Information Economics and Policy 47 (2019) 27–37

Contents lists available at ScienceDirect

Information Economics and Policy

journal homepage: www.elsevier.com/locate/iep

AI meets labor market: Exploring the link between automation and skills

Emilio Colombo[a,*], Fabio Mercorio[b], Mario Mezzanzanica[b]

[a] Università Cattolica del Sacro Cuore and CRISP, Italy
[b] University Milano-Bicocca and CRISP, Italy

ARTICLE INFO

Article history:
Available online 29 May 2019

JEL classification:
J24
J63
C81

Keywords:
Machinelearning
Web vacancies
Skill analysis
Automation

ABSTRACT

This paper develops a set of innovative tools for labor market intelligence by applying machine learning techniques to web vacancies on the Italian labor market. Our approach allows to calculate, for each occupation, the different types of skills required by the market alongside a set of relevant variables such as region, sector, education and level of experience. We construct a taxonomy for skills and map it into the recently developed ESCO classification system. We subsequently develop measures of the relevance of soft and hard skills and we analyze their detailed composition. We apply the dataset constructed to the debate on computerization of work. We show that soft and digital skills are related to the probability of automation of a given occupation and we shed some light on the complementarity/substitutability of hard and soft skills.

1. Introduction

It is indisputable that in the past few decades significant forces and factors have dramatically changed the nature and characteristics of the labor market in both advanced and developing countries. Technical progress, globalization and the re-organization of the production process with outsourcing and offshoring have radically altered the demand for certain skills.[1] In addition, population aging in advanced economies intensifies the need for continued training, and is likely to affect the structural demand for certain competences, in particular those related to the health and care of the elderly.[2] The overall impact of these factors on the labor market is multifaceted. On the one hand several jobs are disappearing while new jobs are emerging; of these some are simply a variant of existing jobs, others are genuinely new jobs that were nonexistent until few years ago. On the other hand the quantity and quality of the demand for skills and qualifications associated to the new labor market has changed substantially. New skills are needed not only to perform new jobs but also the skill requirements of existing jobs have changed considerably. Which occupations will grow in the future and where? What skills will be demanded the most in the next years? Those are the questions that are at the forefront of the policy debate both among economists and policymakers. In order to address these questions specific data need to be collected. This calls for new tools that integrate and complement existing labor market instruments for grasping the complexity and the variability of new labor market trends.

In this paper we develop a set of innovative tools for labor market intelligence by applying machine learning techniques to web vacancies on the Italian labor market. Those tools are at the forefront of research in computer science and are able to address a number of technical and methodological challenges dealing with large volumes of unstructured data, mainly in textual form. In particular they are specifically designed for analyzing firms' skill needs. In this way we can calculate the skills required by the market, alongside a set of relevant variables such as region, sector, education and level of experience.

Our approach allows to shed light on a number of issues. First we can calculate, for each occupation, the different types of skills required and their frequency. Furthermore we are able to classify those skills into a standard classification system and assess the relevance of digital skills and of soft-hard skills constructing measures of specific "skill degree". We then apply the dataset constructed to the debate on computerization of work. We show that soft and digital skills are related to the probability of automation

* Corresponding author.
E-mail addresses: emilio.colombo@unicatt.it (E. Colombo), fabio.mercorio@unimib.it (F. Mercorio), mario.mezzanzanica@unimib.it (M. Mezzanzanica).

[1] See Bhagwati and Panagariya (2004); Feenstra (1998); Acemoglu (1998, 2002); Acemoglu and Restrepo (2017); Acemoglu and Restrepo (2018); Bessen (2018); Autor et al. (1998); Autor et al. (2003), Card and DiNardo (2002).

[2] see Freeman (2006) De Grip and Van Loo (2002).

https://doi.org/10.1016/j.infoecopol.2019.05.003
0167-6245/

of a given occupation and we shed some light on the complementarity/substitutability of hard and soft skills.

The remainder of the paper is structured as follows. Section 2 analyses the advantages and limits of using web vacancies with respect to other more traditional methods. Section 3 describes the tools and the methodology used, Section 4 presents the results. Finally Section 5 concludes.

2. Skills for the labor market of the future: new tools needed?

As stressed in the introduction, mega-trends such as globalization, technical progress and population aging are having a profound impact on the labor market. Albeit the overall effect is extremely complex we can roughly identify two main dimensions of it. One that we can call the extensive margin pertains to the creation of new jobs and the destruction of existing ones. This is probably the most debated issue at the center of the policy debate. Until few years ago the major culprit was identified in globalization with outsourcing and offshoring of jobs, recently however the focus of the attention has been centered on technology. New technologies allow the automation of an increasing number of tasks traditionally performed by individuals. While initially this effect was mainly concentrated in routine based activities both manual (assembly, logistics etc.) and clerical (administration, reporting etc.), with the advent of big data, artificial intelligence and of the Internet Of Things the possibility that activities considered to be too complex to be performed by a machine or by a software, could now be automated has become more concrete.

The second dimension operates along the intensive margin. In addition to create and destroy jobs, technology is changing profoundly existing jobs, in particular the tasks and their skill requirements modifying considerably the skill-mix employers require and placing greater emphasis on soft skills such as problem solving, ability to work in team, communication abilities etc.

In terms of overall effect the intensive margin is likely to be more important than the extensive one as it potentially affects the entire stock of the labor force. However to analyze it, is necessary to develop tools that are able to measure the characteristics of the jobs, their skill requirements and how these change over time and across occupations.

So far firms' skill needs have been assessed through skill surveys[3] which have the merit of providing a comprehensive picture of skill use and needs, but suffer from some relevant major drawbacks. First they are expensive, particularly when considering the opportunity cost of the burden of time for respondents. Second they are cumbersome, as they often take several weeks to be implemented. As a consequence surveys are low frequency tools generally implemented annually. Probably the most important limit of surveys is the approach they follow. They are top-down tools which need to be designed first and subsequently implemented. For this reason the type and the quality of information collected necessarily follows from the initial design. Regarding skills, there are specific questions about them, and the list of skills is generally pre-defined.

Interestingly the same forces that are changing the labor market such as AI and machine learning, are giving us new tools for analyzing the workforce and to solve some of the problems outlined above. In this paper we apply AI techniques to online vacancies and develop a new tool for assessing firms' occupation and skill needs which has interesting prospects. It follows a bottom-up approach that is entirely data-driven. The initial data collected contains all the information that individual firms post on the web. This large amount of data is subsequently filtered and processed using appropriate techniques to obtain the required information. In this way the tools help to categorize a pre-existing information set, but they do not pre-classify the information itself. The type of skills to be classified are those that emerge from the data, not those pre-defined in a questionnaire. This is particularly useful for the identification of soft skills and certain occupation-specific skills that surveys often ignore. For example the O*NET survey, by far the most comprehensive, is able to classify 35 different skills, the UK skill survey 23 skills and the Italian survey only 12. Using a bottom up approach in our data we are able to identify more than 1000 specific skills that can be subsequently grouped in different macro categories. The data driven approach is extremely flexible and allows us to raise new questions and consequently to expand continuously the spectrum of knowledge of the observed phenomena. This feature is particularly important for the detection of *emerging skills*, as it is possible to go back on previous data in order to re-assess them.[4]

Regarding the cost, web based tools tend to have a high fixed cost related to the initial development and implementation of the relevant software and algorithms, however there is the scope for large economies of scale as the same tools can be applied to an increasing number of vacancies. In addition, subject to sufficient computation power, the implementation lag is minimal and there is no need to involve workers or entrepreneurs. These tools are automatically implemented by machines that can operate at any time or on any date, allowing information to be collected almost in real time.

