust" so that possibly anything is becoming data.

Fox and co-authors, in their early assessment of the matter, pointed to three different approaches to define data (Fox *et al*, 1994).

i. Data as "raw facts"
ii. Data as a triple <entity, attribute, value>
iii. Data as a result of measurement or observation.

A recent literature has taken up the issue of clarifying the notion of data adding details to this classification, particularly in view of the complexities brought about by the diffusion of big data (see e.g. Floridi, 2008; Borgman, 2015; Frické, 2015; Leonelli, 2015 and 2019; Hjørland, 2018; Gellert, 2022). A full review of this literature is out of scope for the present paper: for the sake of my argument I will stick to the i-iii views, which are still today the most common ones, recalling how the recent debate has contributed to clarify their differences.

The view of data as "raw facts" is the closest one to the Latin etymology of the term, *datum*. Data is "what is given", is a set of raw facts about a phenomenon which are the base of our arguments or elaborations. This is a very general definition of the term valid in common speech but also a pretty common one, for instance, in the studies within information science using the "DIKW Pyramid" conceptual map (Data, Information, Knowledge, Wisdom) in which data is the raw material on which information is built[3] (Zins, 2007). The most recent literature discussing what data are from an epistemological point of view however tends to critique this view, stressing that the process of scientific discovery does not produce "objective knowledge" about phenomena, since the scientific process itself is "theory-laden", a concept originally put forth by Pierre Duhem and recently applied to the issue of defining data by Leonelli (2015). Looking at actual scientific praxis, it has also been noted that data are always variously "cooked" within the circumstances of their collection, storage, and transmission, so that the "raw" label is actually an oxymoron (see the essays collected in Gitelman, 2013, particularly the one by Bryne and Poovey recounting Fisher's "data scrubbing" in its pioneering contributions to financial modelling). Besides the "raw" label, also the very idea of a factual piece of information "trades one difficult concept (data) for an equally difficult one (facts)" (Floridi, 2008).

The view of data as "raw material" is actually a reasonable one from the point of view of information systems design. As an example, when designing a datawarehouse, there is a clear point of entry for the data the system is based

[3] Quering "Data" in InfoScipedia, a large database of terms and definitions in Information Science and Technologies, 15 in 68 entries explicitely stress the "raw" aspect of data. At https://www.igi-global.com/dictionary/, retrieved October 4th, 2023.

on, and all the procedures for their elaboration and visualisation take them as "what is given". Also in this context, however, definition i) misses the necessary details to discuss in a rigorous way, among others, the quality dimension of data, which was one of the starting point of the alternative approach proposed by Fox and co-authors in their seminal contribution. The same authors provided a more recent outline of this view, widely accepted in the database community, which is actually a collection of definitions of several related items. Within this view, a *datum* or *data item* is an ordered triple <*e*, *a*, *v*>, asserting that the entity *e* has the value *v* for its attribute *a* (Redman *et al*, 2017).

The examples provided in previous section all fit well in this definition. In the *Literary Digest* poll, entities are individuals, the attribute is the intention to vote, the values are the actual intentions to vote of each individual. The case of *Google Flu Trends* is pretty similar, as long as we see web searches as individuals putting queries into a search engine "ballot". The only difference is that the possible values of the attribute (the web search) is not the close list of candidates of an electoral campaign, such as in a multiple choice question, but free text such as in an open-ended one. The Artemis project is coherent with this definition, too: The entities are the infants and the automatic medical readings populate a set of several <*attribute*, *value*> couples.

This formal definition of data is also pretty similar to the one provided by the Royal Society, together with definitions of information and knowledge in a reduced version of the DIKW model, which define them as "numbers, characters or images that designate an attribute of a phenomenon" (Boulton *et al*, 2012, p 12, cit. by Leonelli, 2015). While this view of data has the double appeal of being intuitive and to avoid the use in a circular way of the notion of fact, it has actually the same limits of definition i). In concrete research situations data are never an abstract object defined in terms of their intrinsic properties, rather, they are defined in terms of their function within specific processes of inquiry. This is an argument used by Leonelli (2015) in order to argue in favour of a "relational" definition of data. The stress on the process of inquiry brings again the issue of measurement: Also assuming there is such thing as a fact, it cannot consist of just the value of an attribute, it needs information about the way it was collected. This kind of description, that in the database terminology would be classified as metadata, is an integral part of

approach iii), which defines data explicitely as a result of measurement or observation.

An early and interesting account of data as observations can be found in Yovit's seminal paper trying to define the field of Information Science (Yovit, 1969). His starting point are "observable actions", i.e., quantities which are physical in nature such as the position of an aircraft, the result of a scientific experiment, a new product developed by a firm. As such, they are neither information nor data. In order to become data they must be transformed by a function (which Yovit calls the "T function"), which is fundamentally a measuring device which transforms the observable actions to data. Fox and co-authors, although acknowledging the profound importance of the way data is obtained, object that there are common examples of data which are not obtained by observation but are "assigned" to an entity, such as someone's name or social security number. As I will argue below, this is an objection which is easily takled with, for the sake – so to speak – of not throwing out the baby (the measurement issue) with the bath water.

Hjørland (2018) provides a recent assessment of this view, summing up the many contributions which described the nature of data not as "given" but as *capta*, i.e., "taken" and constructed, including the nice seminal posing of the matter by Jensen (1950):

> It is an unfortunate accident of history that the term *datum* (Latin, past participle of *dare*, 'to give') rather than *captum* (Latin, past participle of *capere*, 'to take') should have come to symbolize the unit-phenomenon in science. For science deals, not with 'that which has been given' by nature to the scientist, but with 'that which has been taken' or selected from nature by the scientist in accordance with his purpose.

This is why the metadata describing the process by which data has been collected, including the purposes and perspectives specific to the research activity which originated them, are not an accessory but rather an indispensable element for the use and re-use of databases, particularly in big data research (see Leonelli, 2014, and the discussion in Hjørland, 2018).

What is then "taking data" in statistics? It is a process by far more elaborate that the operation of a single, however complex measuring device to record physical or biological measures. Adrian Smith, in his presidential

address to the Royal Statistical Society in 1996, discussed what is the possible contribution of statistics for the development of an evidence-based society, in which "informed quantitative reasoning" is the base of public debate and of decision-making in government and business. He proposed a view of statistics as "the science of doing science", "whose role is to provide theory and protocols to guide and discipline all forms of quantitative investigatory procedure" (Smith, 1996). He went on proposing sort-of a check list of the tasks needed to produce reliable quantitative evidence, including:

1. The framing of questions
2. Design of experiments or surveys
3. Drawing up protocols for data collection
4. Collection of data
5. Monitoring compliance with protocols
6. Monitoring data quality
7. Data storage, summarization, presentation
8. Stochastic modelling
9. Statistical analysis
10. Model criticism and assumptions assessment
11. Inference reporting
12. Use of results for prediction, decision-making or hypothesis generation

What is missing in a view of data as "given", such as in definition i), and what is hidden in a formal view of them such as in definition ii), are all the activities from 1 to 6, characterising how statistical data are gathered, which is a mix of statistical theory and of technical competencies and process management. In principle, if the big data revolution had just to do with the size of data but all the phases of data procurement followed an adequate standard, no harm is out there. One could even forget, so to speak, statistical inference: recalling *The end of theory* in previous paragraph, "With enough data, the numbers speak for themselves". Strictly speaking, however, the size of data has to do mainly with point 2 in Smith's list, about the sampling of the population, but all other points are equally important to generate reliable evidence.

In fact, while the three examples in previous section were not easily distinguishable from the point of view of i) and ii) notions of data, the accent on the data production process cast more clear differences among them. Weak theory was indeed one of the piftalls in the *Literaty Digest* case, due to a poor choice of the population frame (point 2), together with a complete absence of monitoring of unit non-response (point 6). In the case of Google Flu Trends, in

the absence of a full documentation of the process, we can say few about points 2 to 6 – which is already a crucial "missing-metadata" issue for their use to inform decision-making. About point 1 on the framing of questions, what is sure is that it was entirely unspecific. Technically, the "question" was "What are you serching for on the web?", with completely unstructured, open-text "answers". Since the question was not directly addressing the theoretical construct of interest, no guarantees about the stability of its relation with actual disease spreading was granted.

In the case of the Artemis example, reading the methodological articles describing the project gives a sense of a huge and specific work on the design and implementation of all phases of data-procurement. From the point of view of Smith's account of how statistical evidence is produced, the Artemis big data project is way more a text-book example of "traditional" statistical enquiry with respect to the *Literary Digest* sampling survey.

4. A CONCEPTUAL MAP FOR THE USE OF FOUND DATA

I here present a simple conceptual map of data as *capta* that serves as a basis to discuss the use for scientific purposes of data not gathered under our direct control.

I take from Yovit (1969) the idea that data are produced by means of some "*T* function" applied to "observable actions". About the latter, the qualification "observable" is a bit of a tautology, while "actions" can be restrictive, so I will say in a more generic way that the research activity has to do with some phenomenon of interest, which I call φ. To denote in a more generic way also the process going from φ to a numerical representation of it I will use the term "map". Indeed, "function" and "map" are terms often used interchangeably, but the latter delivers more directly the idea that we are representing some aspects of the phenomenon of interest, as in geographical maps. Besides, sometimes the term function is used defining at the same time a specific codomain – such as a function mapping into $\Re$ – while in times of big data the codomain is often unstructured, such as in sound, textual or image repositories. For the same reason, I will not adopt at this level the representational approach stating that the result of the map is a triple $< e, a, v >$. Rather, the mapping of the phenomenon under study produces a couple that pairs φ to some digital representation of it. I hence define data as a labelled set of digits resulting from

the application of a map on a phenomenon under study, as in figure 1 below. When the set of digits does actually not possess any structure induced by the map, we may define this kind of data as a digital copy of φ, as in the case of textual or image data, whose repository is usually termed as a datalake instead then a database. Such unstructured data can enter *as are* into the stochastic and statistical modelling phases in Smith's check list, as in the case of GFT, or in the case of the application of sentiment analysis techniques to financial news or for public opinion mining (see e.g. Saberi and Saad, 2017, and Man *et al.*, 2019).

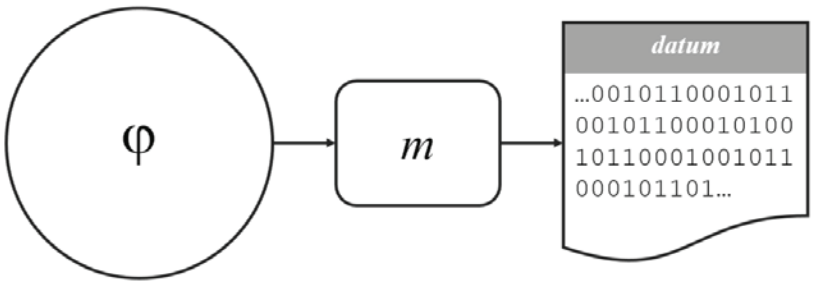


Figure 1: Data as a digital mapping of a phenomenon of interest.

Alternatively, an explicit activity of feature extraction may be implemented before statistical modelling to annotate the digital copies and obtain structured information. In this case, a second map is applied to the digital copy instead than on φ, in order to extract an information set in the form <*attribute*, *value*>, as in Figure 2. In this case, as is common in database jargon, we may talk of a "structure-on-read" schema (Cackett *et al.*, 2013).

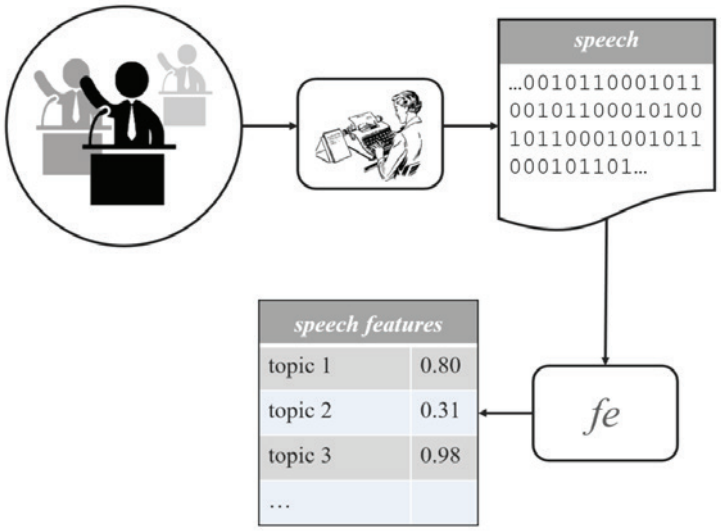


Figure 2: Structure *on read* mapping.

The map may also represent (or measure) the phenomenon of interest directly along a grid of attributes. This is the case of traditional statistical surveys, where the values of the attributes are the answers provided by individuals to an interview, and it is also the case of most administrative data, where e.g. social security contributions paid in favour of workers are registered into transactional databases (figure 3). Internet of Things (IoT) streams of data and the clinical readings in the Artemis project falls in this category, and also in this case it is useful to think to sensor readings as answers to specific questions: The very characteristic of the so-called "structure-on-write" schema is that the data production process tracks exactly the construals needed by the data producer[4].

[4] Note that considering the map *m* as a set of questions avoids the point made by Fox and co-authors that some attributes are "given" instead than "measured", such as social security numbers or names. Strictly speaking we are not measuring them with some device, we just ask.

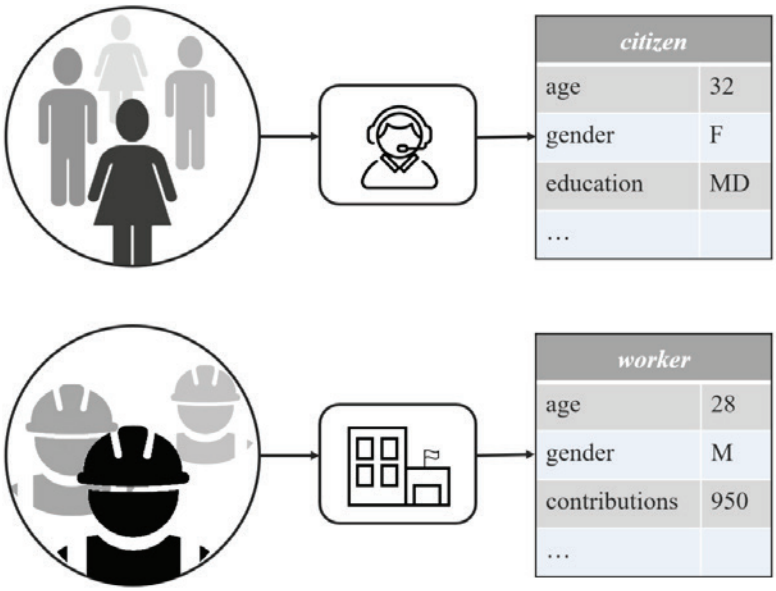


Figure 3: Structure *on write* mapping

Let us use the conceptual map to consider the case in which the purposes of a researcher are not aligned to the ones of the data producer. In general terms, following again Smith's check-list, the main critical aspects involve the survey design, the data quality monitoring and the framing of questions (see e.g. Johnson and Moore, 2005; Wallgren and Wallgren, 2007).

The issues about survey design have to do with what is the scope of the phenomenon of interest mapped by *m*. The representativity of the "found" data with respect to the interests of the researcher is presumably one of the most difficult issues in using, e.g., social media data to study public perceptions, and a key issue also with economic data of administrative source, e.g. in dealing with informality. From this point of view, however, provided that the map *m* is coherent with the purposes of a researcher, the proposed conceptualization does not solicit further considerations.