There is however one limitation of online vacancy analysis: they may not represent the entire population. It is well known that some occupations and sectors are not present in web advertisements or that in some regions or areas digital tools are not widespread enough to encourage employers to post vacancies online. Although this problem is likely to decrease over time as the use of web for job advertising becomes widespread,[5] and although there are statistical tools that can address the representativeness problem,[6] nevertheless this is a major issue that has to be addressed when using online vacancies. For this reason rather than substituting existing labor market tools, online vacancy analysis should complement them in providing in depth and real time information for policy analysis.

3. Data and methodology

3.1. Data

The source of the data is Wollybi,[7] a project that collects online vacancies in Italy from job-portals since February 2013. For internal data consistency we concentrate on 2016 and 2017 containing approximately 2 millions vacancies from which the tools described in the next section allow to extract the relevant information such as location, sector, education, skills etc. The sources are the major portals that advertise vacancies and include newspaper websites, job boards and employment agencies. In order to maximize the quality of the data we concentrated on primary sources, ne-

[3] There are several examples of skill surveys. Some such as the OECD survey of adult skills (PIAAC), the O*NET questionnaire or Cedefop's European Skill Survey are worker based, therefore capture the worker assessment of the use of certain skills in the workplace. Others such as the UK Employer skill survey, or the the Excelsior survey in Italy, are employer survey and capture directly firms' skill needs.

[4] This contrasts with skill surveys, where the information set is determined and not modifiable, and therefore can only be used to answer pre-defined questions identified during the survey design phase.

[5] Indeed there is ample empirical evidence showing that online job search is crowding out alternative search channels (Kroft and Pope, 2014).

[6] See Colombo et al. (2018) for a recent application.

[7] See www.wollybi.com.

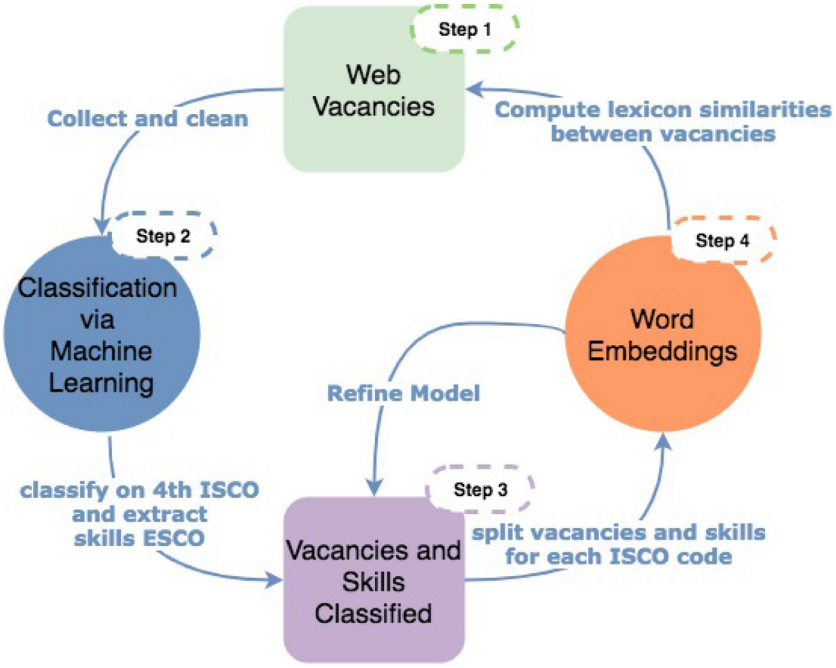

Fig. 1. Overview of the framework proposed.

glecting secondary sources such as aggregators (e.g. websites that re-post vacancies retrieved from other websites).[8]

3.2. Methods

In the construction of the dataset a number of challenges had to be overcome. The data has to be scraped from the websites validated and cleaned. Subsequently the relevant information has to be extracted from unstructured text and classified.

From a methodological point of view, we follow the KDD approach (Knowledge Discovery in Databases, see (Fayyad et al., 1996)), a framework that is now a benchmark in extracting useful and reliable knowledge from raw data in real-life scenarios. It requires to apply a number of steps that are represented in Fig. 1.

3.2.1. Source selection and cleaning

First, data sources need to be selected and assessed. Each website is characterized by its own data structure therefore each web source has been evaluated and ranked on the basis of identified criteria (typology, size, presence of the most important variables, quality of the information, update frequency, etc.). Once the sources have been identified and selected, the data are scraped and stored accordingly. Subsequently the data have to be transformed and cleaned. Roughly speaking, transformation implies modifying the data structure and the content from the original structure to a common framework. The presence of several sources implies several data models characterized by different structures, level of detail etc. It is therefore imperative to construct a common framework to which different data models can be harmonized. During this process, the quality of the data is assessed and cleansing activities are executed. In our context, this task deals mainly with the identification of duplicated job vacancies posted on different web source as well as job vacancies published multiple times on the same site; these tasks have been performed applying AI algorithms (see, e.g. Hernández and Stolfo (1998); Mezzanzanica et al. (2015); Boselli et al. (2014); Lovaglio and Mezzanzanica (2013).

3.2.2. Classification via machine learning

The subsequent steps pertain the proper information extraction from the content of the vacancies. The data is mainly textual and is generally composed by a mixture of structured and unstructured fields. Structured fields refer to a specific category as job title title (e.g. occupation), location, etc. Unstructured fields are the more general description of the content of the vacancy. This is often the most precious part of the data because it contains information about skill requirements, education, experience etc. The major problem is that both structured and unstructured fields contain information in form of unstructured text, that is text not organized into a standardized classification system or taxonomy. For example the field “title” of the vacancy contains the job title which is usually either described using natural language or using websites’ own taxonomy; these descriptions have to be mapped into a standard taxonomy such as SOC, ISCO etc. This problem is addressed by building a classifier: a function that maps (i.e. classifies) a data item into one of several predefined classes. In case of the job title the items are web job vacancies whilst the classes are the ones from the ISCO 4th level hierarchy.

Given the size of the data this task has been implemented through machine learning. Several algorithms have been tested, and Support Vector Machines resulted the best in terms of classification accuracy–higher than 93% (Boselli et al., 2017; 2018a; 2018b).[9]

More specifically text categorization aims at assigning a Boolean value to each pair $(d_j, c_i) \in D \times C$ where D is a set of documents and C a set of predefined categories. A *true* value assigned to (d_j, c_i) indicates document d_j to be set under the category c_i, while a false value indicates d_j cannot be assigned under c_i. In our LMI scenario, we consider a set of job vacancies $\mathcal{J}$ as a collection of documents each of which has to be assigned to one (and only one) ISCO occupation code. We can model this problem as a text classification problem, relying on the definition of Sebastiani (2002). Formally speaking, let $\mathcal{J} = \{J_1, \ldots, J_n\}$ be a set of job vacancies,

[8] The sources are all private as the website of the Italian PES at present contains too few vacancies and is rarely updated.