The issues about data quality have to do with the functioning of the map. The main point here is that while in statistical surveys data quality is mainstreamed in a consistent way, in the case of data collected for different purposes not all attributes may bear the same interest to the data producer, and

this usually has a large impact on quality. As an example, incomes are typically recorded with high accuracy in Tax and Social Security records, since it is a key information for the administrative purpose of the agencies collecting data; while the information about education – when recorded at all – has usually lower levels of quality (see e.g. Stüber *et al.*, 2023, and Adriaans *et al.*, 2020, for a study on German data and a review). Also from this point of view, the proposed conceptualization does not solicit further considerations with respect to the general literature about data quality.

Both representativeness and quality, then, may be dealt with during the statistical analysis of the data, e.g. discussing in a proper manner the external validity of the results and modelling errors in variables (as in Griliches' argument).

The third potential issue is about the framing of questions. One of the main disadvantages in using administrative data is that it is the data producer who chooses which questions to ask and typically its interests are very different with respect to the researchers' ones (Wallgren and Wallgren, 2007). A more subtle issue is that also the statistical units which are mapped may be different, and/or the "questions" used in the map may not measure exactly the concepts the researcher is interested in (see again Johnson and Moore, 2005, and Wallgren and Wallgren, 2007; and the thourough study in Kapteyn and Ypma, 2007).

This latter issue cannot be dealt with during the statistical analysis of the data, but requires a pre-processing of the data in order to bring their information content closer to the researcher's purposes. We can modify our conceptual map to represent this situation as in Figure 4. The primary map m is the one used by the data producer to obtain the information needed for its administrative purposes. The researcher, when having the possibility of directly survey the phenomenon of interest, would have used a different map more coherent with its purposes (called k in the figure), collecting possibly different attributes on different entities. In order to use the administrative data for the researcher's purposes, then, a secondary map is needed to tranform the original data in order to obtain an information set as close as possible to the desired one. The secondary map ideally would be $m^{-1}k$, in which case the use of found data would be equivalent to a direct survey of the phenomenon under study.

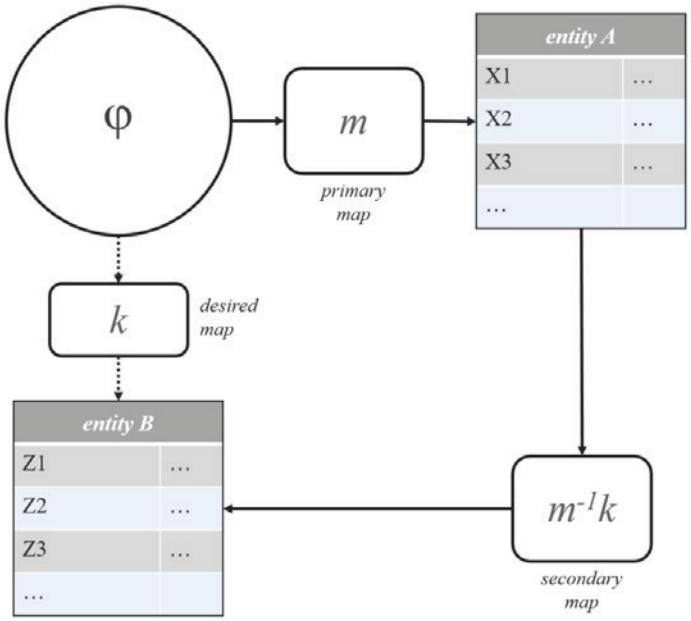


Figure 3: Pre-processing of found data to mimick a direct survey of φ.

In a sense, the secondary map is an "interview" posed not to the individuals in φ but to the administrative data, whose objective is to recover the kind of information that would have been obtained with a direct survey of φ applying the desired map *k*.

5. INTERVIEWING SOCIAL SECURITY RECORDS FOR LABOUR MARKET RESEARCH

In this section I will illustrate the need for a secondary mapping of administrative data in order to use them for research purposes, using as an example the WHIP-Health database on work and health biographies in Italy. The database is developed by the University of Turin and the Epidemiological Service of the Piedmont Region on behalf of Italian Ministry of Health, in cooperation with the many public institutions that have given access to their administrative data. It collects information of a 7% sample of the Italian population, with a longitudinal coverage of about 50 years as far as working careers are concerned, while the information on workers' health covers a more

limited period (from mid Nineties for work injuries and professional diseases and from early 2000s for hospital dismissal forms).

I will focus here in particular on work biographies and the framing of questions issue[5]. The section of the database dealing with working careers is based on the integration of more than 20 different files provided by INPS, the Italian Social Security Administration, relative to different typologies of work arrangements, different welfare provisions and employers' data; plus a database on work contracts' openings, closures and transformations provided by the Ministry of Welfare (MoW). For the sake of simplicity, to exemplify the framing of questions issue I will consider separatedly those regarding the definition of the units of observations (the entities) and those regarding the attributes construals, even though they are indeed linked with each other.

5.1 THE STATISTICAL UNITS

The main entities of interest for labour market studies are the two sides of the labour market itself, i.e. the individuals who offer their work services and the businesses and other institutions who may employ them.

With regard to the former, the use of administrative data entails a first hurdle. In a survey, each entity is extracted by sampling a population register in which a unique personal identifier is included. The data collection process implemented with a survey follows what we might call a top-down flow, in which one starts with the entities and then all the information of interest about them is collected with an interview. With a top-down schema, the integrity of the statistical unit definition and the association with its attribute-values is granted *by design*. In the case of administrative records, the data collection process typically follows a reversed, bottom-up flow. It is administrative transactions which are collected first – such as social security contributions collection or welfare benefits payment – and thanks to a personal identifier such as the fiscal code the data are then associated to the correct statistical unit. The integrity of the statistical unit definition hence rests entirely on the data quality of personal identifiers. Although in current information systems they possess a

[5] See e.g. Contini and Trivellato (2005) for a wider description and for applications of the work histories section of the database, and Bena *et al.* (2012) and Ardito *et al.* (2017) for the section on health biographies.

high level of certification, in longitudinal databases collecting retrospective data from legacy archives the very identification of individuals may be critical, leading to matching errors which corrupt the accuracy usually associated to administrative data (Kapteyn and Ypma, 2007). In this case, the secondary map that may be used is a probabilistic match trying to find apparently different individuals which, due to their personal characteristics and their career patterns, can be identified as the same individual. In the case of WHIP this kind of procedure was run up to the early 2000s, before INPS itself, in cooperation with the Tax Administration, started a systematic data cleansing activity of the fiscal codes used to identify persons in the individual registers.

With regard to employers, the situation is more complex. From an administrative point of view the employer is a legal entity, and no ambiguities are out there: The employer is the entity identified by the unique fiscal code which is paying the social security contributions. From a theoretical point of view, however, a researcher would like to analyse data about employers without being forced to a strictly legal definition of them. Depending on the research question, the interest could be e.g. on the local units of a firm, on firms belonging to a group, or on legal transformations of firms. In the case of Italian data all these details are not present, but some of these topics can be handled with a secondary mapping of the original data. Consider as an example relevant events such as ownership transfers**,** mergers and acquisitions, spin-offs and legal transformations. With the "legal entity" point of view of administrative data, all these events produce apparent firms' start ups and closures, even when there has not been a substantial discontinuity in the life of a firm. This generates "spurious" firm demography events, which hinders both the study of firm dynamics and a correct measurement of workers mobility and of job creation and destruction.

To cure the WHIP data about employers we implemented an algorithm which uses information about the flows of clusters of workers across different legal employers to identify longitudinal relationships in longitudinal business data (Pacelli and Revelli, 1995; see also Hethey-Maier and Schmieder, 2013). In the case of US business data, a similar correction in the statistical unit identification involved about 10-13% of all apparent worker flows (Benedetto *et al.*, 2007); in the current WHIP release this share is lower when compared to all job separations (5-8%), but it is a huge quota of apparently direct job to job transitions (between 40% and 50% in mid 2000s).

5.2 ATTRIBUTES CONSTRUAL

There are several attributes in the social security archives WHIP is based on which require a mapping from the original measures to construals closer to the interest of researchers. As an example, most income variables are not based on the net or gross measure of them – which are the ones typically of interest to researchers – and until recently the sector of activity and skill level were measured using old and/or non-standard classifications. These edits however, although important, do not raise particular methodological issues.

I will focus instead on a topic much investigated in current empirical literature about labour market dynamics, i.e., the long run trend towards an increase in precarious work arrangements. The issue of defining "precarity" is indeed a complex one both from a theoretical and an empirical point of view. A recent assessment of the literature found no widely accepted definition of it, while many operationalisations of the concept were actually an accommodation to the available data (Kreshpaj *et al*, 2020). Whatever the definition, however, it is fair to say that the duration of an employment relationship is a key dimension for the empirical operationalization of the concept.

To measure this important attribute within a sample survey, you can simply ask, as in ILO current recomendations: "*How long have you been employed by your current employer?*". Let us consider instead what are the "questions" available in our source data, as posed by INPS and by the MoW. The latter administration takes as a reference the legal aspect of the relationship, i.e. the labour contract between the employer and the employee. The legal basis for the data collection is the requirement for employers to communicate to the MoW all their hirings and firings, plus eventual tranformations of contracts, including their dates. We could then use this kind of information to directly answer our question about the job duration.

For the years prior to 2009 however MoW data are not available, hence one has to resort to the INPS' source. The collection of social security data does not originate from hirings and firings communications, but on contribution payments' forms, which entail a completely different data transaction[6]. The

[6] Starting from 2005, the digital transmission of contributory data has been radically changed, improving some critical aspects that I'm going to present. Since the WHIP data cover dependent employment starting from 1987, however, all the

issue here is that the relation between contributory forms and labour contracts is actually of a many-to-many type: as an example, when employing a seasonal worker with two different contracts in January and then in December of a given year, a firm would compile one single contributory form. The rationale is: The administration needs just to add the contributions to the previdential account of the worker, no matter if they originated from one or more contracts[7]. In this example, to trade a contributory form for a job contract would underestimate the mobility of workers between contracts, and precarity. From the other side, if a worker stays in the job but there is a change in some aspects of the job itself – e.g. the province of work – the employer has to compile two different contributory forms, leading in this case to an overestimation of workers mobility and precarity. In this case, to derive information about job contracts one needs a complicated remapping of the data, aimed at splitting/joining contributory forms in order to identify (in a probabilistic way) the job contracts which generated the contributory spells.

The situation, however, is more complex than this: similarly to the discussion regarding the entity "firm", one should also consider whether the object of interest for the researcher is really the legal aspect of a job (*i.e.* the employment contract) or some other construal. For the sake of simplicity, let us consider the MoW data, which measure with high accuracy hirings and firings, and consider a worker who had two successive contracts of one year with the same employer. Taking at its face value the MoW datum, at month 13 we would classify the worker as having a tenure of one month, instead than 13, which, depending on the research question, may not be appropriate. A similar issue has been considered in the US' Current Population Survey. Prior to 1983, CPS supplements on tenure asked workers "*When did you start working at your present job?*". The term "job" is itself an ambiguous one: A worker employed for 10 years promoted to a managerial position 1 year prior to the survey may have been counted as having 10 years or 1 year of tenure, depending on

arguments exposed still apply for the procedures handling the older decades of the work careers.

[7] Also the start of the labour contract is actually recorded in contributory forms but it is affected by a huge missing data problem, presumably because of the low administrative relevance of it and of the many-to-many nature of the relation, which implies a non univocal association between contributory forms and work contracts

whether s/he interpreted the tenure with the current employer or in the managerial position. The Bureau of Labor Statistics then switched the wording of the question to a formulation closer to the ILO one quoted above, creating a break in the job tenures' time series (US Bureau of Labor Statistics, 2022).

This is not a minor issue, since successive labour contracts with the same employer are a very common situation in the Italian labour market, with an increasing trend. It is actually now common to hire workers even on a daily basis, with the activation of two or three labour contracts within the same week-end. In the MoW data sample available in WHIP, the share of hirings with contracts which lasted one single day was 12% in the years 2014-2019, and one out of five had a complete tenure within one week. In this case, to consider different legal arrangements as if they were independent one from the other would overestimate the mobility of workers between contracts, and precarity.

To my knowledge, the issue about what is the precise notion (or the set of notions) of employment spell that should be used in labour studies has been rarely discussed, contrary, e.g., to the notion of unemployment. A reference for the debate can be found in the ILO *Employment Relationship Recommendation*, 2006 (EER). The EER is actually focused on the identification of a subordinate condition in cases of casual work arrangements and of concealed employment relationships, but it provides also a reference for a clarification of the concept. The EER operationalise it suggesting that "the determination of the existence of such a relationship should be guided primarily by the facts relating to the performance of work and the remuneration of the worker, notwithstanding how the relationship is characterized in any contrary arrangement, contractual or otherwise, that may have been agreed between the parties". The EER goes on requiring that the relationship has a certain duration and continuity, but with a very loose interpretation of these concepts. In Europe, prevailing interpretations consider as a unique employment relationship sequences of several short-term contracts, even when comprising intermittent working and non-working periods (Risak *et al*, 2013).

Although it apparently delivers a poorer information with respect to MoW data, the recording present in INPS data was actually sufficient to approximate a definition of employment relationships close to the ILO one. To derive this kind of information – i.e. to answer to the question "*How long have you been employed by your current employer?*" – instead of sticking to the information

about the start and end of each work contract one has to "interview" the administrative records, looking for sequences of contributory spells with a certain continuity involving the same individual and the same employer, in order to ascertain how long this employment relationship has been lasting.

6. CONCLUDING REMARKS

Administrative data are an early instance of the family of non statistical sources collectively labelled as big data, delivering structured information with a volume and velocity which granted them a long standing role in official statistics and academic research. The advent of a wave of novel data sources such as IoT and social media data triggered a discussion which is shedding light on issues which were already present in the literature but have long been under-explored, particularly as regards their use for academic research.

In this paper I reviewed this recent debate particularly about the clarification of the very notion of data. A point which has been stressed by many authors is the importance of explicitely viewing data as *capta*, i.e. as an outcome of research activities and not only as a raw material for subsequent analysis. An immediate consequence of this is the importance of documenting not only what data are from a formal point of view – which entities and which attributes they represent. All production process has to be documented, both to make clear what are the theoretical perspectives adopted in their collection and as a basis to facilitate their re-use.

I went on proposing a conceptual map of data coherent with this view, which I used as a basis to focus on the re-use of administrative data for economic research. The main point I stressed is the fact that the different purposes of the institution who "made" the data and the researcher who "found" them reveal themselves already in the framing of the questions at the very beginning of the data production process. This implies that both the statistical units and the attributes available in the data may differ in a substantial way from the ones of interest for the researcher. Using as an example the WHIP database on work and health biographies in Italy, I discussed the case of firm data demography, which in social security files is typically based on a legal definition of employers, implying an over-estimation of firm closures and openings and consequently of job creation and destruction; and the case of

tenure estimation, which again is often based on the legal definition of work contracts, implying a mismeasurement of workers mobility and precarity.

The relevance of these issues rests in the need of studying the functioning of labour markets avoiding the perils of data driven research. From a descriptive point of view, it is indeed interesting to have statistics about the flows of contracts based on their legal definition. Also for research, they may be important e.g. for studies about collective bargaining and the evaluation of labour market legislation. From a perspective focused on employment precarity, a measurement of labour market flows based on the legal representation of them may instead be significantly misleading. An employment relationships with a restaurant or a hotel with a weekend commitment may be based on a single vertical part-time contract or on a recurrent sequence of very short ones. We may well evaluate their relative stability in different ways, but from the point of view of tenure, of human capital accumulation and firm-specific experience, they are hardly distinguishable. In two studies about the impact of tenure and experience on work safety, a definition of employment relationship closer to the ILO construal allowed to detect that one of the health costs of precarious work is mediated by short average tenures and the shift between different employers and tasks (Giraudo *et al.*, 2016, and Bena *et al.*, 2013). A measure of the same risks based on a legal definition of tenure would have implied a mis-classification of workers, hiding or attenuating this potential health spillovers of labour market flexibilization. Similarly, a measure of unemployement based on a tentative reconstruction of the statistical definition of it allowed to identify a causal impact of long unemployment spells of cardiovascolar health, which was cancelled out sticking to the apparently precise administrative measure of it based on unemployment benefits recipiency (Ardito *et al.*, 2017). As Tukey put it in his seminal work which anticipated modern data science, it is "far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise" (Tuckey, 1962).