[9] The algorithms have been trained on a training set of 1k web job vacancies classified by us and validated by labor market experts.

the classification of $\mathcal{J}$ under the ISCO classification system consists of |O| independent problems of classifying each job vacancy $J \in \mathcal{J}$ under a given ISCO occupation code o_i for $i = 1, \ldots, |O|$. Then, a *classifier* is a function $\psi : \mathcal{J} \times O \to \{0, 1\}$ that approximates an unknown target function $\dot{\psi} : \mathcal{J} \times O \to \{0, 1\}$. Clearly, as we deal with a single-label classifier, $\forall j \in \mathcal{J}$ the following constraint must hold: $\sum_{o \in O} \psi(j, o) = 1$.

Classification has been implemented for occupations (ISCO), location (NUTS) education (ISCED), industry (NACE), and skills (see below).

3.2.3. Skill extraction.

The next step involves the extraction of skills required in each vacancy through the analysis of its text. This is achieved through linguistic models that allow to identify the portion of the text relevant for the analysis, extract information about skills by disambiguating different texts (i.e. the word "design" can refer to a skill or to an action which is not skill related), computing word similarities and dismissing spurious information (e.g. "ideal candidate").

This goal is achieved by incrementally building a taxonomy of extracted words recognized as potential skill. Specifically, the system uses the n-gram[10] Document Frequency (DF), i.e., the number of vacancies where the n-gram is found. The result is a list of n-grams that identify skills together with their synonymous, that is potential skills. We then use string similarity functions[11] to map potential skills to ESCO ontology. The mapping has been validated by experts. Such Information Extraction techniques have proved to be helpful for this purpose in several applications related to labor market (see, e.g., Lee (2011); Singh et al. (2010); Yi et al. (2007)).

3.2.4. The role of ESCO.

In classifying occupation titles and skills we have used the taxonomy contained in ESCO. ESCO is a multilingual classification system for European skills, competences, qualifications and occupations developed by the European Commission. The ESCO occupation classification corresponds to the International Standard Classification of Occupations (ISCO-08) up to the 4th digit level. ISCO has a hierarchical structure similar to the SOC system used in the US. The skills pillar of ESCO is not hierarchical but is organized as a graph tanking into account both the occupation specificity of hard skills and the transversal nature of soft skills. This paper uses version 1.0 of ESCO released in the summer of 2017.

Although ESCO provides a classification of occupations that corresponds to the ISCO classification in the next sections of the paper we will use data organized along the SOC classification. Therefore we mapped the ISCO codes of our classification into the corresponding SOC codes. This procedure is not without problems as it is well known that there is not a one to one correspondence between SOC and ISCO (see (Hoffmann, 2003)). In case of multiple correspondences we have attributed the same SOC code to multiple ISCO.

For each occupation (at 4 digit level) we have analyzed the skills required and calculated the soft skill degree, the hard skill degree and, in the latter group, the ICT degree. The *skill degree* is the frequency of occurrence of skills of a certain category in a given occupation.[12] For example if an occupation is characterized by a hard skill degree of 32% it means that 32% of the skills found in all vacancies that refer to that specific occupation belong to the hard category.

In this way we can construct a measure of skill relevance or skill importance which is entirely data driven.[13]

4. Results

4.1. Hard, soft skills and occupations

The methodology described in Section 3 allows to extract skills from vacancies. The use of ESCO taxonomy allows to group skills into categories that allow a better representation and analysis. The first major distinction is between hard and soft skills. Hard skills are typically job-specific skills and competences that are needed to perform a specific job or task (examples are knowledge of specific software or instruments, specific manual abilities etc.) Soft skills, on the other hand, are more transversal in nature and refer to the capacity of individuals to interact with others and the environment (examples are communication skills, problem solving etc.). Within hard skills we further distinguish between *digital* skills and *non-digital* skills. Digital skills encompass a range of different abilities that allow an individual to use ICT tools at different levels, from the use, manipulation and interaction with standard ICT tools down to the design, implementation and deployment of complex ICT systems and services.

Given the relevance of digital skills for today's labor market we augmented the ESCO classification by further constructing the following sub-groups.

- ***Information brokerage skills.*** Refer to the ability to use ICT tools and platforms for data exchange and communication (e.g. social media);
- ***Basic ICT skills.*** Refer to the ability to use some standard ICT applications for supporting the individual professional activities (e.g. use of spreadsheet or word processing software);
- ***Applied/Management ICT skills.*** These skills refer to tools and software used within the organization for supporting management, operational and decision making processes (e.g. administrative software);
- ***ICT technical skills.*** Refer to solutions, platforms and programming languages that are strongly related to ICT-specific professions (e.g. programming languages, advanced ICT softwares).

Within soft skills we identify the the following sub-categories.

- ***Thinking skills.*** Refer to the ability to apply mental processes and reasoning to solve complex problems, to increase knowledge and to perform complex tasks.
- ***Social interaction*** Refer to the ability to develop interact and engage with colleagues, clients and customers.
- ***Application of knowledge*** General application of skills commonly used in the workplace and in learning; knowledge of the organization and the working environment.
- ***Attitudes and values*** Refers to a person's work style, preferences and work-related beliefs that underpin behavior so that knowledge and skills are applied effectively. Examples are ability to adapt to change, to work independently, to meet commitments etc.[14]

Fig. 2 provides a quick overview of the distribution of skills over occupation aggregated by large occupation group (for simplicity we have used SOC 1 digit). For each occupation we have calculated the number of times each skill is mentioned by each category and subcategory. In the top left panel we distinguish between hard and

[10] An n-gram is a contiguous sequence of n items from a given sequence of text or speech.

[11] Levenshtein distance, Jaccard similarity, and the Srensen-Dice index have been employed in this phase.

[12] Whenever a single SOC occupation corresponded with multiple ISCO occupations we calculated the soft, hard and digital degrees by averaging across ISCO occupations.

[13] This is a major difference with respect to alternative measures such as those provided by the O*NET system which are based on expert judgment.

[14] These skills do not refer to a person's character.