It is a bit of a paradox that the clarification of the notion of data has received several contributions coming from biology, literary and information science and philosophy, while statistics and data science are among the few fields in which we do not ask ourselves what data are, and when we do it we stick to a definition of it close to the prevailing one in database theory. This

may be because the process of statistical evidence production was already grounded on established theories and practices which materialized themselves into specialised institutes for the delivery of statistical surveys, as summed up by Smith (1996).

The advent of big data and the diffusion of stand-alone uses of administrative data requires to put back the accent on the data production process and on the purposes and perspectives on which it is based on. As Griliches put it, the first question we should pose is not which models we can run on the data, but what are the data telling to us, since it is not us who framed the questions.

This is certainly not an easy task. Among the challenges to fully grasp the opportunities of big data, there are the issues linked to the data protection legislation ruling the processing of personal and sensitive data. The two conflicting aims of privacy protection and data needs for scientific research currently translates in a trade-off between the accessibility to anonymised microdata distributed as public use files, but poor in information detail, and the richness of confidential data that is however accessible – when accessible at all – only on the premises of the data holder. Safe and accessible data has to pay the price of a substantial reduction in the information content available in the source data (Trivellato, 2019), and this hinders not only their use but also an effective activity of data wrangling and understanding.

The regulations about personal data is actually a wider issue, posing a potential obstacle to the full exploitation also of census and survey data particularly in the field of economic policies evaluation (see Crato and Paruolo, 2019, for a recent assessment). A further point, more specific to the use of administrative data, is what we can define a "missing-metadata" issue. It is a common experience for practitioners in the field that data are provided "as are", without a full clarification of the concepts besides a bare schema of the database of origin and of the query which was used to fetch data. The point has not a straightforward solution, since the framing of questions, in the case of public administrations, is embedded into complex layers of different laws, decrees and regulations more than on choices from the part of actual data managers. To "interview" administrative data, trying to design a secondary map of the data answering the needs of a specific research question, can then serve a dual purpose. It is the necessary step to derive information from the data as coherent as possible with the construals of interest, and a possible basis for a

clear and comprehensible documentation of the data themselves, using as a way of documenting data the traditional and well known form of a questionnaire, listing the "questions" posed not to actual respondents but to the administrative source data.

REFERENCES

Adriaans, J., Valet, P., and Liebig, S. (2020). Comparing administrative and survey data: Is information on education from administrative records of the German Institute for Employment Research consistent with survey self-reports? *Quality & Quantity*, *54*(1), 3–25. https://doi.org/10.1007/s11135-019-00931-4

Al-Sai, Z. A., Abdullah, R., and Husin, M. H. (2019). Big data impacts and challenges: A review. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 150–155. https://doi.org/10.1109/JEEIT.2019.8717484

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. https://www.wired.com/2008/06/pb-theory/

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., & Shapiro, M. D. (2014). *Using Social Media to Measure Labor Market Flows* (Working Paper 20010). National Bureau of Economic Research. https://doi.org/10.3386/w20010

Ardito, C., D'Errico, A., Leombruni, R., Ricceri, F., & Costa, G. (2020). Life expectancy inequalities and their evolution in Italy. How these impact on the equity of the pension system? *European Journal of Public Health*, *30*(Supplement_5), ckaa165.764. https://doi.org/10.1093/eurpub/ckaa165.764

Ardito, C., Leombruni, R., Mosca, M., Giraudo, M., & d'Errico, A. (2017). Scar on my heart: Effects of unemployment experiences on coronary heart disease. *International Journal of Manpower*, *38*(1), 62–92. https://doi.org/10.1108/IJM-02-2016-0044

Balazka, D., & Rodighiero, D. (2020). Big data and the little big bang: An epistemological (r)evolution. *Frontiers in Big Data*, *3*. https://www.frontiersin.org/articles/10.3389/fdata.2020.00031

Bena A, Giraudo M, Leombruni R., and Costa G. (2013). Job tenure and work injuries: A multivariate analysis of the relation with previous experience and differences by age. *BMC Public Health*, *13*, 1–9. https://doi.org/10.1186/1471-2458-13-869

Bena, A., Leombruni, R., Giraudo, M., & Costa, G. (2012). A new Italian surveillance system for occupational injuries: Characteristics and initial results. *American Journal of Industrial Medicine*, *55*(7), 584–592. https://doi.org/10.1002/ajim.22025

Benedetto, G., Haltiwanger, J., Lane, J., & McKinney, K. (2007). Using worker flows to measure firm dynamics. *Journal of Business & Economic Statistics*, *25*(3), 299–313. https://doi.org/10.1198/073500106000000620

Blount, M., Ebling, M. R., Eklund, J. Mikael., James, A. G., McGregor, C., Percival, N., Smith, K., & Sow, D. (2010). Real-time analysis for intensive care: development and deployment of the Artemis analytic system. *IEEE Engineering in Medicine and Biology Magazine*, *29*(2), 110–118. https://doi.org/10.1109/MEMB.2010.936454

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. The MIT Press. http://catalogue.bnf.fr/ark:/12148/cb44284618q

Cackett, D., Bond, A., & Gouk, J. (2013). *Information Management and Big Data: A Reference Architecture.* Oracle Corporation.

CEDEFOP. (2019). *Online Job Vacancies and Skills Analysis*. CEDEFOP. https://www.cedefop.europa.eu/en/publications/4172

Chetty, R. (2012). Time trends in the use of administrative data for empirical research. *34th Annual NBER Summer Institute. Cambridge, Mass.(July 9–27, 2012)*.

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*, 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015

Conti, R., Michele, B., Boscolo, S., Puccioni, C., Ricchi, O., Tedeschi, S., & Fabrizi, E. (2023). The Italian Treasury Dynamic Microsimulation Model (T-DYMM): Data, structure and baseline results. *Italian Department of the Treasury Working Paper Series WORKING PAPERS*, *1*.

Contini, B., & Revelli, R. (1986). Natalità e mortalità delle imprese italiane: Risultati preliminari e nuove prospettive di ricerca. *L'industria*, *2*, 195-.

Contini, B., & Revelli, R. (1987). The process of job creation and job destruction in the Italian economy. *Labour*, *1*(3), 121–144. https://doi.org/10.1111/j.1467-9914.1987.tb00122.x

Contini, B., & Trivellato, U. (2006). *Eppur si muove. Dinamiche e persistenze nel mercato del lavoro italiano*. Il Mulino.

Crato, N., & Paruolo, P. (A c. Di). (2019). *Data-Driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design*. Springer International Publishing. https://doi.org/10.1007/978-3-319-78461-8

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V. R., Tsou, A., Weingart, S., & Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*, *66*(8), 1523–1545. https://doi.org/10.1002/asi.23294

Engstrom, R., Hersh, J., & Newhouse, D. (2022). Poverty from space: Using high resolution satellite imagery for estimating economic well-being. *The World Bank Economic Review*, *36*(2), 382–412. https://doi.org/10.1093/wber/lhab015

Floridi, L. (2008). Data. In W. A. Darity (A c. Di), *International Encyclopedia of the Social Sciences*.

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, *165*, 234–246. https://doi.org/10.1016/j.ijpe.2014.12.031

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information Processing & Management*, *30*(1), 9–19. https://doi.org/10.1016/0306-4573(94)90020-5

Frické, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology*, *66*(4), 651–661. https://doi.org/10.1002/asi.23212

Gauvain, J.-L., Lamel, L., & Adda, G. (2000). Transcribing broadcast news for audio and video indexing. *Communication of the ACM*, *43*, 64–70. https://doi.org/10.1145/328236.328148

Gellert, R. (2022). Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms? *Regulation & Governance*, *16*(1), 156–176. https://doi.org/10.1111/rego.12349

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), Articolo 7232. https://doi.org/10.1038/nature07634

Giraudo M, Bena A, Leombruni R., & Costa G. (2016). Occupational injuries in times of labour market flexibility: The different stories of employment-secure and precarious workers. *BMC Public Health*, *16(1)*, 1–11. https://doi.org/10.1186/s12889-016-2834-2

Gitelman, L. (2013). *Raw Data Is an Oxymoron*. MIT Press.

Griliches, Z. (1984). *Data Problems in Econometrics* (SSRN Scholarly Paper 300716). https://papers.ssrn.com/abstract=300716

Griliches, Z. (1985). Data and econometricians—The uneasy alliance. *The American Economic Review*, *75*(2), 196–200.

Griliches, Z. (1986). Comment on Behrman and Taubman. *Journal of Labor Economics*, *4*(3), S146–S150.

Harford, T. (2014). Big data: A big mistake? *Significance*, *11*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2014.00778.x

Hethey-Maier, T., & Schmieder, J. F. (2013). *Does the Use of Worker Flows Improve the Analysis of Establishment Turnover? Evidence from German Administrative Data* (Working Paper 19730). National Bureau of Economic Research. https://doi.org/10.3386/w19730

Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., & Lyberg, L. E. (2020). *Big Data Meets Survey Science: A Collection of Innovative Methods*. John Wiley & Sons.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Hjørland, B. (2019). Data (with big data and database semantics). *KO Knowledge Organization*, *45*(8), 685–708.

Jensen, Howard E. 1950 "Editorial note." In H.P.Becker *Through Values to Social Interpretation: Essays on Social Contexts, Actions, Types, and Prospects*. Durham, NC: Duke University Press, vii–xi

Johnson, B., & Moore, K. (2005). Consider the source: Differences in estimates of income and wealth from survey and tax data. *Special Studies in Federal Tax Statistics*, 77–99.

Kapteyn, A., & Ypma, J. Y. (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics*, *25*(3), 513–551. https://doi.org/10.1086/513298

Khaouja, I., Kassou, I., & Ghogho, M. (2021). A survey on skill identification from online job ads. *IEEE Access*, *9*, 118134–118153. https://doi.org/10.1109/ACCESS.2021.3106120

Koopmans, T. C., Vining, R., & Hastay, M. (1995). 'Measurement without theory' debate (Review of Economics and Statistics, vol. 29, 1947, pp. 161–72 (cut); vol. 31, 1949, pp. 77–93 (cut); and Journal of the American Statistical Association, vol. 46, 1951, pp. 388–90). In D. F. Hendry & M. S. Morgan (A c. Di), *The Foundations of Econometric Analysis* (pp. 491–524). Cambridge University Press. https://doi.org/10.1017/CBO9781139170116.046

Kreshpaj, B., Orellana, C., Burström, B., Davis, L., Hemmingsson, T., Johansson, G., Kjellberg, K., Jonsson, J., Wegman, D. H., & Bodin, T. (2020). What is precarious employment? A systematic review of definitions and operationalizations from quantitative and qualitative studies. *Scandinavian Journal of Work, Environment & Health*, *46*(3), 235–247.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Leonelli, S. (2014). What difference does quantity make? On the epistemology of big data in biology. *Big Data & Society*, *1*(1), 2053951714534395. https://doi.org/10.1177/2053951714534395

Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science*, *82*(5), 810–821. https://doi.org/10.1086/684083

Leonelli, S. (2019). What distinguishes data from models? *European Journal for Philosophy of Science*, *9*(2), 22. https://doi.org/10.1007/s13194-018-0246-0

Llorente, R., Morant, M., Macho, A., Garcia-Rodriguez, D., & Corral, J. L. (2015). Demonstration of a spatially multiplexed multicore fibre-based next-generation radio-access cellular network. *2015 17th International Conference on Transparent Optical Networks (ICTON)*, 1–4. https://doi.org/10.1109/ICTON.2015.7193681

Lohr, S. L., & Brick, J. M. (2017). Roosevelt predicted to win: Revisiting the 1936 Literary Digest poll. *Statistics, Politics and Policy*, *8*(1), 65–84. https://doi.org/10.1515/spp-2016-0006

Lusinchi, D. (2012). "President" Landon and the 1936 Literary Digest poll: Were automobile and telephone owners to blame? *Social Science History*, *36*(1), 23–54. https://doi.org/10.1017/S014555320001035X

Man, X., Luo, T., & Lin, J. (2019). *Financial Sentiment Analysis (fsa): A survey*. 617–622.

McGregor, C., Catley, C., James, A., & Padbury, J. (2011). Next generation neonatal health informatics with Artemis. *Studies in Health Technology and Informatics*, *169*, 115–119.

Pacelli, L., & Revelli, R. (1995). Trasformazioni societarie, scorpori, fusioni: Un metodo di individuazione mediante dati di fonte Inps. Biffignandi S. Martini M., Il registro statistico delle imprese.

Petrelli, A., Sebastiani, G., Di Napoli, A., Macciotta, A., Di Filippo, P., Strippoli, E., Mirisola, C., & d'Errico, A. (2022). Education inequalities in cardiovascular and coronary heart disease in Italy and the role of behavioral and biological risk factors. *Nutrition, Metabolism and Cardiovascular Diseases*, *32*(4), 918–928. https://doi.org/10.1016/j.numecd.2021.10.022

Polgreen, P. M., Chen, Y., Pennock, D. M., & Nelson, F. D. (2008). Using internet searches for influenza surveillance. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *47*(11), 1443–1448. https://doi.org/10.1086/593098

Redman, T. C., Fox, C., & Levitin, A. (2017). *Data and Data Quality*. Routledge Handbooks Online. https://doi.org/10.1081/E-ELIS4-120008897

Risak, M., Rauws, W., Sredkova, K., & Portmann, W. (2013). *Regulating the Employment Relationship in Europe: A Guide to Recommendation No. 198*. ILO/ELLN.

Saberi, B., & Saad, S. (2017). Sentiment analysis or opinion mining: A review. *Int. J. Adv. Sci. Eng. Inf. Technol*, *7*(5), 1660–1666.

Shin, S.-Y., Kim, T., Seo, D.-W., Sohn, C. H., Kim, S.-H., Ryoo, S. M., Lee, Y.-S., Lee, J. H., Kim, W. Y., & Lim, K. S. (2016). Correlation between national influenza surveillance data and search queries from mobile devices and desktops in South Korea. *PLoS ONE*, *11*(7), e0158539. https://doi.org/10.1371/journal.pone.0158539

Smith, A. F. M. (1996). Mad cows and ecstasy: Chance and choice in an evidence-based society. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *159*(3), 367–383. https://doi.org/10.2307/2983324

Soren, S., & Peter, F. (2018). *What is the State of the Manufacturing Sector in Mozambique?* (Availability Note: Information provided in collaboration with the RePEc Project: http://repec.org).

Stüber, H., Grabka, M. M., & Schnitzlein, D. D. (2023). A tale of two data sets: Comparing German administrative and survey data using wage inequality as an example. *Journal for Labour Market Research*, *57*(1), 8. https://doi.org/10.1186/s12651-023-00336-9

Triplett, J. E. (2007). Zvi Griliches' contributions to economic measurement. In *Hard-to-Measure Goods and Services: Essays in Honor of Zvi Griliches* (pp. 573–589). University of Chicago Press. https://www.nber.org/books-and-chapters/hard-measure-goods-and-services-essays-honor-zvi-griliches/zvi-griliches-contributions-economic-measurement

Trivellato, U. (2019). Microdata for social sciences and policy evaluation as a public good. In N. Crato & P. Paruolo (A c. Di), *Data-Driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design* (pp. 27–45). Springer International Publishing. https://doi.org/10.1007/978-3-319-78461-8_3

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67.