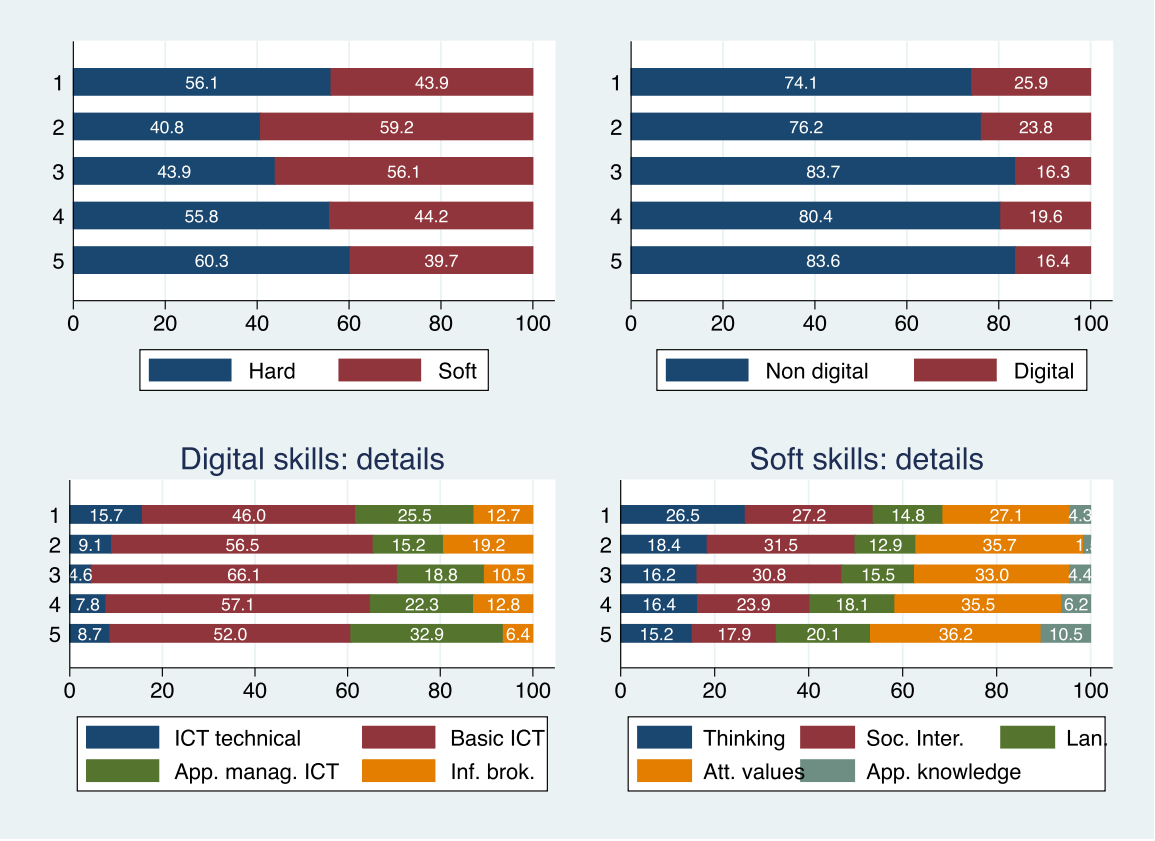

Fig. 2. Distribution of skills typology by SOC 1 digit group.

soft skills. Not surprisingly hard skills tend to be more relevant for technical and production related occupations (groups 1, 4 and 5) while soft skills are relatively more demanded in service related occupations (groups 2 and 3). Within hard skills digital ones account approximately for 20% (top right panel). The bottom panels present the detailed distribution of the components of digital and soft skills. In the former basic ICT skills account for approximately 50% of the total with ICT technical being more relevant in technical occupations. In the latter attitudes and values and social interaction are the more relevant category (approx. 30% each) while thinking skills are more relevant in group 1 and 2 (managerial and high skill occupations).

As stressed previously one of the major advantages of our dataset is it richness even at a high level of granularity. Fig. 3 shows box plots of the distribution of the soft skill degree by 2 digit SOC group.[15] The figure reveals some interesting patterns. First, despite the great deal of heterogeneity of the within group distribution, soft skills are pervasive and for several occupations are more relevant than hard ones. Even in highly technical occupations (e.g. group 15: Computer and Mathematical Occupations) soft skills are highly demanded. Second, as expected, soft skills tend to be less important in low skill occupations (groups 5) than in high skill ones (groups 1–2). Third, for some occupations the distribution of the soft skill degree tend to be more concentrated than in others. This could be the result of a narrower set of competences and tasks required or also by the fact that occupations mostly affected by technological progress and globalization (Goos et al., 2014) require larger demand for social skills to cope with this change. The next section will investigate more thoroughly these aspects. Fig. 4 shows box plots of the distribution of the digital skill degree by occupation groups. Overall digital skills are less pervasive than soft skills although a moderate degree of digital skill is required by almost every occupation even the less technical ones.

The granularity of the data at territorial level allows also to analyze regional differences in skill demand within occupations. This is an interesting piece of information because a systematic difference in skill requirement within occupation would signal the presence of a segmented labor market. Italy is known to be characterized by a strong regional divide between the more advanced North and the less developed South. In order to control for all possible confounding factors we have use the full data set and regressed the demand for different category of skills of each vacancy on a dummy for the North and a set of controls including experience, education dummy for sectors and occupation.[16] The value of the coefficient of the regional dummy represent the regional difference in demand for a certain skill within occupation controlling for differences in sectors, education and experience.

Fig. 5 shows that there is a greater demand for hard skills (mainly non digital ones) and of Basic and applied management ICT skills in the North. No statistically significant difference emerges for soft skills nor for experience. The only feature in higher demand in the South is education. This could be explained by a dual system not only in the labor market but also in the educational system which in the North is able to deliver the hard skills requested by firms mainly through the secondary and post-secondary VET system. On the contrary in the South the lack of specific hard skills provided by formal education induces firms to ask for a higher level of education. Overeducation in the South is therefore a response to the skill mismatch.

[15] Here each observation is the skill degree for a given occupation at 4 digit level.

[16] Errors are clustered at occupation level

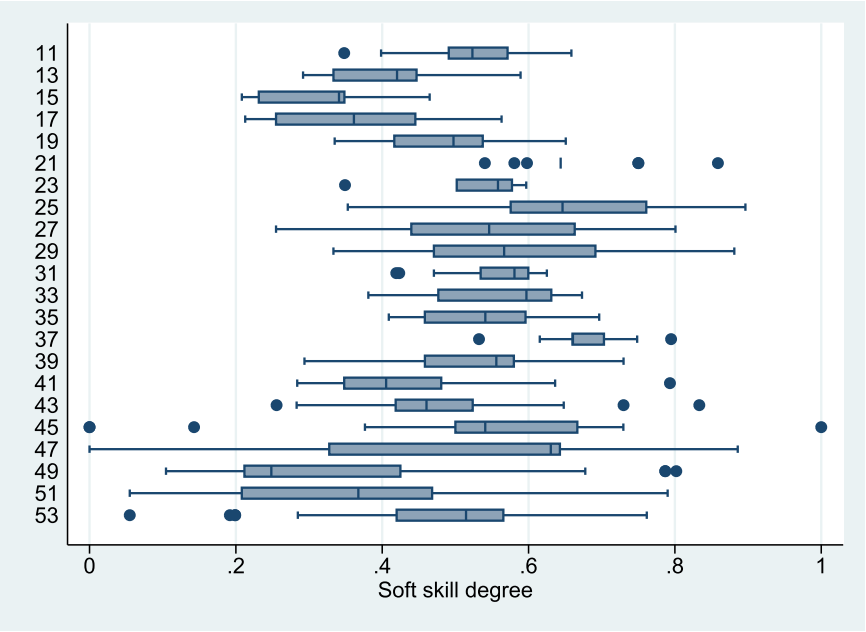

Fig. 3. Distribution of soft skill degree by SOC 2 digit group.