US Bureau of Labor Statistics. (s.d.). *Employee Tenure Technical Note—2022 A01 Results*. Retrieved 8 october 2023, https://www.bls.gov/news.release/tenure.tn.htm

Wallgren, A., & Wallgren, B. (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*. John Wiley & Sons.

Yovits, M. C. (1969). Information science: Toward the development of a true scientific discipline. *American Documentation*, *20*(4), 369–376. https://doi.org/10.1002/asi.4630200421

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, *58*(4), 479–493. https://doi.org/10.1002/asi.20508

doi.org/10.26398/IJAS.302

HOW TO IMPROVE EMPLOYABILITY IN ENGINEERING STUDENTS: A STUDY ABOUT STUDENT CAREER PLANNING AND PERCEPTION OF LABOR MARKET

Elena Barzizza[1]**, Nicolò Biasetton, Riccardo Ceccato**
Department of Management Engineering, University of Padova, Vicenza, Italy

Monica Fedeli, Concetta Tino
Department of Philosophy, Sociology, Pedagogy and Applied Psychology, University of Padova, Padova, Italy

Abstract *Socio-economic factors such as unemployment among young adults have influenced graduates' perception of the labor market and their career planning. In this context, education plays a central role and in particular there is the need to understand which are the most important soft skills and competences that universities should promote to enhance employability among students. The aim of this paper, therefore, is to investigate engineering student perception of labor market and career planning. Data was collected by means of a questionnaire administered to engineering students of the University of Padova and analyzed by applying suitable machine learning models to investigate the relationship between student perception of the labor market and career planning, and some other factors, i.e. general information, self-perceived employability, career proactivity, and career control. Results of the analysis and their educational and social implications are presented and discussed.*

Keywords: *Employability, Engineering, Machine learning, Labor market, Career planning.*

1. Introduction

Understanding higher education students' perceptions of labor market demands has an effect on students' self-perceived employability, career proactivity, and career control. This is particularly of interest from an educational point of view since it can give useful insights about what students need to better develop their employability. In the literature there has been much discussion about graduate employability. It is referred to as "the capacity to gain initial employment, maintain employment and obtain employment if required" (Hillage and Pollard, 1998), or the ability to find and retain a graduate-level job or move between jobs if required (Yorke, 2010). In some studies, individual employability is considered to be a complex dimension made up of various factors, such as professional

[1]elena.barzizza@phd.unipd.it

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.
Competing Interests: The Author(s) declare(s) no conflict of interest.

identity, spanning social connectedness, work and life experience, and career self-management (Jackson, 2016; Tomlinson, 2017). Other research focuses on the ability of graduates to develop transferable skills (Cascio, 2019) to guarantee them a 'sustainable employability' (Monteiro and Ceu Taveira M. & Almeida, 2019). It is a construct related to a) 'career adaptability', i.e. "the tendency affecting the way an individual views his or her capacity to plan and adjust to changing career plans [...] especially in the face of unforeseen events" (Rottinghaus et al., 2005), and b) the agency of graduates and their adaptive and proactive resources necessary to move into employment (Montgomery and Cote, 2003). In employability literature, the effect of external factors such as labor market demands, economic trends and recruitment characteristics, has also been investigated (Hillage and Pollard, 1998; Rothwell and Arnold, 2007), but little research has been carried out on student perception of the labor market during the pandemic, and the effect on their career planning.

Youth unemployment has always been a key issue in policies of the European Commission (2009) and the Organization for Economic Co-operation and Development (OECD, 2012) that aimed to promote strategies and economic reforms that provide a highly qualified labor force with the research and development capabilities to contribute to innovation. Despite this, unemployment among young adults has continued to be a result of various factors: globalization, technological development, rapid transformation of work and professions; the distance and virtual conditions of jobs; boundaryless careers (Lo Presti and Pluviano, 2016) with consequent professional instability (Gevaert et al., 2018; Ingusci et al., 2016); the misalignment between skills that young people have at the end their learning path and those required by the labor market; the growth of the overeducation phe-nomenon with young people being underpaid, and their skills and competences undervalued (Bol et al., 2019; Duncan and Hoffman, 1981; Romero et al., 2017). The onset of the COVID-19 pandemic contributed to a worsening of this precarious situation and to increasing the dynamism of labor market, where the digital disruption and organizational and production changes prevail over any stability (Kaneklin and Gilardi, 2007) and, independently of their attained education, people must be ready to move from one role and set of activities to another, and abandon the traditional long-term employment idea. Before the pandemic, vari-ìous studies were carried out on student perceptions of the labor market, and on career planning/control behaviors, proactivity and self-perceived employability. Various results in the literature were registered in relation to students' labor market perceptions. Tomlinson (2008) highlighted how, given the complexity of the

labor market, students question the value of academic qualifications for finding jobs. Equally, Roulin and Bangerter (2013) reported that students recognize the importance of developing positional advantage to stand out and increase employment possibilities. This negative perception of the labor market was also identified in Jackson and Tomlinson (2020), in which students appear to be aware of the competitiveness of the world of work. On the other hand, some studies show that students consider the challenging labor market to be an opportunity rather than a threat (Deoitte, 2018). Malizia (2016) revealed that 60% of Italian graduates seem to have positive labor market perceptions in terms of the flexibility of their work and the correspondence between training and job obtained, but are less satisfied with the economic aspects because the job they obtain is often less paid than the quality of their professional profile merits. Students' self-perceived employability is linked to their individual beliefs regarding successful gaining of employment. Rothwell et al. (2008) investigated this by incorporating individual self-beliefs, student perceptions of university reputations and credibility of their chosen field of study into their measure, as well as the state of the labor market, demonstrating that self-perceived employability depends on both internal and external factors. Jackson and Wilton (2017) identified that positive self-perceived employability was linked to low levels of awareness of the uncertainty of the labor market. Career control, career planning, and proactivity are further evidence individual career behaviors. The ability to control one's career is an expression of self-regulation strategies and personal control (Coetzee and Stoltz, 2015; Savickas and Porfeli, 2012). Perceptions of career control can differ between graduates in employment and those who are not, but can also depend on the type of degree obtained (Deoitte, 2018). In fact, professional degree programs (e.g., Engineering, Health) compared to generalist degree programs (e.g., Art, Humanities) are linked to better employment outcomes (Karmel, 2015; Tino, 2021). Career planning is based on the formulation of goals and strategies to achieve the aspired-to career. Determinants of career control were identified as resources, agency, personal motivation and expectations (Lent and Brown, 2013), but also internal locus of control and self-efficacy (Fugate et al., 2004). Proactivity is associated with career initiative, attitude towards change and learning (Spitzmuller et al., 2015), towards finding opportunities and persevering for the achievement of goals (Bateman and Crant, 1993). It is related to entrepreneurial intent (Zampetakis and Moustakis, 2006, 2007) and individuals' adaptability to their environments (Crant, 2000). Despite the large number of interesting studies on different career concepts, only one was found to look at the relationship between student labor market percep-

tions, self-perceived employability, career planning/control, and proactivity (Jackson and Tomlinson, 2020). Although the authors provided empirical results on student perceptions of labor market registering a holistic impact on the most important career dimensions, they did not consider some personal and contextual factors as predictors of student career behavior, factors that are recognized in the literature as determinants of people's career orientation. These factors are the role of the university and high school learning experience, career modelling, family support, expectations, and personal career interest (Buday et al., 2012; Dasgupta and Stout, 2014; Ferry et al., 2000; Kim and Seo, 2014; Lent and Brown, 2013; Lent et al., 1994; Vargas et al., 2018). In fact, employers prefer prestigious universities in which students should be better prepared and ready for the world of work; students in the most highly rated universities have higher expectations in terms of employment in brand organizations (Rothwell et al., 2008). School or university learning experiences nurture students' interests and reinforce outcome expectations through a continuous internal (self-efficacy) and external (from others) recognition process. Career outcome expectations support people's behaviors and motivation when facing challenges. They are connected to people's beliefs about consequences of activity engagement and in terms of anticipation of some results (e.g., money, social recognition and approval, self-satisfaction) (Kim and Seo, 2014). Finally, career modelling and family support influence students' academic and career development, as well as career aspirations (Dasgupta and Stout, 2014). Based on the previous rationale, an Italian version of Jackson and Tomlinson (2020) questionnaire was developed and administered. It aimed to: investigate how labor market perceptions influence student career behaviors; explore student engage-ment in career planning in the context of a threatened world of work; underline students' perceived needs to develop their employability in terms of soft skills and capabilities.

Knowing how students perceive the current labor market and its effect on their career planning and career control will provide useful information for all involved stakeholders: (i) students will have information on their career development process and the achievement of their career outcomes by a leaded reflection on the alignment between their perceptions and the real opportunities and needs of labor market. It's a reflective process that helps them to become more aware of their career planning and control; (ii) universities can reflect upon their level of success, on the impact on graduate employment outcomes, and upon the possibility of introducing work-based teaching methods developed in partnership with industry to better prepare students for their future careers and employment (Dee-

gan and Martin, 2017; Frison, 2015; Tino, 2018); (iii) employers, who can learn about student career motivations and concerns which will assist when recruiting and engaging with graduates.

Thus, by exploring some career concepts, and using the scale developed by Jackson and Tomlinson (2020), this study sought to investigate higher education students' perceptions of the current labor market and the effect on their employability, pursuing two research objectives which essentially represents our data challenges:

(i) to explore student perceptions of current labor market demands and the factors that determine them;

(ii) to know what elements affect student career planning.

Collecting this data will support our understanding of the knock-on effects of labor market demands, student labor market perceptions, and student career planning and control. Furthermore, it will shed light on the opportunities for higher education institutions to design career development learning paths that match graduate skills and knowledge with labor market demands and align graduate employability competences with university-to-work transition processes. Of course, in order to achieve these objectives, it is essential to conduct appropriate statistical analyses with the aim of accomplishing our research goals. Our study was carried out at an Italian university using data collected from a survey administered to students at the faculty of Engineering.

The paper's structure is as follows: Section 2 presents the questionnaire, along with a description of the data. In Section 3, we explain the machine learning approach, as well as the results in Section 4. Final remarks are provided in Section 5.

2. Data and measure

2.1. Data description

The survey was administered to students of the faculty of Engineering. The software Limesurvey was used to collect data between March and May 2021. Students were sent a personalized link to their university email account and were selected from the final year (third-year) of undergraduate courses, from both years of Master's degree courses, and from the last three years of single-cycle courses (i.e. 5-year courses resulting in a Master's degree). These cohorts were selected by virtue of the belief that students at these stages of their university journey are

more aware of their career goals and the challenges of university-to-work transition than the other undergraduates. In comparison to other similar studies on employability in Engineering, such as those by Chou and Shen (2012), Idkhan et al. (2021) and Howell et al. (2023), which considered samples ranging from 130 to 530 respondents, our sample size is notably larger and may provide more robust insights.

2.2. Definition of the adopted measure

This study adopts both the English version of the scale (Jackson and Tomlinson, 2020) and the Italian version (Tino, 2021), developed through back translation (Brislin, 1970). The measure aims to register the relationships between student labor market perceptions and career planning, proactivity and employability.

In particular, the following will be considered:

- If there are greater negative perceptions of the labor market associated with
  - a lower level of self-perceived employability
  - a lower level of career control
  - a greater level of proactivity
  - greater commitment to developing positional advantages;
- If there is a positive association between
  - Self-perceived employability and career planning
  - Proactivity and greater career planning
  - A greater level of career control and career planning
  - A greater role of contextual factors and career planning
  - A greater role of personal factors and career planning.

The measurement is made up of 5 dimensions (perceived labor market conditions, self-perceived employability, career control and planning, proactivity, and developing positional advantage) and 28 items on a 5-point disagree-agree Likert scale. It also includes items on participants' characteristics such as gender, age, residency, study field and stage, level of employment, and parental occupation. The latter was useful to identify students' socio-economic status. With respect to Jackson and Tomlinson's study, additional personal and contextual variables were

considered because of their effect on individual career behaviors (Lent and Brown, 2013): family support, that considered the level of encouragement and recognition provided by families; the role of school focusing on the curriculum, in-school and out-of-school experiences, and student-teacher relationships; the role of the university, investigating the influencing role of the student-teacher relationship and learning experiences on career choices and self-awareness; career modelling, focusing on the career role and experiences of parents; expectations in relation to social recognition, career and job finding opportunities; personal career interests focusing on decision-making processes supported by a personal interest in a field of study, or specific careers and personal development. Testing of the psychometric properties of the scale was presented in a previous study (Tino, 2021) in which the inter-item consistency of scales was tested using Cronbach's alpha ($\alpha > 0.7$). Further operationalized contextual and personal factors considered for the group of engineers were: skills in the world of work, developed skills, skills to be promoted by the university and activities to be promoted by the university to support employability.

2.3. Descriptive statistics

Regarding the sample utilized for the analysis, the majority of participants were aged 18 to 25 (constituting 84% of the sample), although a notable portion was aged 26 to 30 (13%). Moreover, a significant proportion of the sample identified as male (74%), while a smaller segment consisted of international students (3%). The majority were enrolled as students, yet a quarter of the sample comprised student workers, indicating that some participants already had prior experience in the labor market. In the questionnaire, participants respondents were asked to express their opinions on two key performance indicators (KPIs): the current state of the labor market and career planning. Additionally, they provided insights on various other dimensions including proactivity, perceived employability, career control, and skills that they consider important. Moreover, their decisions regarding course of study/career paths, expectations, the role of school and university on curriculum and career choices, the development of a positional advantages and the activities that universities should promote were investigated. The subsequent paragraph presents the primary findings from the descriptive analysis of these themes. Initially, descriptive statistics and graphical representations of the KPIs are provided, followed by an examination of the remaining aspects.

The first KPI regards the perception of the current state of the labor market. Analysis of the responses showed that 64% of respondents believed there was

a fairly high risk of being employed in a job for which they are overqualified. About 40% of students were concerned about competition and uncertainty in the labor market. The 51% of sample agreed on the difficulty of finding work that students would like to do. See Figure 1 for more details. The second KPI under consideration pertains to career planning. Respondents agreed that they often thought about how to plan their future career and how to explore all potential career possibilities, they are also aware of the future career choices they have to make. They also agreed that they strive to improve their employability. See Figure 2 for more details. In both Figure 1 and Figure 2, responses were provided on a Likert scale ranging from 1 to 5. A score of 4-5 indicates "Agree," 3 denotes "Uncertain," and 1-2 signify "Disagree".

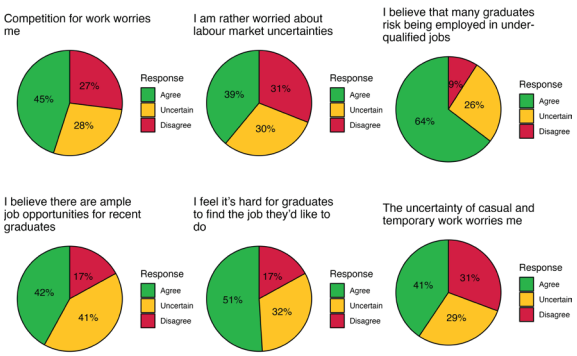


Figure 1: Responses on perception of the current state of the labor market

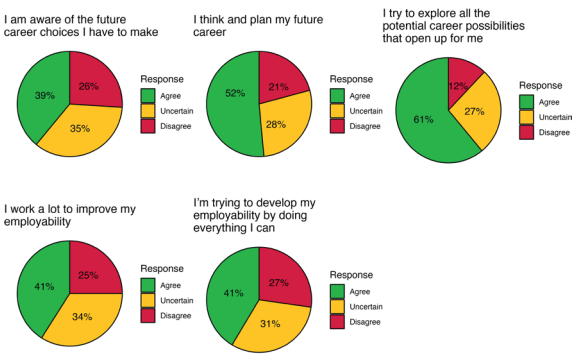


Figure 2: Responses on career planning

Let us now shift our focus to the other aspects addressed in the questionnaire. With regard to proactivity (see Figure 3), respondents considered having a career to be an important aspect of their life (81% of students agreed). Most of the students agreed that they think a lot about their future career (66%) and they are excited to start their career path (75%). With respect to the employability perception (see Figure 3), descriptive statistical analysis revealed that 72% of the sample believed they possessed the skills required by the labor market. They felt confident about the ability to compete in the world of work compared to other graduates (64% of students agreed) and they were convinced that they would be able to find work in their field of study (58%). With regard to career control (see Figure 3), freedom to choose one's own career path was an aspect on which the sample agreed (89%). Moreover, being responsible for one's career and in particular for one's successes or failures were also important to the sample (more than 70% agree with these aspects). A consistant number of students (39%) were uncertain on being able to handle their career setbacks.

Proactivity	
importance_having_career	Having a career is important to me
thinking_future_career	I think a lot about my future career
enthusiasm_start_new_career	I'm excited to start my career path
Employability perception	
skills_importance	I believe that my skills and experiences will be required by the world of work
belief_in_obtaining_job_vs_others	I think I'll be able to compete with other graduates to get a job
belief_in_obtaining_job_after_graduation	I'm sure I'll find work in my field after I graduate
Career control	
career_responsibility	I feel like I'm responsible for my career
career_responsibility_management	Each worker is responsible for managing his or her career
career_failure_management	I think I'll be able to handle setbacks in my career
freedom_in_career_choice	The freedom to choose my career path is important to me
responsibility_for_career_success_failure	I am responsible for the success or failure of my career

Figure 3: Variables considered in the questionnaire

Part of the questionnaire concerned the analysis of skills and activities that respondents consider important in the world of work (see Figure 4).

The skills considered most important were, in order of importance, analytical thinking and innovation, ability to solve complex problems, critical thinking, active learning, creativity, originality and initiative. Confirming this, 44% of the sample believed that the university should promote the development of analytical thinking, 40% agreed that the university should encourage development of the ability to solve complex problems, 42% agreed that the university should encourage the development of critical thinking and active learning, and 43% believed that the university should promote creativity, originality and initiative. More information on the skills that respondents think should be promoted at university can be found in Figure 5.

Skills in the world of work	
analytical_thinking_and_innovation	Analytical thinking and innovation
active_learning_and_strategies	Active learning and strategies
solving_complex_problems	Resolving complex problems
critical_thinking	Critical thinking
creativity_originality_initiative	Creativity, originality, initiative
leadership	Leadership
technologies_use	Use of technologies
technological_design_and_programming	Technological design and programming
Resilience_stress_tolerance_flexibility	Resilience, stress tolerance, and flexibility
reasoning_ideation	Reasoning and ideation
emotional_intelligence	Emotional intelligence
persuasion_negotiation	Persuasion and negotiation
Skills to be promoted by University	
university_analytical_thinking_and_innovation	Analytical thinking and innovation
university_active_learning_and_strategies	Active learning and strategies
university_solving_complex_problems	Resolving complex problems
university_critical_thinking	Critical thinking
university_creativity_originality_initiative	Creativity, originality, initiative
university_leadership	Leadership
university_technologies_use	Use of technologies
university_technological_design_and_programming	Technological design and programming
university_resilience_stress_tolerance_flexibility	Resilience, stress tolerance, and flexibility
university_reasoning_ideation	Reasoning and ideation
university_emotional_intelligence	Emotional intelligence
university_persuasion_negotiation	Persuasion and negotiation

Figure 4: Variables considered in the questionnaire

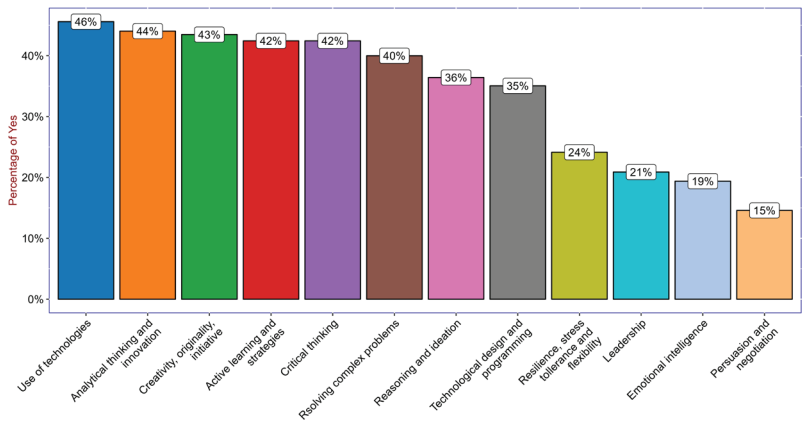


Figure 5: Skills that respondents believe should be promoted at the university

Another set of questions in the questionnaire concerned factors that determine the choice of course of study and career (see Figure 6). Regarding choice of course, the respondents were in agreement that the chosen course of study represents an opportunity for personal development (75%) and represents the desire of following a precise career path (52%), the choice is linked to a personal interest (80%), and that the knowledge and skills provided make it possible to find a good job (77%). Moreover they agreed that family always recognized their ability (74%) and encouraged their choice of study and career (72%). On the other hand,

they seemed to disagree more that the choice is linked to parental influence (66%) or extracurricular experiences (66%).

Among expectations linked to choice of course of study (see Figure 6), the samples agreed with the belief that the chosen course made it possible to find the job best suited to the respondent's preferences (70%). Around 60% agreed that the chosen course made it possible to occupy a position with an active role in society and a socially recognized career.

Choice of university path	
choice_based_on_interest	My choice of studies was determined by my personal interest in the field of study
choice_based_on_desire	The choice of my study path was driven by the desire to follow a specific career path
choice_as_personal_opportunity	The chosen course of study represents for me an opportunity for personal development
good_chances_finding_job	The knowledge and skills acquired during my studies offer me good chances of finding a job
family_encouragement_on_choice	My family encouraged my choices of studies and career
parental_career_influence_on_choice	My parents' career models influenced my choices
choices_based_on_experiences	Extracurricular experiences encouraged by my family helped determine my choices
family_recognition_of_abilities	My family has always recognized my abilities
Expectations	
expectations_for_socially_recognised_career	The chosen path will allow me to pursue a socially recognized career
expectations_for_active_role_society	The chosen path will allow me to occupy a position with an active role in society
expectations_for_pleasant_job	The chosen path will allow me to carry out the work I like the most

Figure 6: Variables considered in the questionnaire

Figure 7 contains questions related to the role of school and university in the choice of curriculum and career. The sample agreed that school activities promoted skills awareness (48%). Moreover the curriculum of the secondary school generated an interest in the choice of the university path (52% of students agreed). On the other hand, regarding the role of the university in relation to career choices, the sample agreed that theory/practice learning experiences strengthen career choices (46%) and generally strengthen the acquisition of awareness of personal skills (43%). They were not in agreement that the relationship with some professors at university strengthen career choices (44%) - only 30% agreed - as well as the relationship with professor at the secondary school strengthen university choice (56% disagreed). The same for the influence due to the peer groups (44% disagreed).

School role	
high_school_role	The secondary school curriculum generated my interest in the chosen path
awareness_abilities	School learning experiences helped me gain awareness of my skills
skill_building	The extra-curricular learning experiences helped boost my skills
University role	
theory_practice_learning_experience	The learning experiences at university that combine theory-practice help strengthen my career choices
learning_experience_improve_awareness	The learning experiences I am experiencing at university strengthen my ability to become aware of my abilities
peer_group_guided_choices	Relations with the peer group have shaped my choices
high_school_teachers_relationship	Relationships with some secondary school teachers influenced my choice
university_professors_relationship	Relationships with some faculty members at the university help strengthen my career choices

Figure 7: Variables considered in the questionnaire

Finally, with respect to the development of a positional advantages (see Figure 8), the sample disagreed that the participation in activities such as student, local and sport club, volunteering and career events, can help to develop a positional advantages (more than 50% of students disagreed in all the aspects considered in Figure 8). While the activities that universities should promote, in order of perceived importance, include: ensuring learning experiences that integrate theory and practice, fostering learning through work-based activities, and creating dialogue with the professional world to provide information about study plans (for the complete list, refer to Figure 8).

Positional advantage	
participation_student_club	Participation in clubs or student associations
participation_local_club	Participation in local club
participation_sport_club	Participation in sport club
volunteering	Participation in voluntary activities
participation_career_events	Participation in career events (seminars, job fairs, etc.)
Activities to be promoted by University	
promote_theory_practice_learning_experience	Ensure learning experiences that combine theory and practice
promote_work_based_activities	Foster learning through work-based activities
promote_dialogue_with_working_world	Create dialogue with the world of work giving information about study plans
provide_experiences_for_measuring_knowledge	Provide experiences for measuring their knowledge and skills
support_students_developing_crossfunctional_skills	Support students in developing their cross-functional skills
support_in_gaining_awareness	Support students in gaining awareness of their skills
promote_skill_active_job_seeking	Promoting skill development for active job seeking
support_development_professional_plan	Supporting students in developing a personal professional development plan
provide_career_coaching_services	Providing career coaching services
support_selfefficacy_development	Supporting students' self-efficacy development
provide_career_path_to_insipire	Providing students with exposure to role models and career paths to inspire them

Figure 8: Variables considered in the questionnaire

3. Machine Learning

3.1. KPIs and Drivers considered

Data collected through the described questionnaire should be properly anal-ysed to meet the objectives of this study, in particular there is the need to find a way to explore student perception of current labor market demands and the factors that determine them and also to understand what elements affect student career planning. Therefore, the focus of the statistical analysis is to understand which drivers have the greatest influence on the two considered groups of KPIs, one group looking at perception of the current state of the labor market (6 items) and the other looking at student career planning (5 items). Figure 9 details the referenced KPIs. We address these issues by leveraging a machine learning approach that involves a feature selection algorithm and some machine learning algorithms for classification.

Perception of the current state of the labour market	
graduate_opportunities	I believe there are ample job opportunities for recent graduates and graduates
concern_about_competition	Competition for work worries me
risk_underqualified_job	I believe that many graduates risk being employed in under-qualified jobs
difficulty_finding_job	I feel it's hard for graduates to find the job they'd like to do
job_uncertainty	The uncertainty of casual and temporary work worries me
job_market_uncertainty	I am rather worried about labour market uncertainties
Student career planning	
planning_future_career	I think and plan my future career
future_career_choices	I am aware of the future career choices I have to make
explore_career_possibilities	I try to explore all the potential career possibilities that open up for me
improve_employability	I work a lot to improve my employability
develop_employability	I'm trying to develop my employability by doing everything I can

Figure 9: KPIs considered in the model

Using the set of variables described in Section 2, we obtained a subset of variables by applying a feature selection algorithm, namely Boruta. This algorithm allowed us to understand which variables impacting upon the KPIs of interest are the most relevant. Certainly, in many practical classification scenarios, we often encounter a large number of features. However, it's not uncommon for a portion of these features to be irrelevant for the classification problem (Kursa et al., 2010) and their relevance may not be evident in advance (Kursa and Rudnicki, 2010a). That's why the use of a selection algorithm can be useful. In short, the Boruta algorithm is based on the random forest classification algorithm which is used to iteratively classify features as "important" or "unimportant" based on their significance, for a more detailed explanation of the steps see Kursa and Rudnicki (2010a). In particular, we used the `Boruta()` function from the R package Boruta (see Kursa and Rudnicki (2010b) for more details), setting the maximal number of importance source runs to 300 (i.e. `maxRuns = 300`) and using the Random Ferns importance function to obtain attribute importance (i.e. `getImp = getImpFerns`). The relevant identified drivers are listed and described in figure 10.

3.2. Machine learning model description

Our analysis considers two sets of response variables, or KPIs, and a set of input variables, i.e. the drivers (see Figure 9 and Figure 10). The aim of this analysis is to firstly understand which drivers have a relevant impact on the outcomes, and secondly understand which type of impact the drivers have (positive, negative or quadratic impact) with regard to the KPIs. We've treated all the KPIs as binary variables. Initially, they were collected as ordinal variables on a Likert scale of 1 to 5. However, we transformed them into binary variables using class **T2B** for responses equal to 4 and 5, while responses falling outside this range were categorized as **Others**. Essentially, the response variables are now considered binary

General information	
gender	Gender
age	Age group
study_stage	Year of course attended
highest_parental_occupation	Highest employment position occupied by one of your parents
student_working_status	Student working status
Employability perception	
skills_importance	I believe that my skills and experiences will be required by the world of work
belief_in_obtaining_job_vs_others	I think I'll be able to compete with other graduates to get a job
belief_in_obtaining_job_after_graduation	I'm sure I'll find work in my field after I graduate
Career control	
career_responsibility	I feel like I'm responsible for my career
career_responsibility_management	Each worker is responsible for managing his or her career
career_failure_management	I think I'll be able to handle setbacks in my career
freedom_in_career_choice	The freedom to choose my career path is important to me
responsibility_for_career_success_failure	I am responsible for the success or failure of my career
Positional advantage	
participation_student_club	Participation in clubs or student associations
participation_career_events	Participation in career events (seminars, job fairs, etc.)
Skills to be promoted by University	
university_analytical_thinking_and_innovation	Analytical thinking and innovation
university_active_learning_and_strategies	Active learning and strategies
university_solving_complex_problems	Solving complex problems
university_critical_thinking	Critical thinking
university_creativity_originality_initiative	Creativity, originality and initiative
university_technologies_use	Use of technologies
university_reasoning_ideation	Reasoning and ideation
university_technological_design_and_programming	Technological design and programming
Choice of university path	
choice_based_on_desire	The choice of my study path was driven by the desire to follow a specific career path
good_chances_finding_job	The knowledge and skills acquired during my studies offer me good chances of finding a job
parental_career_influence_on_choice	My parents' career models influenced my choices
choices_based_on_experiences	Extracurricular experiences encouraged by my family helped determine my choices
Expectations	
expectations_for_socially_recognised_career	The chosen path will allow me to pursue a socially recognized career
expectations_for_active_role_in_society	The chosen path will allow me to occupy a position with an active role in society
expectations_for_enjoyable_job	The chosen path will allow me to carry out work I enjoy
Activities to be promoted by University	
promote_theory_practice_learning_experience	Ensure learning experiences that combine theory and practice
promote_work_based_activities	Foster learning through work-based activities
promote_dialogue_with_working_world	Create dialogue with the world of work giving information about study plans

Figure 10: Drivers considered in the model

outcomes. As such we used a classification model to predict the responses of the KPIs split between the two created classes: T2B or Others.

We consider several machine learning models which may suit the purposes of the analysis. In particular we consider: support vector machines with radial ba-sis function kernel (`svmRadial`), random forest (`rf`), generalized linear model (`glm`) and `glmnet`. To optimize hyperparameters and identify the best-performing model, we leverage 5-fold cross-validation. We use cross-validated AUC as the performance metric, choosing the model with an AUC closest to 1. We repeat this model selection procedure for each KPI in our analysis (see Figure 11).

perceptions of the current state of the labour market				
	glm	**svmRadial**	**rf**	**glmnet**
graduate_opportunities	0,7300	0,7343	0,7369	0,7463
concern_about_competition	0,7042	0,6989	0,6990	0,7144
risk_underqualified_job	0,6155	0,5981	0,5938	0,6139
difficulty_finding_job	0,6779	0,6860	0,6935	0,6902
job_uncertainty	0,6625	0,6684	0,6630	0,6752
job_market_uncertainty	0,7231	0,7174	0,7133	0,7190
students' career planning				
	glm	**svmRadial**	**rf**	**glmnet**
planning_future_career	0,7287	0,7317	0,7282	0,7395
future_career_choices	0,7227	0,7413	0,7390	0,7414
explore_career_possibilities	0,7180	0,7187	0,7102	0,7278
improve_employability	0,7786	0,7879	0,7699	0,7824
develop_employability	0,7568	0,7584	0,7511	0,7602

Figure 11: AUC values for all the considered ML models

The least absolute shrinkage and selection operator (Lasso) classification (within the framework of *glmnet*) is selected since it shows the highest AUC value in most cases.