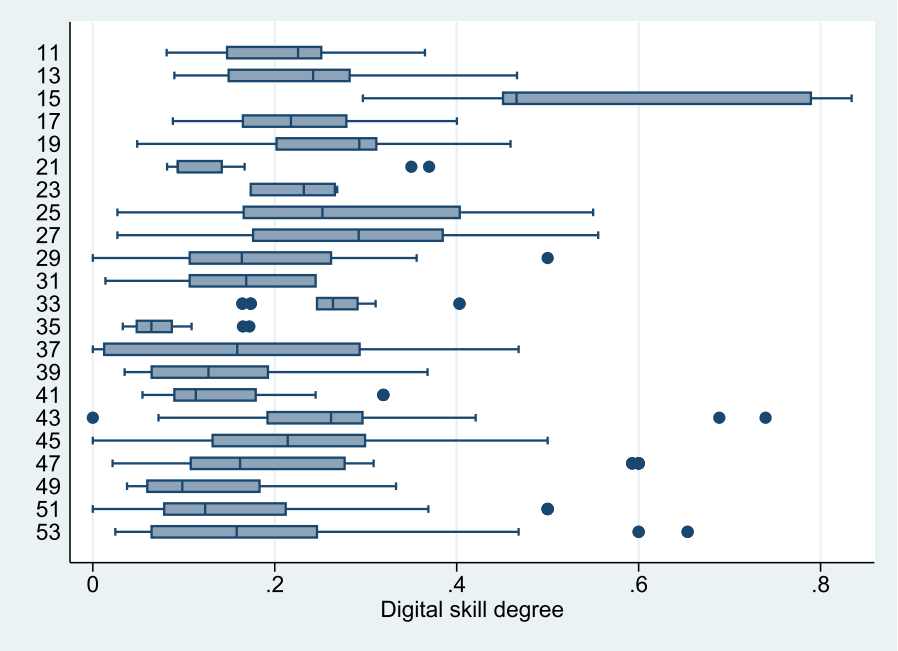

Fig. 4. Distribution of digital skill degree by SOC 2 digit group.

4.2. Technology and jobs

In this section we use the dataset constructed and described in the previous sections to improve our understanding of one of the most pressing problem that is currently affecting advanced economies. In a famous study, Frey and Osborne (2017) estimate the probability of computerization for a large number of detailed occupations in the US. The study spurred a considerable debate about the impact of the new technologies on the labor market.[17] Frey and Osborne (2017) identify the risk of computerization on the basis of the characteristics of selected occupations. These characteristics pertain three major domains that a group of experts identified as the major bottlenecks that computers and artificial intelligence face in completely automating a job. These are perception manipulation (manual dexterity, finger dexterity, cramped workspace), creative intelligence (originality, fine arts), social intelligence (social perceptiveness, negotiation, persuasion, care and assistance for others), and have been assessed on the basis of the O*NET description of occupations. The limit of this approach is that it restricts the information domain to what is available from the official classification description. Using web vacancies it is possible to add the depth and variety of the information set of individual vacancies. In order to grasp the value added of this approach we have used the dataset of Italian web vacancies assigning to each

[17] For a more conservative estimate see Arntz et al. (2016).

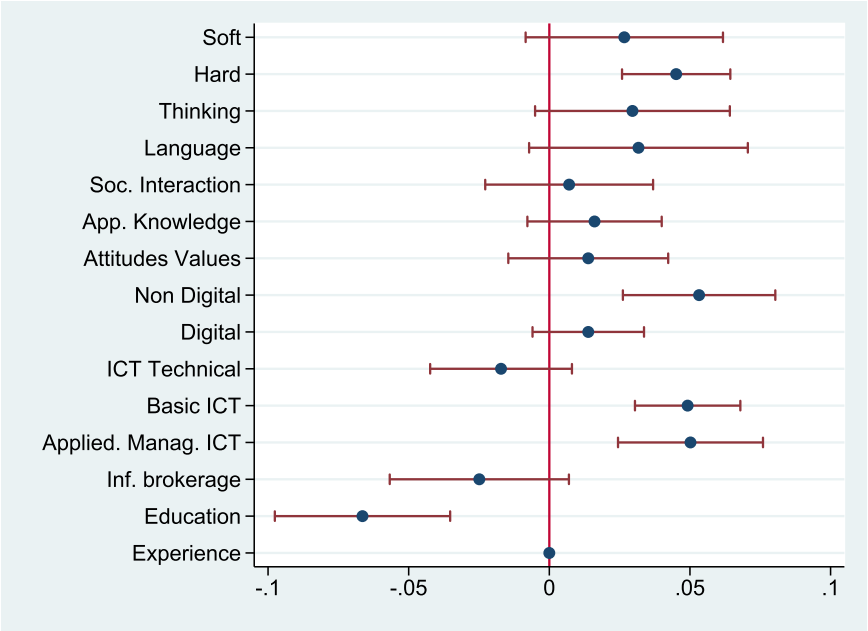

Fig. 5. North South difference in skill requirement within occupation. *Note*: The figure reports the standardized coefficient and the 95% CI of regressions for each skill on a dummy for North. Values above 0 denote a prevalence of that particular skill within occupation in the Northern region. Controls include experience, education, a set of industry and occupation dummies. Errors are clustered at occupation level.

Table 1
Explaining the probability of automation. Aggregate evidence.

	(1)	(2)	(3)	(4)	(5)	(6)
Soft	0.024		0.564**	0.023		
	(0.083)		(0.189)	(0.197)		
Thinking		−0.822**				
		(0.176)				
Soc. Interaction		−0.320*				
		(0.149)				
App. Knowledge		−0.087				
		(0.196)				
Attitude values		−0.079				
		(0.192)				
ICT Technical		−0.004				
		(0.261)				
Basic ICT		−0.106				
		(0.166)				
App. Man. ICT		0.969**				
		(0.253)				
Inf. brokerage		−0.380				
		(0.308)				
Hard					−0.022	−0.569**
					(0.199)	(0.196)
Digital					0.287	−0.001
					(0.175)	(0.147)
Edu.	−0.622**	−0.562**	−0.507**	−0.660**	−0.648**	−0.513**
	(0.078)	(0.078)	(0.116)	(0.116)	(0.117)	(0.117)
Exp. low	0.161	0.209	0.025	0.463*	0.453*	0.032
	(0.104)	(0.107)	(0.136)	(0.190)	(0.190)	(0.140)
Exp. mid	−0.196	−0.222	−0.101	−0.710**	−0.635*	−0.173
	(0.123)	(0.126)	(0.179)	(0.246)	(0.250)	(0.188)
Exp. high	−0.474	−0.404	−0.805*	0.552	0.457	−0.746*
	(0.252)	(0.254)	(0.339)	(0.446)	(0.450)	(0.346)
Const.	−1.319	1.473	0.510	−5.619	−5.091	1.108
	(2.449)	(2.438)	(3.590)	(3.740)	(3.744)	(3.592)
R2	0.312	0.357	0.294	0.361	0.361	0.298
N	892	890	444	448	446	444