A common solution with a binary response is to use a logistic regression model:

$$p(\mathbf{x}) = P(Y = \text{T2B}|\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \mathbf{x}^T\boldsymbol{\beta})}} \tag{1}$$

where, $Y \in \{\text{T2B}, \text{Other}\}$ is the response variable and $\mathbf{x}$ is the matrix of predictors.

In the presence of a large number of potentially correlated predictors, we can use the Lasso penalization to improve the performances of the logistic regression model. Lasso regularization helps in preventing overfitting by penalizing the complexity of the model. Lasso indeed introduces a penalty term that can drive certain coefficients to exactly zero: a feature selection is therefore performed helping in identifying the most relevant predictors and removing irrelevant and redundant ones. Moreover, it can be effective in dealing with multicollinearity: through the process of penalizing certain coefficients to zero, Lasso tends to pick only one variable from a set of highly correlated predictors.

The logistic regression objective function with the Lasso penalty term is given by:

$$l(\beta_0, \boldsymbol{\beta}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] + \lambda \sum_{j=1}^{V} |\beta_j| \tag{2}$$

where the term $\lambda \sum_{j=1}^{V} | \beta_j |$ represents the penalization term defined as the sum

of the module of the model parameters. For the estimation of the parameters a weighted least squares iteration within a Newton step is then adopted (Friedman et al., 2010; Tibshirani et al., 2012).

4. Results and Discussion

In this section we show the results of application of machine learning models to understand the effect and impact of the drivers considered in Section 3.1 compared on the two KPIs under examination: perception of the current state of the labor market and student career planning.

The third to eighth columns of the figure represent a KPI and the values of the coefficient for each relevant driver in the model (see Equation 1).

4.1. Results: perceptions of the current state of the labor market

Figure 12 shows the results of the application of the machine learning model for the KPIs associated with perception of the current state of the labor market. Specifically, it displays the coefficients of application of the LASSO classification model.

	Driver	graduate opportunities	concern about competition	risk underqualified job	difficulty finding job	job uncertainty	job market uncertainty
General information	genderMale	0,26	-0,004			-0,42	-0,1
	age26-30				0,09		
	stage studyYear5				0,09		
Employability perception	skills importance	0,08					
	belief in obtaining job vs others	0,07	-0,37			-0,05	-0,04
	belief in obtaining job after graduation	0,41	-0,26	-0,2	-0,51	-0,37	-0,42
Career control	career responsibility	0,02		-0,06			
	career responsibility management	0,08			-0,09		
	career failure management				-0,02	-0,08	
	freedom in career choice				0,14		
	freedom in career choice^2			0,02	0,003		
	responsibility for career success failure	0,17					
Skills to be promoted by University	university reasoning ideationYes			-0,04			
Choice of university path	choice based on desire					0,05	
	good chances finding job	0,13		-0,07	-0,06		
Expectations	expectations for socially recognised career					0,13	
	expectations for enjoyable job			-0,04	-0,04		

Figure 12: Perceptions of the current state of the labor market: coefficients of the drivers in the model

4.2. Results: student career planning

Figure 13 shows the results of application of the machine learning model for the KPIs associated with student career planning. Again, it displays the coefficients of application of the LASSO classification model.

	Driver	planning future career	career future choices	explore career possibilities	improve employability	develop employability
General information	age26-30				0,07	0,03
	age31+				0,25	0,05
	stage studyYear5				0,12	
	student working status				-0,26	-0,05
Employability perception	skills importance		0,03		0,09	0,03
	belief in obtaining job vs others		0,11	0,04	0,15	0,07
	belief in obtaining job after graduation		0,05			
Career control	career responsibility		0,60	0,05	0,13	0,11
	career responsibility management				0,10	0,06
	career failure management		0,23	0,20	0,20	0,16
	freedom in career choice	0,14	0,09		0,23	0,12
	responsibility for career success failure		0,08			
Participation	participation career events	0,08	0,11	0,21	0,40	0,35
	participation student club				0,02	0,37
Choice of university path	choice based on desire	0,10	0,26	0,03	0,19	0,16
	choices based on experiences		0,04		0,02	
Expectations	expectations for active role society	0,04	0,05	0,07	0,11	0,12
	expectations for enjoyable job	0,08		0,10		
	expectations for enjoyable job^2		0,02			0,01

Figure 13: Student career planning: coefficients of the drivers in the model

4.3. Discussion and interpretation

4.3.1. Discussion and interpretation: perception of the current state of the labor market

Regarding perception of the current state of the labor market (Figure 12), the employability perception impacts especially on graduate opportunities, concern about competition and job/job market uncertainty. The career control variables have an impact especially on difficulty of finding a job and on the graduate opportunity while the choice of the university path have an impact on almost all the kpis except for the concern about the competition and the job market uncertainty. Finally expectations have an influence especially for the job uncertainty.

The findings in Figure 12 highlight two key aspects related to student labor market perceptions. On the one hand, some factors depend on individual skills and responsibility. On the other, some depend on external dimensions. Indeed, in relation to individual factors, students (above all males) have a high level of self-awareness because they consider skills development to be one of the conditions that guarantee them the opportunity to find a job. Furthermore, self-efficacy and internal locus of control allow them to consider themselves to be responsible for their own career success or failure. These factors are recognized by Fugate et al. (2004) as determinants of individuals' employability. Furthermore, belief in finding a job after graduation can be strongly linked to the field of study; indeed, engineering is recognized as a degree with better employment outcomes (Deloitte, 2018; Karmel, 2015), helping graduates obtain higher salaries and offering more favorable job prospects. The external dimensions with a generally negative impact are labor market competitiveness, the risk of being employed in a job for

which they are overqualified and job uncertainty, showing that students seem to be aware of competition in the world of work (Jackson and Tomlinson, 2020) and that students' self-perceived employability depends both on internal and external factors (Rothwell et al., 2008). Additionally, the negative effect observed with belief in finding job after graduation could call into question the value of academic qualifications for finding jobs (Tomlinson, 2008), and require students to improve their professional profile by participating in different activities to gain a positional advantage (Roulin and Bangerter, 2013).

4.3.2. Discussion and interpretation: student career planning

A brief glance at Figure 13 shows us the drivers with an important impact on the KPIs.

Specifically, employability perception has a moderate impact on future career choices as well as on the enhancement/development of employability. Career control drivers exhibit a high impact on future career choices and on improvement of employability while exhibits a moderate impact on both exploring career possibilities and improving employability. The participation to career events and students club impact in particular on improvement/development of the employabiliity while the choice of the university path has an impact especially on the career future choices and the improvement of employability. Finally, the expectations seem to have an influence on all the kpis considered and in particular on the improvement/development of the employability.

Other important aspects emerge from the student career planning findings (Figure 13). Students consider themselves to be the actors of their vocational path, capable of choosing and managing their own career, consistently with the previous results. This ability is an expression of students' self-regulation strategies, personal control and management of their vocational future (Coetzee and Stoltz, 2015). Each KPI analyzed within student career planning included four relevant factors with a positive impact: (i) expectations for enjoyable job - according to Hackman and Oldham (1980), if the characteristics of a job meet the jobholder's needs, they will be internally motivated and perform well. Although participants look for a job that makes them feel satisfied, the desire to play an active role in society could be connected to other factors; (ii) the need for social recognition and approval (Kim and Seo, 2014). Recognition plays an important role in constructing a career identity. Professional identity foresees a real negotiation between the educated and competent people and their environment through a process of socialization, made of constant exchanges with others, implementation of concrete

actions in different life contexts and situations. It is a developmental process that allows individuals to think about their professionalism, their position and social function (Damian, 2014); (iii) the expectation of playing an active role in society is connected to the desire to give back and contribute to society generating change in the world. It is a way to express human agency and self-efficacy that is related to the beliefs of people that events can be effectively managed through their choices and decisions; (iv) participation in career events or student clubs emerges as a dominant effect in all KPIs. Once more, students recognize the importance of developing positional advantage in order to stand out and increase employment possibilities. It mirrors student confidence and strong sense of control over their career.

The threats that participants seem to perceive in relation to labor market uncertainties and the risk of being employed in a job for which they are overqualified underline the responsibility of universities to care not only about students' employment readiness but also about their university-to-work transition paths. Indeed, students seem to have positive career planning prospects and skills but these may not be enough to navigate the complexity of the labor market. Universities must play an important role in supporting them in this transition process, teaching them how to evolve, and guaranteeing them authentic experiences and new teaching approaches that go beyond traditional and disciplinary methods (typical of each subject; they allow to maintain the control of knowledge in a field of study). In this sense, universities should become centers for the development of students' socio-professional identity (Grimaldi, 2016), where the aim is to guarantee authentic, real-life work experience-based learning strengthened by continuous business-university dialogue, to explicitly enhance students' adaptability and proactivity when facing labor market challenges. Positive career planning prospects do not guarantee a successful university-to-work transition. Facilitating students' proactive engagement with career development in real employment contexts helps them better define their own career profiles, identify resources, such as networking, exposes them to real problem solving, and introduces them to communities where they can test their knowledge, skills, and resilience.

5. Conclusion

This study analysed the engineering student labor market and career planning perceptions giving useful insight about how to enhance employability among engineering students. In particular, the survey investigated some important dimensions namely, perceived labor market conditions, self-perceived employability, ca-

reer control and planning, proactivity, and developing positional advantage. Such information are particularly important for universities for two main reasons: a) higher education institutions (HEIs) can engage students in a university-business dialogue to understand to what extent students' perceptions reflect the real labor market conditions and opportunities; b) the development of a university-business dialogue can support the analysis of labor market expectations and universities' contributions in the development of students' capabilities and skills considered crucial for students' employability. The challenge of our data analysis consist on extracting information about the factors that influence both student perceptions of the labor market and their career planning, especially to provide valuable insights to various stakeholders, namely students, universities and employers. With this aim data analysis was carried out through the application of machine learning model. Findings offer relevant implications for career development learning in HEIs and graduates' university-to-labor market transition paths. According to this new and complex employability scenario, the mission of universities changes, not only in terms of promoting the knowledge and skills needed to gain and maintain a specific job, but also in terms of developing those skills that allow students to move competently through uncertain contexts, and learn how to design or redesign their professional life. Indeed, findings showed that students consider it important to develop a positional advantage in order to better develop a skillful profile. The engineering students in this study seem to perceive themselves to be well-prepared with good potentiality for finding a job thanks to their knowledge and skills and because their field of study offers them many more job opportunities than others. One of the most relevant insights that these findings offer is the definition of a significant picture of the HEIs as systems responsible to act their role according to a general social perspective. Today, promoting knowledge do not mean only to providing students with skills and abilities, but creating learning environments where the construction of knowledge is strongly connected to the external world and the social issues. Why are females still facing challenges connected to the idea of finding positions within some fields (e.g. Engineering, as in this study)? The themes of gender issues, weak university-labor market partnerships, are useful to promote fruitful relationships, innovation in teaching methods and curriculum, and in workplace contexts as well; the impact of personal beliefs on career development and career choices, need to be included in the agenda of an entrepreneurial university with social responsibility. This contributes to provide HEIs with the opportunity to fulfill their third mission for economic and social development. Further studies could consider subsamples of students from differ-

ent backgrounds to understand if there are differences in perceptions of the labor market and career planning and how the educational offer should be adapted in terms of skills promoted and capabilities developed for different students' background. They could also investigate the social and entrepreneurial character of universities.

Acknowledgement

This study was carried out within the MICS (Made in Italy - Circular and Sustainable) extended partnership and received funding from the European Union Next-Generation EU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) - MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 - D.D. 1551.11-10-2022, PE00000004); the MOST Sustainable Mobility National Research Centre and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)-MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4âD.D. 1033 17 June 2022, CN00000023). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

Bateman, T. and Crant, J. (1993). The proactive component of organizational behavior: A measure and correlates. In *Journal Of Organizational Behavior*, 14: 1993.

Bol, T., Eller, C., and C., Van De Werfhorst H. & DiPrete, T. (2019). School-to-work linkages, educational mismatches, and labor market outcomes. In *American Sociological Review*, 84: 275–307.

Brislin, R. (1970). Back-translation for cross-cultural research. In *Journal of Cross-Cultural Psychology*, 1: 185–216.

Buday, S., Stake, J., and Peterson, Z. (2012). Gender and the choice of a science career: The impact of social support and possible selves. In *Sex Roles*, 66: 197–209.

Cascio, W. (2019). Training trends: Macro, micro, and policy issues. In *Human Resource Management Review*, 29: 284–297.

Chou, C. and Shen, C. (2012). Factors influencing employability self-efficacy of engineering students in Taiwan. In *International Journal of Engineering Practical Research*, 1: 10–14.

Coetzee, M. and Stoltz, E. (2015). Employees' satisfaction with retention factors: Exploring the role of career adaptability. In *Journal of Vocational Behavior*, 89: 83–91.

Crant, J. (2000). Proactive behavior in organizations. In *Journal of Management*, 26: 435–462.

Damian, J. (2014). Professional identity, social recognition and entering the workforce of the university student with hybrid education. In *Journal of Educational Psychology-Propositos Y Representaciones*, 2: 44–76.

Dasgupta, N. and Stout, J. (2014). Girls and women in science, technology, engineering, and mathematics: Steming the tide and broadening participation in STEN careers. In *Policy Insights from the Behavioral and Brain Sciences*, 1: 21–29.

Deegan, J. and Martin, N. (2017). *Merging Work & Learning to Develop the Human Skills that Matter*.

Deloitte (2018). Deloitte millennial survey: Millennials disappointed in business unprepared for industry. In *Deloitte Touche Tohmatsu*, 4.

Duncan, G. and Hoffman, S. (1981). The incidence and wage effects of overeducation. In *Economics Of Education Review*, 1: 75–86.

Ferry, T., Fouad, N., and Smith, P. (2000). The role of family context in a social cognitive model for career-related choice behavior: A math and science perspective. In *Journal of Vocational Behavior*, 57: 348–364.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. In *Journal of Statistical Software*, 33 (1): 1.

Frison, D. (2015). *Promuovere university-business dialogue: strategie ed esperienze didattiche di ricerca partenariale*. Pensa multimedia.

Fugate, M., Kinicki, A., and Ashforth, B. (2004). Employability: A psycho-social construct, its dimensions, and applications. In *Journal of Vocational Behavior*, 65: 14–38.

Gevaert, J., De Moortel, D., Wilkens, M., and Vanroelen, C. (2018). What's up with the self-employed? A cross-national perspective on the self-employed's work-related mental well-being. In *SSM-population Health*, 4: 317–326.

Grimaldi, A. (2016). Dall'autovalutazione dell'occupabilita' al progetto professionale. la pratica Isfol di orientamento specialistico. In *Isfol Research Paper*, 30: 1–96.

Hackman, J. and Oldham, G. (1980). Work redesign, Addison-Wesley Publishing Company Reading, Massachusetts, California. In *Development of The Job Diagnostic Survey. Journal Of Applied Psychology*, 60: 159–170.

Hillage, J. and Pollard, E. (1998). *Employability: Developing a Framework for Policy Analysis*. DfEE, London.