Note: OLS regression. Dependent variable: probability of automation. Each variable denotes a degree of intensity of request of each particular skill. Edu defines the fraction of vacancies requiring tertiary education, Exp low, med, high define the fraction of vacancies requiring at lest 2, 3-5 and more than 5 years of experience. Sector and region controls included but not reported. Cols. 3-4 define respectively quartiles 1,2 and 3,4 of the distribution of Hard skill degree. Cols. 5-6 define respectively quartiles 1,2 and 3,4 of the distribution of soft skill degree * denotes significance at 0.05 level, ** at 0.01,

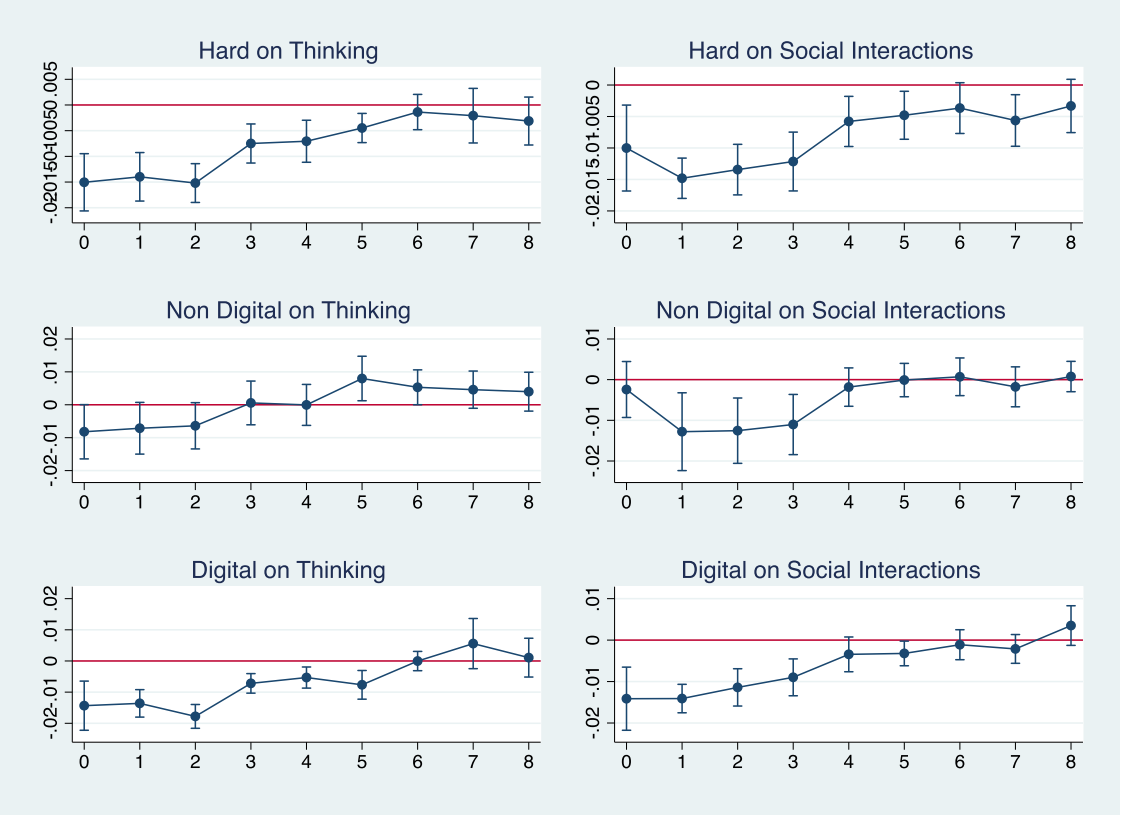

Fig. 6. Interaction effects between skill requirements. *Note*: The figure reports the coefficient and 95% CI of the interaction terms between soft and hard skills for different values of the soft skill variable. Regressions are analogous to the ones in Table 2 and include industry and region dummies, errors are clustered at occupation level.

occupation the probability of automation estimated by Frey and Osborne (2017). We then used these information to enhance the informative content of the classification obtained by Frey and Osborne (2017) by explaining the probability of automation on the basis of the characteristics of skill demand emerging from web vacancies.[18]

The value added of our approach is threefold. First we characterize the probability of automation in terms of detailed skills required by the market allowing a deeper and more complete understanding of the mechanisms that underlie the relationship between automation and jobs. Second we validate, using a data driven approach, results that rest on expert judgment. Third and most importantly we can learn the possible interactions between skills and whether different types of skills are complement or substitutes in a job automability. This type of information is crucial for the education system that can provide the skills that complement computers and robots in the labor market.

We will proceed in two steps. First we will present results at occupation level, subsequently we will use a more disaggregated approach using individual vacancies. Table 1 reports the results where we have aggregated variables at occupation level.[19] Each variable represents therefore a *skill degree*; we also include variables for the level of education, experience in addition to region and industry controls.[20] The first column shows that the probability of automation does not seem to be correlated with the soft skill degree, while it is negatively correlated with the level of education and the degree of experience. Column 2 disentangles among the different sub-components of soft and hard skills. Thinking and social interaction skills are negatively correlated with the probability of automation, in line with the suggestions of the theoretical literature, whereas among hard skills applied management ICT skills are positively correlated with automation. This can be explained with the fact that occupations characterized by the highest probability of automation are often medium level administrative occupations (clerks, tellers, credit analysts) for which applied management ICT skills are required.

The evidence so far suggests the potential advantage of using information on skills, however it is limited by the fact that computing skill degrees at occupation level does not allow to include together the entire skill set. Moreover hard and soft skills may also interact with each other: soft skills can be less relevant for job automability in occupations where hard skills are important, or viceversa. These issues can be analyzed with the full micro dataset to which now we turn.

Table 2reports the results of the analysis performed on the full data set. Now each observation is a vacancy as opposed to an occupation as in the previous table. The dependent variable is the

[18] This exercise rests on the assumption that the impact of the automation process in the US estimated by Frey and Osborne (2017) on occupations is analogous to what is occurring in Italy. Applying US estimates to other advanced economies is a common practice (see for instance the famous paper by Rajan and Zingales (1998) on the relationship between financial development and growth). Our assumption is made easier by the widespread use and pervasiveness of technology in advanced economies.

[19] The table shows that we conduct the analysis with 892 occupations. This number is higher than the 702 occupations identified by Frey and Osborne (2017). This is due to the imperfect correspondence between the SOC (US) and the ISCO (Europe) classification. As stressed in the previous section in case of multiple correspondences we have attributed the same SOC code to multiple ISCO.

[20] As for skills, each variable is constructed as relative importance at occupation level. For instance Education defines the fraction of vacancies in a given occupation that require tertiary education (vs below tertiary), Experience the fraction of vacancies requiring low, medium and high years of experience (vs 0 years). Analogously industry controls identify the relative importance of each industry by occupation.