Howell, S., Hall, W., and Geelan, D. (2023). Exploring the perspectives of engineering undergraduates on employability and employability building activities. In *Higher Education, Skills and Work-Based Learning*, 13: 161–178.

Idkhan, A., Syam, H., and Hasim, A..O. (2021). The employability skills of engineering students: Assessment at the university. In *International Journal of Instruction*, 14: 119–134.

Ingusci, E., Manuti, A., and Callea, A. (2016). Employability as mediator in the relationship between the meaning of working and job search behaviours during unemployment. In *Electronic Journal of Applied Statistical Analysis*, 9: 1–16.

Jackson, D. (2016). Re-conceptualising graduate employability: The importance of pre-professional identity. In *Higher Education Research & Development*, 35: 925–939.

Jackson, D. and Tomlinson, M. (2020). Investigating the relationship between career planning, proactivity and employability perceptions among higher education students in uncertain labour market conditions. In *Higher Education*, 80: 435–455.

Jackson, D. and Wilton, N. (2017). Perceived employability among undergraduates and the importance of career self-management, work experience and individual characteristics. In *Higher Education Research & Development*, 36: 747–762.

Kaneklin, C. and Gilardi, S. (2007). Formare una pratica professionale competente in ambito psicologico: il ruolo dell'università. In *Psicologia Sociale*, 2: 389–408.

Karmel, T. (2015). Skills deepening or credentialism?: Education qualifications and occupational outcomes, 1996-2011. In *Australian Journal of Labour Economics*, 18: 29–51.

Kim, M. and Seo, Y. (2014). Social cognitive predictors of academic interests and goals in South Korean engineering students. In *Journal of Career Development*, 41: 526–546.

Kursa, M.B., Jankowski, A., and Rudnicki, W.R. (2010). Boruta–a system for feature selection. In *Fundamenta Informaticae*, 101 (4): 271–285.

Kursa, M.B. and Rudnicki, W.R. (2010a). Feature selection with the boruta package. In *Journal of statistical software*, 36: 1–13.

Kursa, M. and Rudnicki, W. (2010b). Feature selection with the boruta package. In *Journal of Statistical Software*, 36: 1–13.

Lent, R. and Brown, S. (2013). Social cognitive model of career self-management: toward a unifying view of adaptive career behavior across the life span. In *Journal of Counseling Psychology*, 60: 557.

Lent, R., Brown, S., and Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest. In *Choice, and Performance*, 45: 79–122.

Lo Presti, A. and Pluviano, S. (2016). Looking for a route in turbulent waters: Employability as a compass for career success. In *Organizational Psychology Review*, 6 (2): 192–211.

Malizia, G. (2016). Giovani e mercato del lavoro in uno scenario socio-economico ancora incerto. In *Problemi e Prospettive*, 53–71.

Monteiro, S. and Ceu Taveira M. & Almeida, L. (2019). *Career adaptability and university-to-work transition: Effects on graduates' employment status.* Education+ Training.

Montgomery, M. and Cote, J. (2003). College as a transition to adulthood. In *Blackwell Handbook of Adolescence*, 149–172.

Romero, L., Huertas, I., and Jimenez, M. (2017). Wage effects of cognitive skills and educational mismatch in europe. In *Journal of Policy Modeling*, 39: 909–927.

Rothwell, A. and Arnold, J. (2007). Self-perceived Employability: Development and Validation of a Scale. *Personnel Review*.

Rothwell, A., Herbert, I., and Rothwell, F. (2008). Self-perceived employability: Construction and initial validation of a scale for university students. In *Journal of Vocational Behavior*, 73: 1–12.

Rottinghaus, P., Day, S., and Borgen, F. (2005). The career futures inventory: A measure of career-related adaptability and optimism. In *Journal of Career Assessment*, 13: 3–24.

Roulin, N. and Bangerter, A. (2013). Students' use of extra-curricular activities for positional advantage in competitive job markets. In *Journal of Education and Work*, 26: 21–47.

Savickas, M. and Porfeli, E. (2012). Career adapt-abilities scale: Construction, reliability, and measurement equivalence across 13 countries. In *Journal of Vocational Behavior*, 80: 661–673.

Spitzmuller, M., Sin, H., Howe, M., and Fatimah, S. (2015). Investigating the uniqueness and usefulness of proactive personality in organizational research: A meta-analytic review. In *Human Performance*, 28: 351–379.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R.J. (2012). Strong rules for discarding predictors in lasso-type problems. In *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74 (2): 245–266.

Tino, C. (2018). *Alternanza scuola-lavoro. Le dimensioni-chiave per promuovere partnership strategiche: boundary spanners: un nuovo profilo professionale per le figure scolastiche dell'alternanza scuola-lavoro*. Pearson.

Tino, C. (2021). Relazione tra pianificazione e controllo della carriera, proattivita', self-perceived employability e percezioni-mercato del lavoro di studenti/esse: traduzione e affidabilita' di una scala di misurazione. In *Professionalita' Studi, Studium - Ed. La Scuola - ADAPT University Press*, 262–278.

Tomlinson, M. (2008). The degree is not enough: students' perceptions of the role of higher education credentials for graduate work and employability. In *British Journal of Sociology of Education*, 29: 49–61.

Tomlinson, M. (2017). Forms of graduate capital and their relationship to graduate employability. *Education+ Training.*

Vargas, R., Sanchez-Queija, M., Rothwell, A., and Parra, A. (2018). Self-perceived employability in Spain. *Education+ Training.*

Yorke, M. (2010). Employability: Aligning the message, the medium and academic values. In *Journal of Teaching and Learning for Graduate Employability*, 1: 2–12.

Zampetakis, L. and Moustakis, V. (2006). Linking creativity with entrepreneurial intentions: A structural approach. In *The International Entrepreneurship and Management Journal*, 2: 413–428.

Zampetakis, L. and Moustakis, V. (2007).Entrepreneurial behaviour in the Greek public sector. *International Journal of Entrepreneurial Behavior & Research.*

doi.org/10.26398/IJAS.285

INCLUSION OF PEOPLE WITH DISABILITIES IN THE LABOUR MARKET: DISENTANGLING INDIVIDUAL EMPLOYABILITY CHARACTERISTICS

Sara Maiorino, Federico Rappelli
PoliS-Lombardia, Milan, Italy.

Francesco Giubileo
CEQF – Competency Evaluation and Qualification Framework, Italy.

***Abstract**. In this study, we use data derived from different administrative sources to investigate the employment of people with disabilities through targeted placement services in the Lombardy region. Our analysis reveals that within the period considered (2018–2022), there was an increase in the proportion of employed people among those registered with these services. The share of individuals hired on permanent contracts also increased. Among the most relevant characteristics influencing an individual's chances of employment were citizenship, age, education and degree of disability. Individuals' chances of securing a permanent contract were also affected by gender.*

***Keywords**: Labour market; Disability; Targeted placement; Inclusion.*

1. INTRODUCTION

Regulation and promotion of the employment of people with disabilities in Italy is the object of Law no. 68/1999, which delegates to the regions the management of the labour market. The law's implementation, therefore, depends on the regions' ability to efficiently coordinate the various actors involved in the inclusion of people with disabilities in the labour market. The means established by the law to accomplish this inclusion are targeted support and placement services. People with disabilities seeking employment through targeted placement services must be registered on a list held by the competent offices, which manage

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.
Competing Interests: The Author(s) declare(s) no conflict of interest.

the ranking of the beneficiaries of Law 68/99 and supervise placements (Giovannone, 2022).

Despite the law's fairly advanced protection of the right to work of people with disabilities (van der Zwan and de Beer, 2021), difficulties persist: these include prejudice, inefficiencies within employment centres and factors such as the situation of the labour market, affected since 2008 by the long-lasting effects of a serious economic crisis and more recently by the economic consequences of the Covid-19 pandemic.

This research aims to shed light on the inclusion of people with disabilities in the labour market through the targeted placement system over the 5-year period 2018–2022 and to explore how different characteristics influence both the probability of being employed and the likelihood of obtaining a stable employment contract, namely a permanent contract. It contributes to the literature on the topic by outlining not only whether local measures to promote the employment of people with disabilities are effective but also which individual characteristics may hinder the effectiveness of such interventions. Furthermore, the study employs a novel dataset, derived from public administration sources which are not publicly available, reflecting a process subject to strict privacy guidelines. The results raise implications for local policies, such as which subgroups among the large and heterogeneous group of people with disabilities should be addressed by more targeted interventions.

2. BACKGROUND

Parts of the literature on labour market inclusion of people with disabilities focus on the institutional factors affecting entrance into the labour market and particularly on the effect of anti-discrimination laws aimed at reducing barriers to job market inclusion (Jones, 2008), while other contributions focus on variations in the rate of employment between different welfare regimes. In this regard, Tschanz and Staub (2017) examine how disability policy, including labour market integration policies, varies among different welfare regimes in Europe, identifying four distinct models: the encompassing model, prevalent in Nordic countries; the activating and rehabilitating model seen in Central European countries; the preference for social protection largely held in Southern European

countries; and a distinct model followed in Eastern European countries, characterised by few guaranteed civil rights.

When considering the degree to which people with disabilities are included in the labour market, two main indicators have been used to compare countries:

- some contributions examine the employment rate of people with disabilities per se: Jones (2008), reviewing multiple literature sources on the impact of disability on labour market outcomes, and specifically on employment and earnings, found a penalty associated with disability, regardless of the time, the data used and the place in which those data were collected; Heggebø and Dahl (2015), examining unemployment among people with ill health in a pre-crisis (2007) and in the crisis (2011) period in the 28 EU countries, found that this was fairly stable over time and that such people experienced unemployment to a lesser extent than those with good health status in the crisis year, if only in countries with high overall employment rates;
- others use the disability employment gap (DEG), that is, the difference in employment rates between people with and without disabilities (Geiger et al., 2017; van der Zwan and de Beer, 2021). Geiger et al. (2017) performed a cross-sectional comparison of 25 countries, comparing their DEG scores as shown by different surveys. In some cases, they found concomitant results across different surveys on the same country, while in others they found lower DEGs where higher-quality surveys had been conducted, highlighting the need for measures that can more effectively be compared over time and space. Van der Zwan and de Beer (2021) examined how the DEG varied by country and gender and how labour market policies influenced the size of the gap. They found that Southern European countries had comparatively smaller disability gaps and that stricter legislation, such as Italy's mandatory quota system, tended to be beneficial for people with disabilities.

There are 20 EU countries implementing the quota system, including Italy, and those with the lowest DEGs are among them (Richard and Hennekam, 2021).

When we consider the national panorama, and more specifically the targeted placement system, the literature on the topic is quite sparse. A recent contribution (Moscatelli et al., 2024), focusing on targeted placement from the perspective of

companies in Lombardy, uses qualitative analysis to understand how the process could be improved from their viewpoint. The findings outline practical suggestions to foster improvement at the regional level, including boosting the networks of local stakeholders operating in the field of labour market inclusion of people with disabilities and improving the homogeneity of the procedures followed in different regions.

What most of the contributions mentioned here, as well as others focused on the national level, have in common is that they concentrate on the pure correlation between disability and labour market participation, without considering the features of specific jobs and how the individual characteristics of disabled people influence their entrance to and permanence in the job market.

Regarding the national level, in the scientific literature, there is not much evidence concerning labour market entrance realised through targeted placement mechanisms. The sensitivity of the data concerning people with disabilities is among the main reasons for this.

Law 68/99 regulates the so-called reserve quotas. Within the framework of this law, the initial provisions that warrant reflection here pertain to the subjective and objective scope of application.

To delve into the specifics, the obligation to hire applies to both public and private employers who are required to employ workers from disadvantaged groups, as specified by the so-called reserve quotas.

One positive aspect to highlight is that Law 68/99 has significantly expanded the range of businesses subject to hiring obligations, now including those with workforces ranging from 15 to 35 employees, hitherto excluded by the previous regulatory framework.

The criteria for calculating the reserve quota are outlined in Article 4 of Law 68/99, which has undergone significant regulatory changes with controversial practical implications. On the one hand, the range of individuals to be considered for the correct calculation has been expanded. On the other hand, certain categories of workers have been excluded, limiting access to employment for disabled individuals who are ‘external’ to the company.

Through the use of a dedicated fund, Lombardy’s regional administration, in applying the law, finances services ranging from providing technical assistance to companies that need to comply with reserve quotas to tutoring people with disabilities who are seeking employment.

In this context, the aim of the present study is twofold:

(I) Understanding to what extent people with disabilities who choose to register on the targeted placement lists in Lombardy are included in the labour market and whether other specific measures designed to foster their employment are effective.

(II) Delving into the nature and characteristics of their employment (e.g. in terms of duration of contracts and precariousness of positions) and understanding how individual characteristics may impact the efficacy of such measures.

3. DATA AND METHODOLOGY

The research was carried out exploiting data from different data sources: the initial dataset was built by merging data from provincial targeted placement offices (which are in charge of collecting and retaining the records of the beneficiaries of Law 68/99) and data from mandatory communications (*Comunicazioni obbligatorie*), which are made on the creation, modification or termination of an employment contract.

The two sets of data can be described more specifically as follows:

- The data collected from the targeted placement offices contained the records of a sample of people who registered at the offices in 2017. They were collected from five provinces: Milan, Bergamo, Monza and Brianza, Cremona, and Mantua. Despite not being perfectly representative of the whole region, the chosen provinces were among those with the highest number of people registered. The total number of observations collected from the five provinces was 8,099;
- The data collected from the mandatory communications data referred to the same individuals registered in 2017, following their careers over five years, from 2018 to 2022, registering their occupational status and type of contract (if employed) at the end of each year.

The two datasets were merged using the personal identification numbers of the individuals. The privacy concerns raised by this process were addressed by using a password-protected cloud storage system to exchange and merge data.

After matching, each personal ID number was subjected to a pseudonymisation process, allowing for complete anonymisation of the records.

The resulting final dataset, containing observations of 8,099 individuals over five years, was harmonised. Later, two different analyses were performed, utilising multilevel logistic regression models.

We employed a multilevel logistic regression model to take into account the structure of the data, nested within five different provinces, each with its own contextual features, so that we were able to account for the effects of unobserved group characteristics. The model equation took the form.

$$\text{logit}(\Pr(Y_{ij} = 1)) = \beta_0 + \beta_1 x_{1ij} + \ldots + \beta_m x_{mij} + u_j$$

where Y_{ij} is the binary response variable for the i^{th} individual within the j^{th} cluster, x_{1ij} through x_{mij} are the k predictors measured on the i^{th} subject in the j^{th} cluster.

The target variable was different between the two models used:

- First, we tried to understand which individual characteristics had a positive or negative effect on the probability of being employed, with the dependent variable taking the value 1 if the individual was employed at the end of the year considered (observed on 31 December) and 0 if not. The multilevel model considered the structure of the data, nested at the provincial level.
- In a second multilevel logit model, we checked whether the same characteristics influenced the probability of obtaining a stable employment position, namely a permanent contract, with the dependent variable assuming the value 1 if the individual had a permanent contract at the end of a given year and 0 if their contract was not permanent, namely in all other cases (apprenticeship, on-call and all other types of fixed-term contracts). The type of contract recorded was that observed on 31 December of the year under consideration.

The initial variables included in the dataset were citizenship (binary, distinguishing between Italian or foreign), gender, educational level (no education title (ISCED 0), primary education (ISCED 1), lower secondary education (ISCED 2), upper secondary education (ISCED 3), post-secondary non-tertiary education (ISCED 4), or short-cycle tertiary education or higher (ISCED 5) and

year (ranging from 2018 to 2022). Other variables included were degree (percentage) of disability and marital status. As predictors in the model, we included citizenship, gender, age, age-squared (to check whether the age effect was linear or not), degree of disability and educational level (no title or primary; secondary; tertiary or higher). We also controlled for time to check whether a pre- and post-Covid-19 effect was visible in our data.