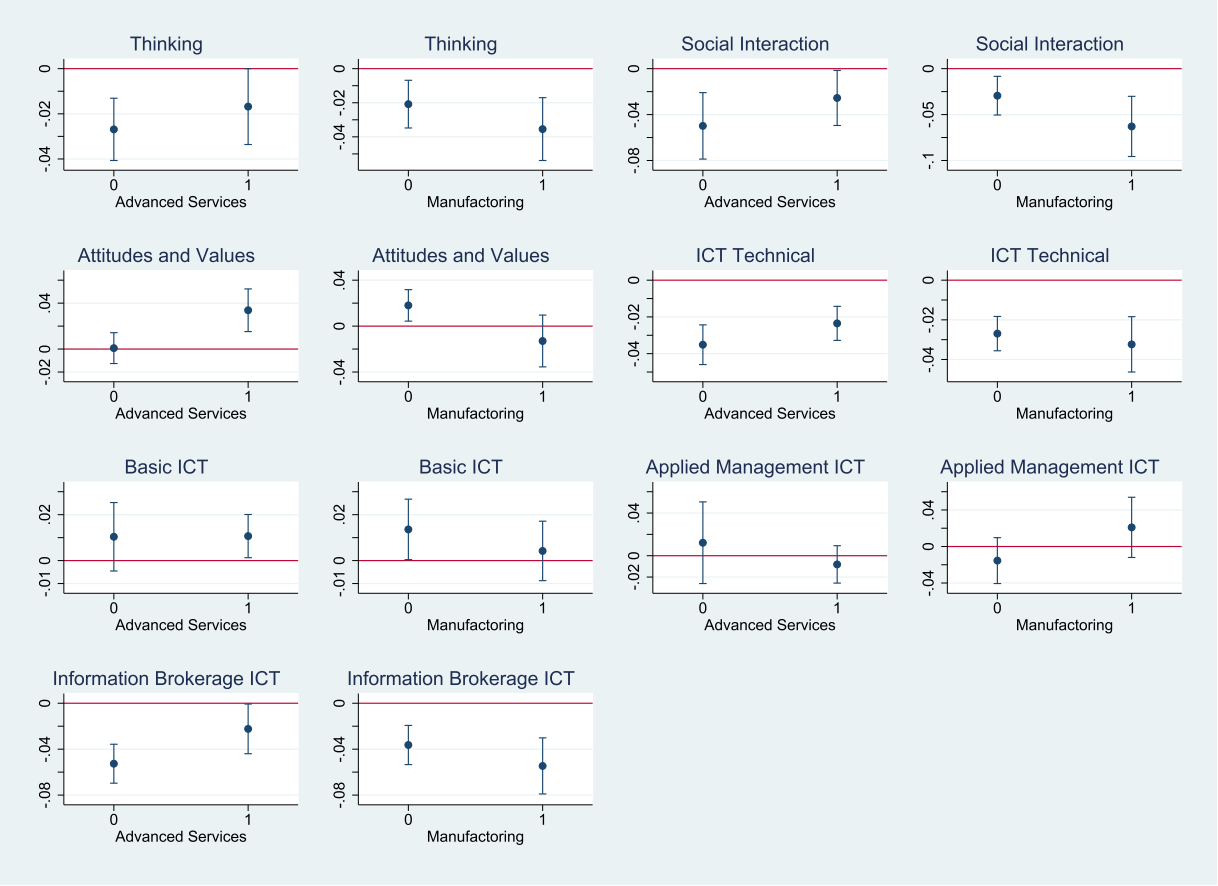

Fig. 7. Interaction effects: sectoral effects. *Note*: The figure reports the coefficient and 95% CI of the interaction terms between soft skills and a sectoral dummy for manufacturing and advanced services. Regressions are analogous to the ones in Table 2 and include industry and region dummies, errors are clustered at occupation level.

probability of automation for each occupation and the explanatory variables are the number of skill mentioned in each vacancy by category, in addition to a variable measuring the degree of experience required and the level of education. The regression is saturated with region and sector dummies while errors are clustered at occupation level. Measuring skill requirements directly and not as a share as in the aggregate regression allows us to include the full set of skills. The table confirms the results of the previous analysis: both hard and soft skills tend to be negatively related to the probability of automation.[21] However while the results on soft skills are confirmed (thinking and social interaction skills are relevant and statistically significant), results on hard skills are different. Non digital hard skills affect negatively the probability of automation while among digital skills ICT technical and information brokerage are the most relevant skills. Thus taking into account the full details of the microdata the relevant digital skills that appear to dampen the probability of automation are advanced ICT skills and skills related to communication and social media. This provides hard evidence to the recent trends outlined by recruiters and experts (McKinsey, 2017).

4.3. Skill interactions

The availability of a granular and detailed dataset allows to explore the possible interactions between skills in explaining the probability of automation for a given occupation. Are soft and hard skills complements or substitute? Is automability affected by them more or less in certain sectors? Table 3 reports the coefficient of the interaction term between soft and hard skills in the regression of Table 2. The table shows that the interaction effect is always positive and significant suggesting that hard and soft skills are substitutes in their relationship with job automability. Fig. 6 explains this relationship with more detail. The figure reports the value and the confidence interval of the hard skills coefficient at different levels of soft skills. At low levels of soft skills the coefficient of hard skills is clearly negative showing a negative effect on the probability of automation. When soft skills are more requested hard skills are not significant or become even positively related to job automability.

Fig. 7 explores interactions at sectoral level. The figure reports the different value of the coefficient of the effect of a given skill on job automability interacting it with a dummy identifying two sectors considered by the literature as the most likely to be affected by technology and AI: manufacturing and advanced services (Information and communication, finance, professional, scientific and technical services). The figure reveals that both social (in particular thinking and social interaction) and digital (in particular ICT technical and information brokerage) skills tend to temper the negative impact of technology on jobs but more so in the manufacturing sector. Soft and digital skills can therefore complement the use of machines and software making the job less substitutable even for occupations that are on average highly automatable.

[21] Clustering of standard errors and controls allow to state that this result holds within occupation and controlling for sector and geographic area.

Table 2
Explaining the probability of automation. Micro evidence.

	(1)	(2)	(3)
Non digital	−0.007**	−0.007**	−0.007**
	(0.003)	(0.003)	(0.003)
ICT technical	−0.025***	−0.025***	−0.024***
	(0.005)	(0.005)	(0.006)
Basic ICT	0.008	0.008	0.009*
	(0.005)	(0.005)	(0.005)
Applied. Manag. ICT	0.012	0.011	0.011
	(0.014)	(0.014)	(0.014)
Inf. brokerage	−0.032***	−0.032***	−0.030***
	(0.010)	(0.010)	(0.009)
Thinking	−0.023***	−0.023***	−0.022***
	(0.006)	(0.006)	(0.006)
Language	−0.007	−0.007	−0.004
	(0.015)	(0.015)	(0.014)
Soc. Interaction	−0.042***	−0.042***	−0.040***
	(0.010)	(0.010)	(0.010)
App. Knowledge	−0.008	−0.008	−0.010
	(0.015)	(0.015)	(0.015)
Attitude values	−0.002	−0.002	0.009*
	(0.005)	(0.005)	(0.005)
Experience		0.003	−0.001
		(0.003)	(0.003)
Education			−0.071***
			(0.015)
R2	0.173	0.173	0.188
N	1299928	1299928	879887

Note: OLS regression. Dependent variable: probability of automation. Sector and region controls included but not reported. Robust standard errors clustered at occupation level. * denotes significance at 0.05 level, ** at 0.01.

Table 3
Soft and hard skill interactions.

	Hard	Digital	Non digital
Thinking	0.001*	0.002**	0.002**
	(0.001)	(0.001)	(0.001)
Social interaction	0.001*	0.003**	0.001*
	(0.001)	(0.001)	(0.001)
Attitudes and values	0.002**	0.003**	0.002*
	(0.001)	(0.001)	(0.001)

Note: Each cell reports the estimated coefficient for the interaction between soft and hard skills in the regression models of Table 2. Controls and additional variables are identical to those in Table 2. Robust standard errors clustered at occupation lavel in parenthesis. * denotes significance at 0.05 level, ** at 0.01.

5. Conclusions

AI and machine learning are not only changing the labor market but are also giving us new tools for analyzing the workforce. In this paper we apply machine learning techniques to web vacancies on the Italian labor market developing a new set of tools for labor market intelligence.

These tools are specifically designed for analyzing firms' skill needs. Our approach allows to shed light on a number of issues. We can calculate, for each occupation, the different types of skills required. We are able to classify those skills into a standard classification system and develop measures of the relevance of soft and hard skills in the latter group, we can drill down detailed digital skills. This allows us to better understand the relationship between labor and automation. Overall this approach provides extremely promising insights that enable to grasp the relevant changes that are affecting jobs and occupations. Web vacancies are thus a useful tool that can complement existing instruments to deliver a more complete picture of the labor market and enable us to make better decisions regarding labor policy.

Acknowledgement

We thank participants to the workshop in Milan and to the conference "The Economics and Policy Implications of Artificial Intelligence" organized by the Technology Policy Institute. The usual disclaimer applies. We thank Tabulaex for granting us access to Wollybi data.

References

Acemoglu, D., 1998. Why do new technologies complement skills? Directed technical change and wage inequality. Q. J. Econ. 113 (4), 1055–1089.
Acemoglu, D., 2002. Technical change, inequality, and the labor market. J. Econ. Lit. 40 (1), 7–72.
Acemoglu, D., Restrepo, P., 2017. Robots and jobs: evidence from US labor markets. Working Paper 23285. National Bureau of Economic Research doi:10.3386/w23285.
Acemoglu, D., Restrepo, P., 2018. Artificial intelligence, automation and work. Working Paper 24196. National Bureau of Economic Research doi:10.3386/w24196.
Arntz, M., Gregory, T., Zierahn, U., 2016. The risk of automation for jobs in OECD countries: a comparative analysis. Social, Employment and Migration Working Papers, No. 189. OECD.
Autor, D.H., Katz, L.F., Krueger, A.B., 1998. Computing inequality: have computers changed the labor market? Q. J. Econ. 113 (4), 1169–1213.
Autor, D.H., Levy, F., Murnane, R.J., 2003. The skill content of recent technological change: an empirical exploration. Q. J. Econ. 118 (4), 1279–1333.
Bessen, J., 2018. AI and jobs: the role of demand. Working Paper 24235. National Bureau of Economic Research doi:10.3386/w24235.
Bhagwati, J., Panagariya, A., 2004. The muddles over outsourcing. J. Econ. Perspect. 18 (4), 93–114.
Boselli, R., Cesarini, M., Marrara, S., Mercorio, F., Mezzanzanica, M., Pasi, G., Viviani, M., 2018a. WoLMIS: a labor market intelligence system for classifying web job vacancies. J. Intell. Inf. Syst. 51 (3), 477–502.
Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., 2014. Planning meets data cleansing. In: Proceedings of the Twenty Fourth International Conference on Automated Planning and Scheduling (ICAPS), pp. 439–443.
Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., 2017. Using machine learning for labour market intelligence. In: Proceedings of the Machine Learning and Knowledge Discovery in Databases–European Conference, ECML PKDD. Springer, Skopje, Macedonia, pp. 330–342. doi:10.1007/978-3-319-71273-4_27. September 18–22.
Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., 2018b. Classifying online Job Advertisements through Machine Learning. Future Generation Comp. Syst. 86, 319–328. doi:10.1016/j.future.2018.03.035.
Card, D., DiNardo, J.E., 2002. Skill-biased technological change and rising wage inequality: some problems and puzzles. J. Labor Econ. 20 (4), 733–783.
Colombo, E., Lovaglio, P., Mercorio, F., Mezzanzanica, M., 2018. A multilevel approach for addressing the representativeness of online job vacancies. Mimeo, University Milano-Bicocca
De Grip, A., Van Loo, J., 2002. The economics of skills obsolescence: a review,. In: de Grip, A., van Loo, J., Mayhew, K. (Eds.), The Economics of Skills Obsolescence, 21. Emerald Group Publishing, pp. 1–26. Research in Labor Economics
Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. The KDD process for extracting useful knowledge from volumes of data. Commun. ACM 39 (11), 27–34. doi:10.1145/240455.240464.
Feenstra, R.C., 1998. Integration of trade and disintegration of production in the global economy. J. Econ. Perspect. 12 (4), 31–50.
Freeman, R.B., 2006. Is a great labor shortage coming? Replacement demand in the global economy. NBER Working Papers 12541. National Bureau of Economic Research, Inc.
Frey, C.B., Osborne, M.A., 2017. The future of employment: how susceptible are jobs to computerisation? Technol. Forecast. Soc. Change 114, 254–280. doi:10.1016/j.techfore.2016.08.019.
Goos, M., Manning, A., Salomons, A., 2014. Explaining job polarization: routine-biased technological change and offshoring. Am. Econ. Rev. 104 (8), 2509–2526.
Hernández, M.A., Stolfo, S.J., 1998. Real-world data is dirty: data cleansing and the merge/purge problem. Data Min. Knowl. Discov. 2 (1), 9–37.
Hoffmann, E., 2003. International statistical comparisons of occupational and social structures. I. In: Hoffmeyer-Zlotnik, J., Wolf, C. (Eds.), Advances in Cross-National Comparison. Springer.
Kroft, K., Pope, D.G., 2014. Does online search crowd out traditional search and improve matching efficiency? Evidence from craigslist. J. Labor Econ. 32 (2), 259–303.
Lee, I., 2011. Modeling the benefit of e-recruiting process integration. Decis. Support Syst. 51 (1), 230–239.
Lovaglio, P.G., Mezzanzanica, M., 2013. Classification of longitudinal career paths. Quality & Quantity 47 (2), 989–1008.
McKinsey, 2017. Job lost, job gained: workforce transitions in a time of automation. Technical Report. McKinsey Global Institute.

Mezzanzanica, M., Boselli, R., Cesarini, M., Mercorio, F., 2015. A model-based evaluation of data quality activities in KDD. Inf. Process. Manag. 51 (2), 144–166. doi:10.1016/j.ipm.2014.07.007.

Rajan, R.G., Zingales, L., 1998. Financial dependence and growth. Am. Econ. Rev. 88 (3), 559–586.

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34, 1–47.

Singh, A., Rose, C., Visweswariah, K., Chenthamarakshan, V., Kambhatla, N., 2010. Prospect: a system for screening candidates for recruitment. In: Proceedings of the Nineteenth ACM International Conference on Information and Knowledge Management. ACM, pp. 659–668.

Yi, X., Allan, J., Croft, W.B., 2007. Matching resumes and jobs based on relevance models. In: Proceedings of the Thirtieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 809–810.