3.1 SAMPLE CHARACTERISTICS AND DESCRIPTIVE STATISTICS

The following table shows the starting characteristics of the sample collected from the targeted placement offices among people registered in 2017.

Table 1: Targeted placement sample's characteristics

Average percentage of disability (%)	66.2	
Average age	42.3	
	N	**%**
Gender		
Men	4,673	57.7
Women	3,426	42.3
Education level		
ISCED 0	26	0.3
ISCED 1	493	6.2
ISCED 2	2,910	36.8
ISCED 3	1,556	19.7
ISCED 4	1,958	24.8
ISCED 5	955	12.1
Citizenship		
Italian	7,450	91.9
Foreigner	649	8.0
Civil status		
Single	11,985	51.2
Married	7,810	33.3
Separated or divorced	2,625	11.2
Widowed	330	1.4

As we can see from Figure 1, the percentage of individuals on the lists who were employed on 31 December was above 40% in each of the years considered. The lowest share of employed individuals was registered in 2018. It subsequently increased (2019) and decreased again in the year of the Covid-19 outbreak (2020). The following year, 2021, recorded a slightly higher percentage, which remained stable in 2022.

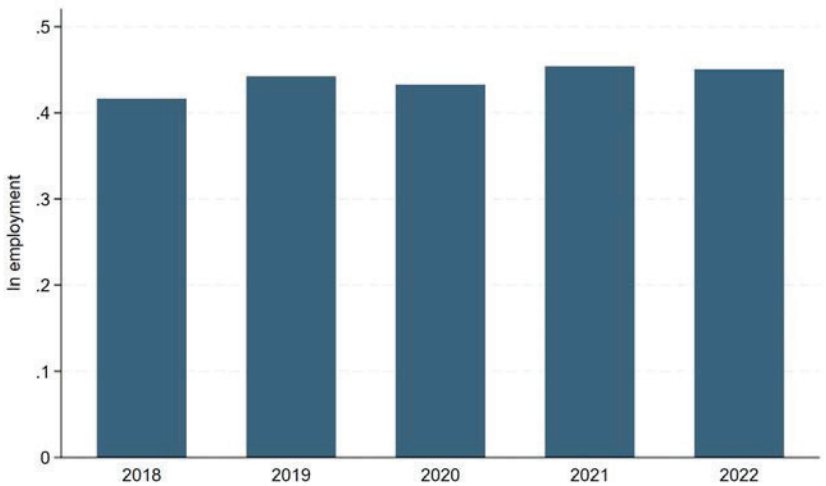


Figure 1: Percentage of employed individuals on the whole sample (Years 2018-2022)

If we look at the percentage of permanent contracts in our sample, this increased progressively each year, from 19.6% in 2018 to 23.6% in 2019 and 24.6% in 2020, reaching 25.5% in 2022.

Two other features of our data that must be highlighted are the following: 39.2% of the sample appear in our data as "never employed", namely they were never employed when the snapshot the 31st of December of each of the observed years was taken. This suggests that these individuals might never have been employed over the five years considered or that they might have been so only for short and non-contiguous periods. The second is that a considerable percentage of the individuals contained in our sample seem to have been subject to a fairly high job turnover: over the five years considered, 16.2% had held at least 2

different positions, 5.1% at least 3 positions, 1% at least 4 and 0.2% at least 5. This phenomenon appears to have been slightly more frequent among men than among women.

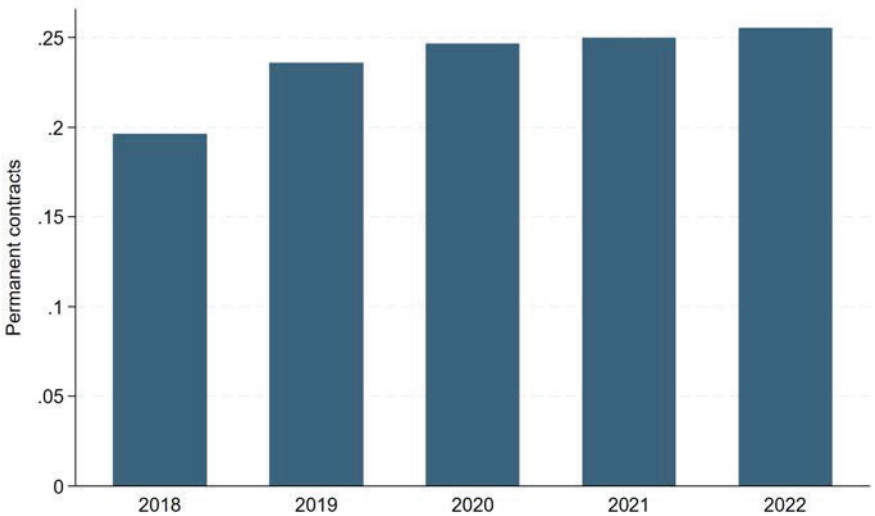


Figure 2: Percentage of employed individuals with a permanent contract on the whole sample. Years 2018-2022.

4. RESULTS

The analysis revealed as the most important in affecting the probability of entrance into the labour market were almost the same as those impacting the probability of obtaining a permanent contract.

Being male had a statistically significant impact on an individual's probability of securing a permanent contract. While the coefficient of gender was not significant for the probability of obtaining a job, women were less likely to obtain permanent contracts: according to our data, men registered on the targeted placement lists were 10.2% more likely to secure a permanent contract than women. Previous studies conducted in Lombardy region showed that, among the

general population, being a male seems to be positively correlated with both the probability of being in employment and the chances of having a permanent contract (PoliS-Lombardia, 2022). It is therefore interesting to note that, for people registered in the targeted placement lists, there is not a substantial difference in the probability of getting a job between males and females.

Being a foreigner resulted in a 42% lower probability of obtaining a permanent contract and a 22% lower probability of access to employment. The magnitude of the coefficient is quite high and may shed light on a double disadvantage effect of disabled immigrants in the labour market: among the general population, in fact, migrants are often employed in low skilled occupations, facing the risk of overqualification more than the risk of unemployment (Riva and Zanfrini, 2013).

We controlled for education: having a tertiary education had a significant positive effect on both the probability of securing a job and that of obtaining a permanent contract (respectively, increasing the chances of getting a job by 67.3% and those of obtaining a permanent contract by 123.6%); secondary education was in both models non-significant compared to the reference category (primary education). If we look at the descriptive statistics provided on the general population by Istat (2022), we can observe a relevant gap in the employment rate of people with primary and people with secondary education, suggesting that a secondary school education title does play a role in the odds of getting a job in this case. It is instead interesting to observe, among people registered in the targeted placement lists, the huge impact of a tertiary education title on the chances of entering the job market and/or being hired with a permanent contract.

Age had a linear and negative relationship with the probability of both events: a 1-year increase in age decreased the likelihood of employment by about 4% and the likelihood of a permanent contract by about 2.1%. The degree of disability had a negative and significant impact on the probability both of obtaining a job and a permanent contract, but the magnitude of this impact was not as great as we might have expected. A 1% increase in the degree of disability decreased the likelihood of getting a job by 1.8% and the likelihood of securing a long-term contract by 1.1%. The relatively small magnitude of the coefficient could perhaps be attributable to heterogeneity in terms of disability types in the sample (our data did not include this information), which may have impacted individuals' chances of securing employment differently. Finally, controlling for time, we observe positive and significant variations over time both in the

probability of entering the job market (even if the significance decreases in 2020, the year of Covid's outbreak) and in the likelihood of obtaining a permanent contract.

There was considerable provincial variation in the chances of getting a permanent contract but not in the likelihood of getting a job.

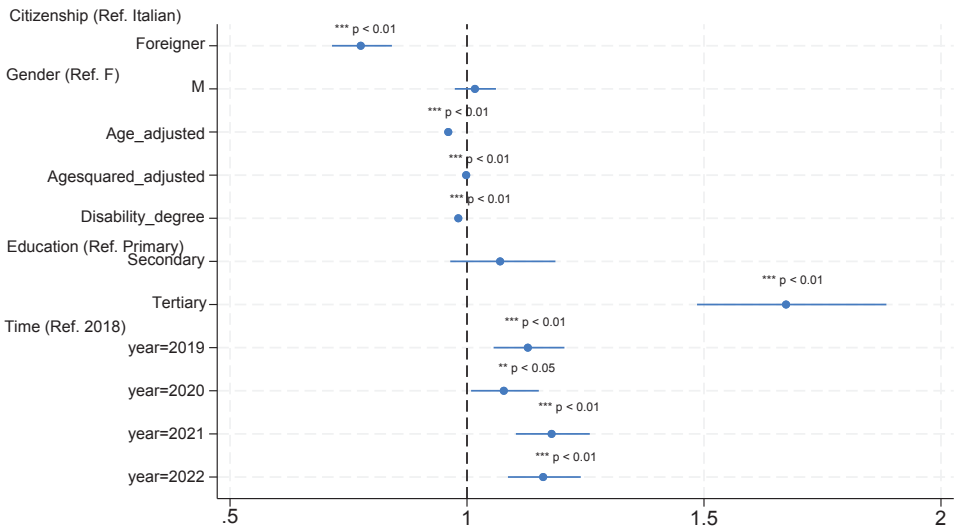


Figure 3: Odds ratio Model 1. Dependent variable: Employed (Yes/No).

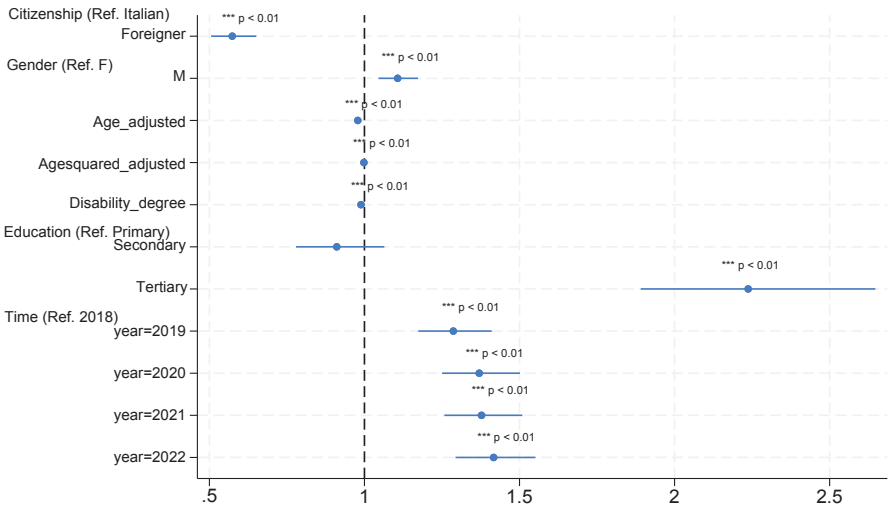


Figure 4: Odds ratio Model 2. Dependent variable: Permanent contract (Yes/No).

5. CONCLUSIONS

In this study, we used data derived from different administrative sources to investigate the employment of people with disabilities through targeted placement services in Italy (more specifically in Lombardy). Our analysis showed that there was an improvement over the years considered both in the probability of finding employment for the individuals registered on the targeted placement lists and in their chances of being hired with stable (permanent) contracts. Among the characteristics with the strongest positive influence on individuals' chances of securing employment were being Italian, having completed tertiary education and having a lower degree of disability. The same applied to individuals' chances of being hired with a permanent contract, except that in this case, gender also played a role, with being male associated with a higher probability of obtaining such a contract. Our results also suggest that geographical location in terms of province might have an impact concerning the type of contracts available to disabled individuals. Eventually, our findings show that, for people in the targeted placement lists, having a tertiary education has a huge impact, with respect to

primary education, both in the odds of being hired and in that of obtaining a permanent contract. Therefore, more attention should be paid on the labour market insertion of people with disability with lower education levels.

These results are useful to inform more targeted policies and measures aimed at improving the extent to which local placement services are effective in achieving the labour market inclusion of people with disabilities. If cases are identified in which a double disadvantage may be at play (such as, for instance, being a foreigner and disabled), more specific interventions may be discussed and designed to address them. Furthermore, this evidence may be useful to broaden the discussion from the solely binary outcome of being "in" or "out" of the job market to the quality and stability of placements available for the heterogeneous subgroups of the people registered on the targeted placement lists.

Thanks to the collection of data from multiple sources, this paper is aligned with one of the commitments made by countries in the 2030 Agenda for Sustainable Development: to obtain better statistics and to allow the monitoring of progress in the labour market inclusion of disabled people (OECD-ILO, 2018).

6. LIMITATIONS AND SHORTCOMINGS

Our research has some limitations. Firstly, the sample of people selected among those registered with the targeted placement services may not be perfectly representative of the whole Lombardy region. Secondly, our data suffer from the same limitations as most administrative data, including a high number of missing values that prevented us from using certain variables and incomplete or inaccurate data items.

Furthermore, some records were improperly entered, rendering data harmonisation impossible in some cases. As a result, we may lack relevant information on characteristics that could be relevant mediators in our analysis, such as the previously mentioned type of disability.

Finally, despite having information over multiple years for single individuals, we only have the observed value on 31 December of each of the years considered, leaving us unaware of any changes occurring within the year.

REFERENCES

Geiger, B. B., Van Der Wel, K. A. and Tøge, A. G. (2017). Success and failure in narrowing the disability employment gap: Comparing levels and trends across Europe 2002–2014. *MC Public Health*. 17(1), 1-7.

Giovannone, M. *L'inclusione lavorativa delle persone con disabilità in Italia.* ILO. https://www.ilo.org/wcmsp5/groups/public/---europe/---ro-geneva/---ilo-rome/documents/publication/wcms_874035.pdf. Last access: 06/03/2024.

Heggebø, K. and Dahl, E. (2015). Unemployment and health selection in diverging economic conditions: Compositional changes? Evidence from 28 European countries*. International Journal for Equity in Health*. 14(1), 1-17.

Istat. (2022). *Labour Force Survey Statistics*. http://dati.istat.it/. Last access: 05/02/2024.

Jones, M. K. (2008). Disability and the labor market: A review of the empirical evidence. *Journal of Economic Studies. 35*(5), 405-424.

Moscatelli, M., Pavesi, N. and Ferrari, C. (2024). Targeted placement for people with disabilities in Italy: A perspective from Lombardian companies. *Equality, Diversity and Inclusion: An International Journal*. *43*(9), 1-17.

OECD-ILO. *Labour Market Inclusion of People with Disabilities.* https://www.ilo.org/wcmsp5/groups/public/---dgreports/---inst/documents/publication/wcms_646041.pdf. Last access: 20/09/2023.

PoliS-Lombardia*. Indagine sull'occupazione femminile e maschile delle imprese in Lombardia con più di 50 dipendenti.* https://www.polis.lombardia.it/wps/wcm/connect/13033778-40bd-4840-9ced-9e39c43e09b8/221351SOC_occupazione_femm_masch_2022_ed2023giugno.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-13033778-40bd-4840-9ced-9e39c43e09b8-oGSRTlT. Last access: 06/03/2024.

Richard, S. and Hennekam, S. (2021). Constructing a positive identity as a disabled worker through social comparison: The role of stigma and disability characteristics. *Journal of Vocational Behavior*, *125*, 103528.

Riva, E. and Zanfrini, L. (2013). The labor market condition of immigrants in Italy: The case of Lombardy. *Revue Interventions Economiques. Papers in Political Economy*. (47).

Tschanz, C. and Staub, I. (2017). Disability-policy models in European welfare regimes: Comparing the distribution of social protection, labour-market integration and civil rights. *Disability & Society*. 32.8: 1199-1215.

van der Zwan, R. and de Beer, P. (2021). The disability employment gap in European countries: What is the role of labour market policy?. *Journal of European Social Policy*. 31(4), 473-486.